Power Systems

Ignacio J. Pérez-Arriaga *Editor*

Regulation of the Power Sector

Springer

Power Systems

For further volumes:
http://www.springer.com/series/4622

Ignacio J. Pérez-Arriaga
Editor

Regulation of the Power Sector

Springer

Editor
Ignacio J. Pérez-Arriaga
Instituto de Investigación Tecnológica
Universidad Pontificia Comillas
Madrid
Spain

ISSN 1612-1287 ISSN 1860-4676 (electronic)
ISBN 978-1-4471-5033-6 ISBN 978-1-4471-5034-3 (eBook)
DOI 10.1007/978-1-4471-5034-3
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2013932462

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To our students

Preface

Power systems regulation is never at rest. I have been working on regulation of the power sector in more than 30 countries and teaching this topic for about 25 years. Every year, when I set out to prepare the classes in my regulation courses and I examine the previous year's slides, I have to discard a fair amount and create fresh ones. New laws and norms have been approved, new successes and failures have been revealed, new regulatory instruments have been proposed and implemented, ideas that looked promising last year are already out-of-fashion today and others that were discarded in the past seem attractive now.

In regulation there are quieter periods and more active ones. Present times appear to demand a particularly active regulatory response. Energy—and electricity in particular—has recently become a fashionable topic. Energy news is now a staple in the media. Fifteen years ago nobody brought energy as a subject for discussion in a social get-together. In family gatherings my brother, a medical doctor, attracted quite some attention with questions about diverse health matters, while my energy activities were thoroughly ignored. Now I successfully compete with my brother.

There are very good reasons for energy—and electricity in particular—to have become a hot topic, and greater interest in energy and electricity regulation consequently followed. The power sector is at the core of the profound transformation that will be needed in the next few decades to decarbonise the world energy model, while supporting the electrification of transportation and heating and incorporating new features such as the strong penetration of renewable generation—distributed in a significant proportion—active demand participation facilitated by the widespread availability of communication and control technologies, massive efforts in energy conservation and efficiency, the creation of new services and business models at the retail level, the integration of local or national electricity markets into larger regional and transnational entities, and the implementation of effective approaches to achieve universal access to electricity.

The design and deployment of the multiplicity of answers to these difficult challenges in a diversity of contexts requires a thorough knowledge of power systems functions and technologies, mastery of the available modelling tools to represent power system behaviour and to support the decision making in all time

ranges—from protection of electromagnetic transients to long-term capacity expansion of the infrastructures of generation and networks—and the capability to assess any specific decision under engineering, economic, legal and even social or behavioural perspectives.

However, during the past two decades, most of the best universities in the world have disregarded research in power systems, thinking that it was a mature, already under control, well-known technology with little more to offer in terms of research… just to find out now that the power system, the "supreme engineering achievement of the twentieth century" according to the US National Academy of Engineering, has to undergo by the mid-twenty-first century the most drastic overhaul that any industry has ever experienced in such a short time: "to become almost fully decarbonized" if we are to maintain any hope of meeting the targets that climate scientists consider necessary to avoid catastrophic impacts of climate change. Not surprisingly, now we see numerous students in our universities interested in learning at the intersection of the engineering, economic, social and legal aspects of power systems because they want to make professional contributions in a variety of energy fields.

Power sector regulation already is and will increasingly be a key ingredient in any approach to the improvement and adaptation of worldwide energy models. As *The body of knowledge* of the Public Utility Research Center (PURC, http://purc.org) at the University of Florida rightly indicates: "There is a growing consensus that the successful development of utility infrastructure—electricity, natural gas, telecommunications and water—depends in no small part on the adoption of appropriate public policies and the effective implementation of these policies. Central to these policies is development of a regulatory apparatus that provides stability, protects consumers from the abuse of market power, guards consumers and operators against political opportunism, and provides incentives for service providers to operate efficiently and make the needed investments."

A broad range of regulatory policies will be needed to unlock the considerable potential of the new technologies, institutional reforms and business models that are anticipated in the near future. These policies include establishing incentives for efficiency and innovation in operation and new investment of electrical facilities, adapting market rules to the advent of new technologies of production, making markets and governmental mandates compatible, removing non-economic barriers, promoting market competition and mitigating market power, developing public–private partnerships, encouraging new business and financial models and subsidising research and development.

I fell in love with power systems when I studied the subject at MIT in the late 1970s and early 1980s with giants like Fred Schweppe and Gerald Wilson. At that time Power Systems was an active area of research in US universities, with topics ranging from stability analysis of large interconnected systems, optimisation of centralised planning of capacity expansion and operation of generation and networks, improved design of electric machinery, development of advanced tariffs to stimulate efficient demand response or integrated resource planning. The traditional

gap between power systems technology, economics and regulation was getting narrower with innovative proposals on spot pricing of electricity, sophisticated mathematical models to support operational and strategic decisions and preliminary visions of restructuring and deregulation, as in the 1984 seminal book *Markets for Power* by Paul Joskow and Richard Schmalensee.[1]

In 1984, equipped with this baggage and with a small group of bright and dedicated graduate students we commenced, at Comillas Pontifical University of Madrid, Spain, the adventure of creating from scratch a research institute, the Institute for Research in Technology (Instituto de Investigación Tecnológica, IIT, in Spanish) primarily devoted to conducting applied research for the electric power industry on a large variety of topics, mostly relying on detailed mathematical models as our method of analysis. Almost 30 years later, IIT, with more than one hundred researchers, is an international reference institution on power systems research and continues the tradition of blending technology, economics and regulation in a comprehensive approach to a large diversity of power system problems.

Accounting for regulation in the technical and economic analysis of power system problems has probably been the distinct trait of the research conducted at IIT during these almost 30 years. Most of the authors of this book have learned power sector regulation in practice, while consulting, teaching or participating in research projects for regulatory authorities, utilities and various institutions, initially in Latin American countries and later mostly in Europe and obviously also in Spain.

The research activities at IIT allowed me to build a body of knowledge that could be transferred to students in graduate courses. I started teaching a doctoral level course on power sector regulation at the Engineering School of Comillas University of Madrid in 1992, based on material from several intensive courses I had taught in Latin American countries during the previous years. In 2000, I was asked by the Council of European Energy Regulators (CEER) to organise a training course for the staff of European energy regulatory authorities. Under my direction, the first edition took place in 2002 as a very intensive 2-week course, where most of the instructors were experienced regulators. Around this course the Florence School of Regulation was created in 2004 within the Robert Schuman Center at the European University Institute in Florence, Italy. A few months earlier, IIT had accepted the proposal of the World Bank, via the Spanish CEDDET Foundation, of teaching an online course for the staff of governmental institutions and regulatory authorities in Latin America. This required the preparation of written tutorial material suitable for self-learning. This course was successfully run for 4 years. Taking advantage of this written material, conveniently translated into English, in the academic year 2005–2006 we turned the intensive short course of the Florence School of Regulation into an annual course that

[1] Meanwhile Chile had pioneered a revolutionary transformation of its power sector in 1981 (unknown at the time to most of the world), which contained most of the ingredients of the models of restructuring and liberalisation that swept the world one decade later.

combined presential and e-learning modules. This course has continued until the present time, under my direction.

Meanwhile, the initial course in regulation at Comillas University became the backbone of a 1-year Master's Program on the Electric Power Industry and later a 2-year Erasmus Mundus Master's Program, sponsored by the European Commission, with the participation of the Universities of Delft and Paris Sud XI. Under the leadership of the IIT-Comillas team, a new doctoral program was created in 2010, the Erasmus Mundus Joint Doctorate in Sustainable Energy Technologies and Strategies (SETS), jointly offered by Delft University, the Royal Institute of Technology (KTH) at Stockholm and Comillas University, in consortium with the John Hopkins University at Baltimore, USA, Paris Sud XI University and the Florence School of Regulation, again under the sponsorship of the European Commission. The course on power systems regulation is the glue that keeps the different subjects in the program together. Different versions of the regulation course, sometimes shorter, other times combined with other topics, have been employed by IIT researchers in diverse teaching activities for regulatory authorities or in collaboration with universities in Chile, Argentina, Colombia, Uruguay, Costa Rica, Italy, Turkey, Russia, Hungary or Lithuania as well as in Master's programs in the Dominican Republic and Peru.

In the academic year 2008–2009, I was invited by the Massachusetts Institute of Technology (MIT) to teach a course on "Engineering, economics and regulation of the electric power sector". This was again an excellent opportunity to test and to improve the same teaching material with a quite different audience. The course has been very well accepted and it has become a permanent course during the Spring term at MIT.

Finally, we have decided to collect our experiences, writings and accumulated knowledge in this book. Contrary to what the term "edited-by" usually suggests—a more or less coherent collection of papers by different authors, which have been put together by an editor—this book is the "corpus" of knowledge distilled by the IIT team of researchers after about 25 years of teaching, research and consultancy in many countries.

This book is a collective work. It is centred around the regulation course that has been taught under multiple formats and in different contexts and that has been enriched by the experiences and the individual research paths of the co-authors of the different chapters. The e-learning material for CEDDET, which was later used at the Florence School of Regulation, has been the starting point. However, this book goes further into the complexities of regulation, and it cannot be considered an introductory text. In many respects it covers topics up to the current frontier of knowledge. This frontier changes quickly and I am sure that this book, if well-accepted, will have to be frequently updated.

The book adopts the point of view of the regulator. This is a logical consequence of having worked and taught frequently the course for the staff of regulatory commissions and public officials and, in the case of two of the authors, for having been commissioners in energy regulatory commissions. An interest in

electricity regulation may have multiple origins: government, staff of regulatory authorities, lobbyists, consumer associations, consultants or professionals of electricity utilities or energy service companies. Because of the institution they represent, sometimes these professionals defend regulatory positions that are not fully aligned with the common social interest. This is not the case of the independent regulator, whose goal must be to promote the public good. This undoubtedly encourages independent and creative thinking in search of what is best for society as a whole, and this is why we have adopted the independent regulator's perspective.

The book is directed at regulators, policy decision makers, business managers and researchers. It is a pragmatic text and we expect that power system professionals and students at all levels will benefit from the stock of blended theory and real-world-derived know-how that the book contains.

The IIT team of researchers has worked extensively in Spain, Latin America, the European Union and North America, but much less in the non-EU Europe, Africa, Asia and Oceania, where very interesting regulatory developments have also taken place. If the book is well-received, we shall try to correct this imbalance in future editions.

Arranged in four parts, the book addresses both traditional regulatory frameworks and also liberalised and re-regulated environments. First, an introduction gives a full characterisation of power supply including engineering, economic and regulatory viewpoints. Part II presents the fundamentals of regulation and the third part looks at the regulation of particular components of the power sector in detail. Advanced topics and subjects still open or subject to dispute form the content of Part IV.

It has been very encouraging how well-accepted the different courses on regulation that have led to this book have been. The book is dedicated to our students, who have overcome the difficulties of the frequently limited teaching material, have accompanied us in the learning of this complex subject and have inspired us with their sharp questions and their enthusiasm. We also thank the rest of our colleagues at IIT, both present and past, who are not explicit co-authors of the book but who have participated and contributed to its content in many ways. We express our gratitude to the numerous companies and institutions that have trusted us. The pioneer regulators of CEER, Jorge Vasconcelos, Jacques de Jong and Pippo Ranci, had the vision of asking for and supporting the training course for energy regulators that has become the flagship of the Florence School of Regulation. Miguel Angel Feito of CEDDET conceived the idea of our first e-learning course, where the very preliminary material for this book was prepared. The editorial staff of Springer have been always supportive of this project, and we thank them for their patience with our delays in delivering the manuscript and their careful editorial work. Tommy Leung from MIT and Paolo Mastropietro at IIT have carefully reviewed several chapters, pinpointing mistakes that otherwise

would had remained unnoticed to the authors. Margaret Clark has done a magnificent job of turning our texts of convoluted Spanish and, even worse, poor English, into elegant English prose. Some sections of the book have not benefited of her careful supervision, as the reader may easily discover.

Madrid, January 2013 Ignacio J. Pérez-Arriaga

Contents

1 Technology and Operation of Electric Power Systems 1
Damián Laloux and Michel Rivier

2 Power System Economics 47
Mariano Ventosa, Pedro Linares and Ignacio J. Pérez-Arriaga

3 Electricity Regulation: Principles and Institutions 125
Carlos Batlle and Carlos Ocaña

4 Monopoly Regulation 151
Tomás Gómez

5 Electricity Distribution 199
Tomás Gómez

6 Electricity Transmission 251
Michel Rivier, Ignacio J. Pérez-Arriaga and Luis Olmos

7 Electricity Generation and Wholesale Markets 341
Carlos Batlle

8 Electricity Tariffs 397
Javier Reneses, María Pía Rodríguez and Ignacio J. Pérez-Arriaga

9 Electricity Retailing 443
Carlos Batlle

10 Regional Markets 501
Luis Olmos and Ignacio J. Pérez-Arriaga

11 Environmental Regulation 539
Pedro Linares, Carlos Batlle and Ignacio J. Pérez-Arriaga

12 Security of Generation Supply in Electricity Markets 581
Pablo Rodilla and Carlos Batlle

13 Electricity and Gas 623
Julián Barquín

14 Challenges in Power Sector Regulation 647
Ignacio J. Pérez-Arriaga

Annex A: Case Example of Traditional Regulation of the Electric Power Sector 679

Annex B: Grandma's Inheritance Theorem 719

Index 723

Chapter 1
Technology and Operation of Electric Power Systems

Damián Laloux and Michel Rivier

Electric power systems are generally regarded to be the largest and most complex industrial systems ever built.

Today's developed societies are inconceivable without electric power. Indeed, in societies where basic human needs such as food, health care, decent housing and education are met, "electricity and running water" are usually the two primary conveniences immediately associated with quality of life. Electricity, with its versatility and controllability, instant availability and consumer-end cleanliness, has become an indispensable, multipurpose form of energy. Its domestic use now extends far beyond the initial purpose, to which it owes its colloquial name ("light" or "lights"), and has become virtually irreplaceable in kitchens for refrigerators, ovens and cookers or dishwashers and any number of other appliances, and in the rest of the house as well, for air conditioning, heating, radio, television, computers and the like. But electricity usage is even broader if possible in the commercial and industrial domains: in addition to providing power for lighting and air conditioning, it drives motors with a host of applications such as lifts, cranes, mills, pumps, compressors or other machine tools. Industrial activity that uses no manner of electricity is simply unimaginable.

Beginning a book on the restructuring and regulation of the electricity industry with a basic description of the electric power system is hardly pointless: on the contrary, it is intended to refresh basic concepts and establish firm grounds on which to build further knowledge. Although the content of this chapter may well be familiar, it will give readers an overview of electricity systems while serving as an introduction to matters that may be decisive for a more comprehensive understanding of the chapters that follow.

The chapter is organised into six sections of varying importance: the core content is found in Sects. 1.3 and 1.4, which address the technology and

D. Laloux (✉)
ICAI School of Engineering, Universidad Pontificia Comillas,
Alberto Aguilera 25, 28015 Madrid, Spain
e-mail: dlaloux@des.icai.upcomillas.es

M. Rivier
IIT, Instituto de Investigación Tecnológica, ICAI School of Engineering,
Universidad Pontificia Comillas, Alberto Aguilera 25, 28015 Madrid, Spain

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_1,

management of electric power systems. The remaining sections revolve around these two, discussing the causes determining and the consequences arising from these issues. Specifically, Sect. 1.1 describes the properties that distinguish electricity from other products, and Sect. 1.2 places electric power systems in their historical context. Section 1.3 describes the components of a power system from generation to consumption, whereas Sect. 1.4 reviews the operation and the decision-making process on different time scales, from long term to real time. Section 1.5 deals with the importance of environmental impact, one of the issues that clearly conditions the future of the industry. The chapter concludes with a brief section on the future of electric power systems: foreseeable progress, the potential of technology, the challenges to be addressed and the radical change visible on the horizon. The intention is not, of course, to predict the future but simply to draw attention to issues that may induce significant change in electric power systems.[1]

1.1 Background

At first glance, electricity may appear to be a commodity much like any other. In fact, that perspective may be behind the revolution that has rocked electric power systems worldwide, as they have been engulfed in the wave of liberalisation and deregulation that has changed so many other sectors of the economy. And yet electricity is defined by a series of properties that distinguishes it from other products and makes it harder implementing such changes in the electricity industry.

The chief characteristic of electricity as a product that differentiates it from all others is that it is not liable, in practice, to being stored. Electricity can, of course, be stored in batteries, but price, performance and inconvenience presently make this impractical for handling the amounts of energy usually needed in the developed world. This may change in the future but, at present, electricity must be generated and transmitted as it is consumed, which means that electric systems are dynamic and highly complex, as well as immense. At any given time, these vast dynamic systems must strike a balance between generation and demand, and the disturbance caused by the failure of a single component may be transmitted across the entire system almost instantaneously. This sobering fact plays a decisive role in the structure, operation and planning of electric power systems, as discussed below.

Another peculiarity of electricity has to do with its transmission: this is not a product that can be shipped in "packages" from origin to destination by the most suitable medium at any given time. Electric power is transmitted over grids in which the pathway cannot be chosen at will, but is determined by the laws of physics (in this case Kirchhoff's laws), whereby current distribution depends on impedance in the lines and other elements through which electricity flows. Except

[1] This chapter is based on Pérez-Arriaga et al. [3].

in very simple cases, all that can be said is that electric power flows into the system at one point and out of it at another, because ascribing the flow to any given path is simply not possible. Moreover, according to these laws of physics, the alternative routes that form the grid are highly interdependent, so that any variation in a transmission facility may cause the instantaneous reconfiguration of power flows; and that, in turn, may have a substantial effect on other facilities. All this renders the dynamic balance referred to in the preceding paragraph even more complex.

Indeed, for all its apparent grandiloquence, the introductory sentence to this chapter may be no exaggeration. The combination of the extreme convenience of use and countless applications of electricity on the one hand and its particularities on the other has engendered these immense and sophisticated industrial systems.

Their size has to do with their scope, as they are designed to carry electricity to practically any place inhabited by human beings from electric power stations located wherever a supply of primary energy, for instance in the form of potential or kinetic energy in moving water or of chemical energy stored in any of several fuels, is most readily available. Carrying electric power from place of origin to place of consumption calls for transmission and distribution grids or networks that interconnect the entire system and enable it to work as an integrated whole.

The sophistication of electric power systems is a result of the complexity of the problem of permanently maintaining the balance of supply and demand, with the characteristics discussed above. This dynamic equilibrium between generation and demand is depicted in the highly regular patterns followed by the characteristic magnitudes involved: the value and frequency of voltage and currents as well as the waveform of these signals. Such regularity is achieved with complicated control systems that, based on the countless measurements that constantly monitor system performance, adapt its response to ever-changing conditions.

A major share of these control tasks is performed by powerful computers in energy management centres running a host of software applications: some estimate demand at different grid buses several minutes, hours, days or months in advance; other models determine the generation needed to meet this demand; yet other programs compute the flow in system lines and transformers and the voltage at grid buses under a number of assumptions on operating conditions or components failure, and determine the most suitable action to take in each case. Others study the dynamic behaviour of the electric power system under various types of disturbance. Some models not only attempt to determine the most appropriate control measures to take when a problem arises, but also to anticipate their possible occurrence, modifying system operating conditions to reduce or eliminate its vulnerability to the most likely contingencies.

This, however, is not all: the economic aspect of the problem must also be borne in mind. The actors that make the system work may be private companies that logically attempt to maximise their earnings or public institutions that aim to minimise the cost of the service provided. In either case, the economic implications of the decisions made cannot be ignored, except, of course, where system security is at stake.

The system nearly always operates under normal conditions, leaving sufficient time to make decisions that are not only safe, but economically sound. Hence, when a foreseeable rise in demand takes place during the day, power should be drawn from the facilities with unused capacity that can generate power most efficiently. The objective is to meet daily load curve needs with power generated at the lowest variable cost. This new dimension in the operation of electric power systems is present in all time scales: from the hourly dispatch of generating plants to the choice of which units should start up and stop and when, including decisions on the use of hydroelectric reserve capacity, maintenance programming and investment in new facilities.

Moreover, all these decisions are made in a context of uncertainty: about the future demand to be met, availabilities of plants and of resources such as water in the reservoirs, wind or sunlight, the prices of the various factors involved in the production process, fuels in particular, and even the legislation in effect when long-term decisions are to be implemented.

1.2 Early History

The first electric light systems, installed around 1870, consisted of individual dynamos that fed the electrical system (arc lamps) in place in a single residence. Joseph Swan and Thomas Edison worked separately on the incandescent light bulb until 1880 and Edison spawned the idea of increasing the scale of the process by using a single generator to feed many more bulbs. The first town in the world to establish a public electricity supply was Godalming, UK, in September 1881: a waterwheel on the river Wey powered a Siemens alternator and a dynamo that fed 7 arc lights and 34 Swan incandescent lights. In 1882, Edison's first generator, driven by a steam turbine located on Pearl Street in lower Manhattan, successfully fed a 100-V direct current (DC) to around four hundred 80-W bulbs in office and residential buildings on Wall Street. Shortly thereafter, London's 60-kW Holborn Viaduct station was commissioned, which also generated 100-V direct current. This local generation and distribution scheme was quickly adopted, exclusively for lighting, in many urban and rural communities worldwide.

The invention of the transformer in France in 1883–1884 revealed, in a process not free of controversy, the advantages of alternating current (AC), which made it possible to conveniently raise the voltage to reduce line losses and voltage drops over long transmission distances. Alternating, single-phase electric current was first transmitted in 1884, at a voltage of 18 kV. On 24 August 1891, three-phase current was first transmitted from the hydroelectric power station at Lauffen (Germany) to the International Electro-Technical Exhibition at Frankfurt, 175 km away. Swiss engineer Charles Brown, who with his colleague and fellow countryman Walter Boveri founded the Brown-Boveri Company that very year, designed the three-phase AC generator and the oil-immersed transformer used in the station. In 1990, the Institute of Electrical and Electronic Engineers, IEEE,

agreed to establish 24 August 1891 as the date marking the beginning of the industrial use and transmission of alternating current.

The transmission capacity of alternating current lines grows fast with voltage, whereas the cost per unit of power transmitted swiftly declines. There was, then, an obvious motivation to surmount the technological barriers limiting the use of higher voltages. Voltages of up to 150 kV were in place by 1910 and the first 245-kV line was commissioned in 1922. The maximum voltage for alternating current has continued to climb ever since. Nevertheless, direct current has always been used as well, since it has advantages over alternating current in certain applications, such as electrical traction in the past and especially electricity transmission in overhead, underground or submarine lines, when the distances are too long for alternating current.

The alternating voltage frequency to be used in these systems was another of the basic design parameters that had to be determined. Higher frequencies can accommodate more compact generating and consumption units; this advantage is offset, however, by the steeper voltage drops in transmission and distribution lines resulting from their use. All of Europe, Africa, Asia, Australia, New Zealand and most of South America adopted a frequency of 50 Hz, whereas Japan, Taiwan, North America and some parts of northern South America opted for a frequency of 60 Hz. Low voltage consumers within most 50 Hz regions are supplied at 230 V, while in 60 Hz regions the voltage is generally 110 V. The International Electrotechnical Commission was created in 1906 to standardise electrical facilities everywhere as far as possible. It was, however, unable to standardise frequency, which continues to divide countries around the world into two separate groups.

The advantages of interconnecting small electric power systems soon became apparent. The reliability of each system was enhanced by the support received from the others in the event of emergencies. Reserve capacity could also be reduced, since each system would be able to draw from the total grid reserve capacity. With such interconnections, generating units able to meet demand most economically at any given time could be deployed. This affords particular advantage when peak demand time frames vary from one system to another and when the generation technology mix (hydroelectric and steam, for instance) likewise differs. In 1926, the British Parliament created the Central Electricity Board and commissioned it to build a high-voltage grid that would interconnect the 500 largest generation stations then in operation.

1.3 Components of a Power System

Electric power systems have developed along more or less the same lines in all countries, converging toward a very similar structure and configuration. This is hardly surprising, bearing in mind the very specific characteristics of the product sold. As mentioned earlier, electricity generation, transmission, distribution and retailing are inevitably conditioned by the fact that generation and demand must be in instantaneous and permanent balance. The relevance of technical factors in

maintaining such large-scale systems in dynamic equilibrium cannot be overlooked. A disturbance anywhere in the system may endanger the overall dynamic balance, with adverse consequences for the supply of electricity across vast areas, even whole regions of a country or an entire country. It is perhaps for this reason that the existence of sophisticated real-time control, supervision and monitoring systems, together with distributed local protection devices, is what chiefly differentiates the configuration and structure of electric power systems from other industrial activities in terms of technology. The functions typical of any industry, such as production, shipping, planning and organisation, are also highly specialised in the electricity industry.

Electric power systems, like any other industry, are made up of production centres, or generating plants; transmission (equivalent to transport or shipping in other industries), i.e. the high-voltage grid (network is also used, equally); distribution, in this case the low-voltage grid or network; and consumption (also termed demand or load), in addition to the associated protection and control systems. More formally, system configuration and structure are as depicted in Fig. 1.1.

Production centres typically generate electricity at 6–20 kV and immediately transform this power into voltages of hundreds of kilovolts (132, 230, 400, 500 and 700 kV are relatively common values) to optimise long-distance transmission over electric lines to the areas where consumption is most intense.

Raising the voltage makes it possible to transmit large amounts of electric power (the entire output from a nuclear-powered generator, for instance) over long distances using reasonably inexpensive overhead lines technology with modest

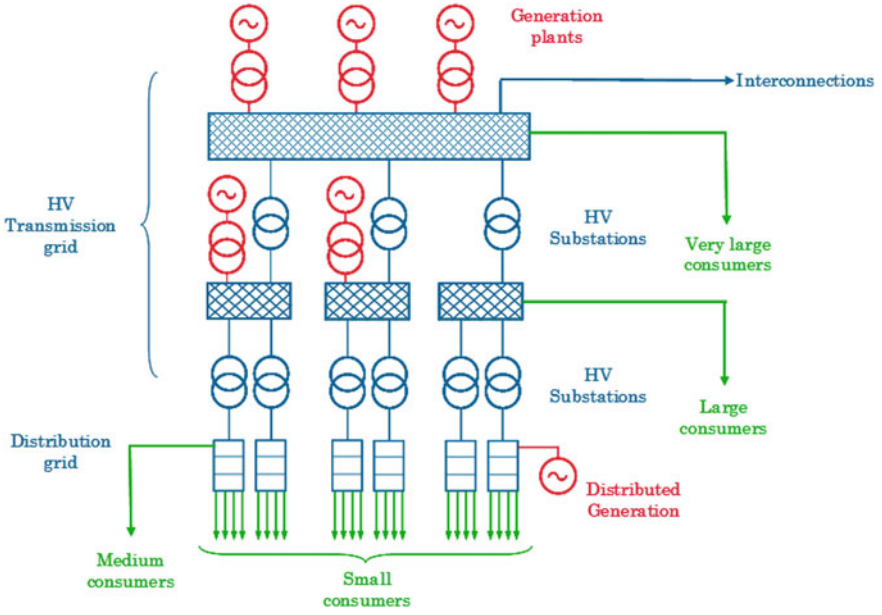


Fig. 1.1 Electric power system configuration and structure

line losses. The transmission grid interconnects all the major production and consumption centres, generally forming a very dense mesh to guarantee high reliability, with alternative pathways for the supply of electric power in the event of failure in a few of the lines.

These electric power transmission motorways are interconnected at communication nodes known as electric substations; the regional grids connect to these substations at a somewhat lower voltage (132, 66 or 45 kV in Spain, for instance) and in turn feed local distribution networks, which bring electric power to consumers at less hazardous voltages, adapted to consumer needs: 20, 15, 6.6, 3 or 1 kV and 380, 230 or 110 V.

Successive substations step the working voltage down in several phases and centralise the measuring and protection devices for the entire transmission grid.

The configuration of these distribution grids is usually radial, with tentacles stretching out to even the most remote consumption points. As the lines are split up at each step, the grids carry less and less power and consequently can operate at lower voltages.

Consumers connect to the voltage level best suited to their power needs, in accordance with the basic principle that the smaller the power capacity, the lower the voltage. This means that highly energy-intensive businesses (iron and steel plants and mills, aluminium plants, railways and the like) connect directly to the high voltage grid; other major consumers, such as large factories, receive power at a somewhat lower voltage, while small consumers such as households, retailers and small factories, are connected to the low-voltage network.

Based on a roughly reciprocal principle, generating stations with a very small output feed their electric power directly into the distribution network, instead of connecting to the high-voltage grid. Such generators run small hydroelectric, photovoltaic, wind, combined heat and power (CHP) or other types of modular power stations, and they are generally termed "distributed generation". Medium size wind or solar farms, from tens up to one hundred megawatts or more, connect to the medium or high voltage distribution network.

The items below focus on these chief components of electric power systems: demand, production, transmission and distribution.

1.3.1 Consumption

Demand growth

Electricity demand has experienced high, sustained growth from the outset. The creation of standards for the electricity "product", in terms of voltage, frequency and current, paved the way for the enormous boom in electricity consumption. This in turn laid the grounds for further standardisation of electric fixtures and facilities, from light bulbs and motors to PCs, dramatically lowering manufacturing costs and enhancing product versatility, making it possible to use a given electrically powered item virtually anywhere.

Table 1.1 Electricity consumption (TWh), 1980–2035 (IEA [2])

Region	1980	2008	2020	2035
OECD	4,739	9,244	10,097–10,488	10,969–12,101
Non-OECD	971	7,575	12,375–13,233	16,660–20,820
World	5,711	16,819	22,472–23,721	27,629–32,922

Electric power consumption is one of the clearest indicators of a country's industrial development, and its increase closely parallels GDP growth. As noted earlier, there are scarcely any production processes or sectors involved in creating wealth that do not require electricity. Furthermore, electric power consumption has also been used as a measure of social development. Electricity consumption per capita and especially the degree of electrification in a country (i.e. the percentage of the population living in homes with electricity) provide a clear indication of the standard of living. This is not surprising, since such basics as lighting, a supply of clean water, refrigerators and other household appliances depend on access to electricity. Table 1.1 shows the growth in electricity consumption by region from 1980 to 2035 considering different scenarios (in the columns showing future projections, the higher end of the range reflects the *current scenario*, with no significant changes in energy policies, and the lower end a more restrictive situation, aiming at achieving the 2 °C global warming goal by limiting the concentration of greenhouse gases in the atmosphere). These data, taken from the IEA World Energy Outlook 2010, assume sustained, planet-wide growth throughout the projection period. The same study anticipates that world electricity demand will continue to increase more sharply than any other final form of energy and in particular points to a threefold increase in electricity demand in China between 2008 and 2035.

Electrification rates and electricity consumption per capita vary widely from one area of the world to another. According to World Energy Outlook estimates, in 2009 the number of people without access to electricity amounted to 20 % of the global population. Its report illustrates the inequities with a striking example: New York City's population of 19.5 million consumes roughly the same amount of electricity per year, 40 TW-hours, as the 791 million sub-Saharan Africans (excluding South Africa). Table 1.2 gives an overview of electricity access by region in 2009.

The growth in electricity consumption is not limited to developing countries; however, it has definitely steadied in developed countries but is certainly not flat. Whilst the industrial world's consumer mentality may be partly responsible for driving such growth, it is nonetheless true that new uses are constantly being found for electric power. The generalised use of air conditioning in developed countries is an obvious example and one that has brought about a radical change in seasonal consumption curves, as explained below.

Moreover, electricity consumption has continued to grow despite substantial improvement in the efficiency of most equipment and processes using electric power, which reduces the input energy needed to attain a given result. More and more voices are being raised calling for the need to rationalise the consumption of electricity and all other forms of energy. Aware of the environmental impact of

Table 1.2 Electricity access in 2009 (IEA [2])

Region	Population without electricity (millions)	Electrification rate (%)
Africa	587	41.9
North Africa	2	99.0
Sub-Saharan Africa	585	30.5
Developing Asia	799	78.1
China and East Asia	186	90.8
South Asia	612	62.2
Latin America	31	93.4
Middle East	22	89.5
Developing countries	1,438	73.0
Transition economies and OECD	3	99.8
World	1,441	78.9

such consumption and the vast amount of natural resources that are literally going up in smoke, these voices rightly call for intergenerational solidarity so they can leave to coming generations an ecologically acceptable planet whose energy resources have not been depleted.

Hence the importance of demand-side management (DSM), a term coined decades ago in the United States to encompass techniques and actions geared to rationalising the consumption of electric power. The aim, essentially, is a more efficient consumption to reduce the enormous investment involved in the construction of new stations and the substantial cost of producing electricity. Demand-side management should, therefore, be an active component of future electric power systems, reflecting real-time electricity costs, including the potential internalisation of environmental costs, which are so often ignored. The role and suitable regulation of this activity is one of the challenges facing the future electricity industry.

A critical objective in this regard is to send consumers, the final and key link in the electricity chain, correct economic signals reflecting the actual costs incurred by electricity consumption at any given time. Pricing should be designed to make consumers aware of the real economic and environmental cost of meeting their power needs, given their consumption patterns in terms of hourly profile and total load. In the medium term, this should stimulate domestic, commercial and industrial users to monitor and actively control electric consumption, in much the same way that discriminatory hourly telephone rates encourage customers to make non-urgent long-distance calls at off-peak times. Similarly, customers will voluntarily reduce electricity consumption by foregoing the most superfluous applications at times when higher prices signal that expensive resources are being deployed or that the margin between the demand and supply of electric power is narrow.

The ability of demand to respond to pricing is normally measured by a parameter known as price elasticity of demand. This is defined as the percentage variation in the consumption of electricity (or any other product) in response to a

unit variation in its price. Generally speaking, electricity demand is scantly elastic in the short term; in other words, the reaction to changes in price is small, although this assertion is more accurate for some types of consumers than others. Such limited elasticity is arguably due to the mentality prevailing until very recently in the electricity industry: continuity of supply was regarded to be a nearly sacred duty, to be fulfilled at any price. Consumers, who were identified as *subscribers* rather than *customers*, were merely passive recipients of the service provided. Advances in communications technology, in conjunction with the liberalisation of the electric and energy industries in much of the world, will radically transform the role of consumers to a more active one, although the change in mentality probably will not happen readily or quickly. Nonetheless, the years to come will very likely witness the coming of age and intensification of the role played by demand in the electricity industry, which will come to carry the same weight as other areas, such as generation. Elasticity will grow, particularly in the short term, and demand patterns will adapt better to system conditions, although this will not necessarily mean a total demand reduction.

Demand profiles

From a technical standpoint, consumption comprises a variety of items. The two most important are power and energy. Power, measured in watts (W), is the energy (Wh) required per unit of time. Power, therefore, is the instantaneous energy consumed; it is also referred to as demand or load. Since electric energy is not stored, electric facilities must be designed to withstand the maximum instantaneous energy consumed, in other words, the maximum power load on the system throughout the consumption cycle. Therefore, not only the total electric capacity needed, but the demand profile over time is especially important in determining consumption. Such profiles, known as load curves, represent power consumed as a function of time. It may be readily deduced that a given value of energy consumed may have a number of related load profiles. Some may be flat, indicating very constant electricity consumption over time, while others may have one or several very steep slopes, denoting highly variable demand. An aluminium plant working around the clock 365 days a year and a factory operating at full capacity only during the daytime on weekdays would exemplify these two types of profiles.

Load profiles are usually repetitive over time. For example, the weekday demand profile is normally very uniform, as is the weekly load profile during a certain season. Therefore, depending on the time scale considered, the load profile to be used may be daily, weekly, monthly, seasonal, yearly or even multi-yearly. Load profiles also have economic significance, as will be seen in the discussion below: for any given demand level, a flat load profile can be accommodated less expensively to than a spiked curve. For this reason, load curves are one of the most important parameters considered in the methods used to set tariffs.

The sum of all the individual consumption curves for an electric power system yields the total daily, weekly, monthly, seasonal, yearly and multi-yearly load curves, each with a characteristic and highly significant power profile. The figures below, taken from the website of the Spanish Transmission System Operator

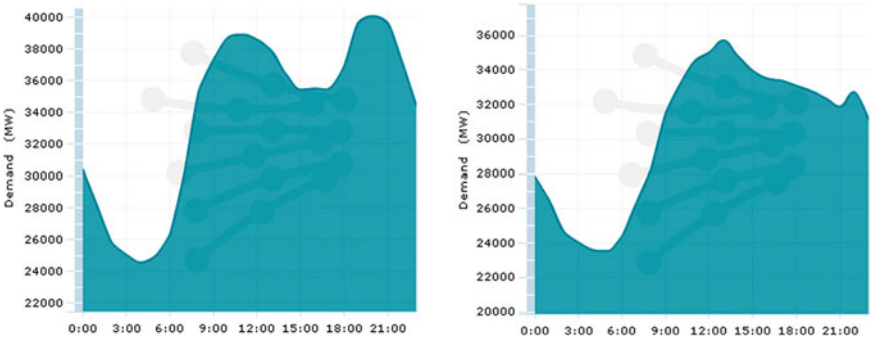


Fig. 1.2 Electricity demand daily pattern in the Spanish system (January 19th, 2011, *left*, and July 20th, 2011, *right*)

(Red Eléctrica de España) show the demand profiles for the Spanish system. Figure 1.2 illustrates how the daily pattern changes with the season (winter vs. summer). Figure 1.3 shows how demand in the working days is significantly larger than in the weekends (Saturday larger than Sunday). Finally, Fig. 1.4 illustrates how demand changes throughout the year, increasing in winter and summer due to the heating and air conditioning consumption. Additionally, in order to illustrate how demand patterns strongly depend on the particularities of each electric power system, Fig. 1.5 show the chronological hourly load for the ERCOT power system. It can be observed how not only the annual pattern presents a significantly different shape, but also the daily ones both in winter and summer.

Demand forecasting is an essential problem to solve in anticipating the conditions under which the system will be operating in the short, medium and long term. The normal procedure is to base the prediction on historical data adjusted to take account of factors affecting the expected load. The most important of these factors include temperature, since many electrical devices are used for heating or cooling; number and dates of working days, to account for the difference in consumption on business days and holidays; and economic growth, in view of the above-mentioned close relationship between economic activity and electricity consumption. Therefore, consumption at any given time can be predicted reasonably well from time series data corrected for foreseeable variations in growth, working days and temperature, taking account as well of special events that may have a substantial effect on demand.[2]

Aggregate electricity consumption can also be represented as a monotonic load curve called the load duration curve, which is particularly useful in certain applications and studies. Such curves represent the length of time that demand exceeds a given load. The line in Fig. 1.6 is the approximate monotonic load curve

[2] Conversely, electricity consumption is a reliable and early indicator of economic crises and recoveries, as shown—once again—in those countries affected by the world economic crisis that started in 2008.

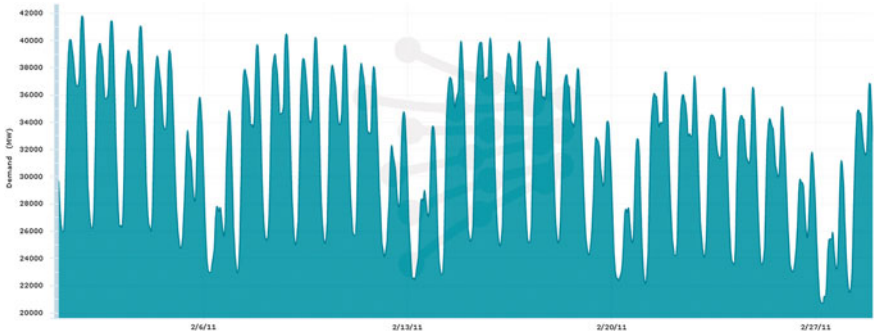


Fig. 1.3 Electricity demand monthly pattern in the Spanish system (February, 2011)

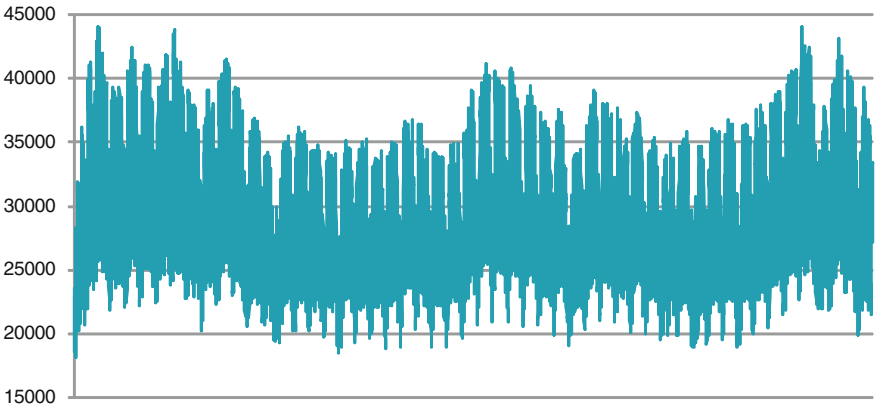


Fig. 1.4 Electricity demand yearly pattern in the Spanish system (2010)

for the Spanish system in 2010: the abscissa values represent time in hours and the ordinate values demand in megawatts. Therefore, each point on the curve indicates the total hours during the year that demand exceeded a given value.

The load duration curve can be plotted directly from the chronological load curve by ranking demand in descending order, and therefore losing any chronological information. The integral of the load duration curve equals the energy consumed in the time frame considered. Note, however, that whereas a given load profile can have only one load duration curve, the opposite is not true.

Although the chronological information contained in load curves is lost in monotonic curves, the latter are widely used for their simplicity. Probabilistic monotonic curves are commonly applied in prospective studies, which are based on demand forecasts subject to some degree of uncertainty; in this case, the x-axis values represent the likelihood that demand will exceed a given value. As in the case of chronological demand profiles, monotonic load curves can be plotted for daily, weekly, monthly, seasonal, yearly or multi-yearly consumption.

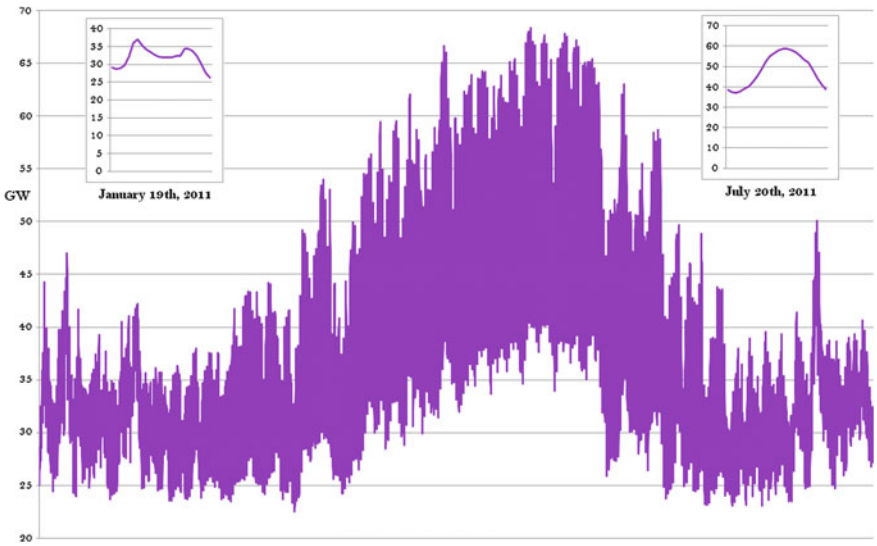


Fig. 1.5 Electricity demand yearly and daily pattern in ERCOT (2011)

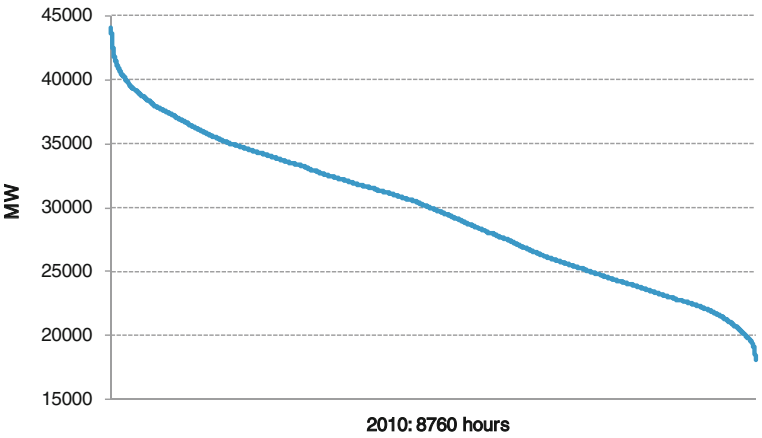


Fig. 1.6 Load duration curve for the Spanish system in 2010

In addition to the power/energy properties discussed at length in the foregoing paragraphs, consumption is characterised by other technical factors. Account must be taken, for instance, of the fact that while real power and energy are consumed in the system, reactive power is also either generated or consumed (usually the latter, since inductive motors, which consume reactive power, generally predominate). This gives rise to a power factor of less than one, which penalises consumption when this feature is included in the tariff, for it entails the circulation of non-productive current and with it ohmic losses and line capacity saturation. Moreover,

consumption may depend on supply conditions such as voltage and frequency, may be static or dynamic and may vary with connection time due to heating or other effects. All of this must be taken into consideration in load modelling.

Quality of service

Electric power consumption may be very sensitive to the technical properties of the electricity supply. Many devices malfunction or simply do not operate at all unless the voltage wave is perfectly sinusoidal and its frequency and magnitude are constant and stable over time.

The performance of electrical devices depends on the quality of the current that powers them. Problems may also arise in almost any type of electrical device when the supply voltage is too low or too high (overvoltage). Computer, motor and household appliance performance may suffer or these devices may even fail altogether if the supply voltage swings up or down.

Most electrically powered equipment, particularly expensive equipment or any regarded to be vital for the proper and safe operation of all kinds of processes, is equipped with fuses, circuit-breakers, switches or protection relays to prevent damage caused by voltage fluctuations outside an acceptable range. The motors that drive the cooling pumps in nuclear power plants, for instance, are fitted with under- and over-voltage protection that may even trip plant shutdown systems, given the vital role of these motors in safe plant operation.

Lastly, outages, whether short or long, are clearly detrimental to service quality. Who has not lost unsaved information representing hours of work on a PC because of an untimely power outage? But power failures can cause even greater harm in industries such as foundries or in chemical or mechanical processes whose interruption may entail huge losses.

In developed countries, where the universal supply of electricity is guaranteed, attention increasingly focuses on quality, as in any other commercial product. Consumption and consumers have become more demanding in this regard, and electricity industry regulation authorities assiduously include quality standards in laws and regulations. Designing the proper signals to combine efficiency with high-quality service is one of the major challenges facing the new regulatory system.

The basic factors that define electricity service quality are set out briefly below:

- *Supply outages*: supply interruptions may have serious consequences for consumers. The duration of such interruptions may be very short, in which case they are called micro-outages, often caused by the reconnection of switches after a short-circuit, or long. If they last more than a few minutes, they are known as sustained interruptions. Normally, the harm caused increases nonlinearly with the duration of the interruption.
- *Voltage drops*: momentary dips in supply voltage caused by system short-circuits or failures, lasting only until the fault is cleared, or due to the start-up of nearby motors with high input demand when first switched on, occasioning voltage drops in the supply network. Some devices are especially sensitive to these drops, particularly motors whose electromagnetic torque varies with the square of the supply voltage.

- *Voltage wave harmonics*: deviations from the fundamental frequency of the voltage sine wave due to the saturation of ferromagnetic materials (in system transformers or generators, for instance) or to the loads themselves; these deviations may also have adverse effects on consumer appliances.
- *Flicker*: low-frequency fluctuations in voltage amplitude normally due to certain types of loads. Arc furnaces and electronic devices with thyristors usually cause flicker, which is detrimental to the proper operation of devices connected to the grid. The solution to this problem is complex, since it depends not on the supplier but on system loads.
- *Overvoltage*: voltage increases caused by short-circuits, faults, lightning or any other event, potentially resulting in severe damage to consumer appliances.

Lastly, it should be added that electric power consumption may vary widely with temperature or other circumstances. What must be borne in mind is, as mentioned earlier, that this demand must be met instantaneously; therefore, the electric power supply system, including power stations and transmission and distribution grids must be designed to detect and respond immediately to such variations. The system must be equipped with sophisticated measurement, control and monitoring equipment and must have reserve generating capacity able to come on stream at all times. Yet, most users flipping switches in their homes or workplaces to turn on the lights or start up an appliance or tool are completely unaware of the host of systems, services and processes needed to provide that service.

1.3.2 Generation

The electricity required to meet these consumption needs is generated at production centres commonly called power plants or stations, where a source of primary energy is converted into electric power with clearly defined characteristics. Specifically, these facilities generate a three-phase, sinusoidal voltage system, with a strictly standardised and controlled wave frequency and amplitude. There are many generation technologies, usually associated with the fuel used. Conventional power stations are divided into hydroelectric, steam turbines—either fossil fuelled or nuclear plants—and combustion turbines, as described below.

The primary source of energy used in hydroelectric stations is water, which is expressed, energetically speaking, in terms of flow rate and height, or "head". Hydroelectric energy is converted by a hydraulic turbine into mechanical energy, characterised by the torque and speed of the shaft coupled to the electric generator. In other words, hydraulic energy is converted into electrical energy in the generator, producing voltage and current at the machine terminals. Because of the source of primary energy used, hydroelectric stations produce less atmospheric pollution than other conventional generation technologies. Another advantage of stations of this type, in addition to the cost of the fuel and lack of pollution, is their flexibility for connection, disconnection and output modification, making them

highly suitable regulating stations for adjusting production to demand needs. Nonetheless, they are costly to build, and ensuring a steady supply of water normally involves flooding vast areas. Lastly, their operation is contingent upon a highly random factor typically: rainfall in the area where they are sited.

The many types of hydroelectric stations can be grouped under three main categories, which are distinguished by system operation.

- Conventional hydroelectric stations are the most common type. Their characteristics are as described in the preceding paragraph.
- Run-of-the-river plants have no storage capacity and consequently have to use water resources as they become available; for this reason, they cannot be used as regulating stations.
- Pumping power stations have a raised reservoir to which they can pump water when electric power is cheaper, and then dump it to a turbine when it is more cost-effective to do so. They can be regarded as an efficient means of storing energy, but not electricity per se.

Fossil fuelled steam thermal power stations, in which the primary energy is provided by coal, fuel oil or gas, are respectively termed coal-fired, oil-fired or gas-fired plants. The operating principle behind these stations is basically as follows:

- the fuel is burned in a boiler to produce high-pressure steam;
- high-pressure steam is converted into mechanical energy in the steam turbine;
- mechanical energy, as in hydroelectric plants, is converted into electric power by the generator.

The thermal efficiency of fossil-fuelled steam power stations, which convert chemical energy to thermal to mechanical to electrical energy, depends primarily on the calorific value of the fuel used. In any event, the highest efficiency reached is never over 45 %. Due to the heat inertia of the boiler, up to about 7 h depending on the type of plant, these stations cannot be readily connected and disconnected or made to follow steep ramps in their output, so they are less flexible in this respect than hydroelectric plants. In light of this circumstance, start/stop studies are conducted on steam power plants to establish operating orders, and the plants are sometimes placed in standby operation, without generating any power whatsoever. Although fuel may be subject to variations in price, in most countries a constant supply is assumed to be routinely available. Therefore, such stations can be used for load following or regulation, subject to their connection inertia and the flexibility in modifying their output.

Two basic types of conventional thermal plant technologies use gas as a fuel. One type is combustion gas turbine plants where, as in jet engines, gas combustion in high pressure air feeds a turbine that produces mechanical energy, in turn absorbed by an AC generator. The other, combined cycle or CCGT (combined cycle gas turbine) plants are today's technology of choice and merit further comment. The operation of these stations, as may be inferred from their name,

involves two types of cycles. In the primary cycle, a compressor attached to the shaft of a combustion gas turbine absorbs air at atmospheric pressure, compresses it and guides it to a combustion chamber where the gas that triggers combustion is likewise injected. The resulting gas expands through the turbine blades to produce mechanical energy. The exhaust gas expelled from the combustion gas turbine, which is still at a high temperature, is used to heat a water steam circuit so that the residual heat of the exhaust gas is exploited to run a classic steam cycle, as described above. Finally, electricity is produced by one or two AC generators, connected to a single common shaft or two separate shafts, one for each cycle. Thanks to the latest advances in ceramics—the materials used to provide thermal protection for the blades—higher temperatures can be reached. Re-using very hot exhaust gases, in turn, results in substantially higher performance in these cycles than in open gas or conventional steam turbine cycles, with thermal efficiency values of up to 60 % in some facilities. This improved performance, together with a considerable reduction in polluting emissions, a high degree of modularity and reasonable investment costs, makes CCGT one of the most competitive generation technologies available.

Nuclear power plants, also known as atomic power plants, essentially consist of a nuclear reactor that produces vast amounts of heat through the atomic fission of uranium. This heat is transferred to a fluid (carbon dioxide, liquid sodium or water) and carried to a heat exchanger, where it is transferred to a water steam circuit. As in steam stations, the rest of the process involves transforming the steam produced into mechanical energy in a steam turbine and then into electric power with an AC generator. There are several drawbacks to the use of nuclear power plants, which make them socially controversial: the magnitude of the catastrophe in the event of an accident, no matter how low the risk, and the problem of eliminating radioactive waste. Another risk is related to nuclear proliferation, i.e. redirection of the technology toward nuclear weapon production or terrorist activities. Finally, these plants also entail a financial disadvantage: electric companies may have to become deeply indebted to defray the very large construction costs involved as well as the long construction time. In light of these difficulties, some countries have imposed a moratorium on the construction of new nuclear power plants. From a system operation standpoint, nuclear power stations are base-load plants, rarely used for regulation because of their comparatively low variable production cost and their operational inflexibility, although this appears to depend much on the plant design. In liberalised electricity markets the typically high margin of electricity prices over nuclear plants variable costs induces companies to run these plants continuously at full capacity. This may change in the future with the possible strong presence in some power systems of highly variable generation with zero variable production cost, such as solar or wind. Inflexible plants and intermittent low-cost generators do not mix well in the absence of large and economically viable storage capacity.

In the present electric power grids, most production currently takes place in so-called conventional stations, i.e. the sort described earlier. The share of other types of power stations in the generation mix has been gradually growing in some areas and countries, however, and has already reached very important penetration levels

in some power systems. The main advantages of these technologies—wind, solar, biomass, geothermal or wave power—are their limited environmental impact and the use of renewable sources of energy.

The renewable source that has experienced the most impressive growth in recent years is wind energy; in fact, CCGT, wind and, most recently, also solar technologies account for very nearly all new generation plants in most systems. Wind farms may be fitted with synchronous AC generators, like the ones used in other types of power stations, or asynchronous machines, which accommodate small variations in speed when the torque fluctuates, to reduce equipment wear. In wind plants with asynchronous generators, capacitors are needed to generate the reactive power consumed by the induction machinery. These stations may be connected to the grid directly or indirectly, i.e. through a rectifier, inverter and filter. With the generation of direct current, plants can work at variable speeds. This comes at a cost, however, in addition to line loss and reliability issues, although the reactive power generated can be controlled by power electronics. Further development of this technology is expected with the implementation of offshore wind generation.

The source of solar energy is hugely abundant but to date the volume of installed capacity is much lower than that of wind energy. Recently, rapid growth has been observed, however, in the installation of photovoltaic cells, which convert solar energy directly to DC current for storage in batteries or to be converted to AC current and injected to the grid. Electricity production with photovoltaic cells is basically indispensable in remote areas with difficult access to other sources of primary energy. The cost of electricity produced with photovoltaic cells, although still quite expensive when compared to conventional technologies, has declined dramatically during the past decade.

Conversely, in concentrated solar steam power stations solar radiation is used to heat a fluid and generate electricity thermodynamically. A variety of alternative solar thermal technologies are in place.

- Parabolic trough stations use parabolic reflectors to focus solar radiation on pipes and heat the receiver fluid they carry. This receiver then releases the heat to a steam turbine cycle, in stations using solar energy alone, or in a variety of cycles in hybrid configurations.
- Central receiver or concentrated solar power tower stations have a field of heliostats (sun-tracking mirrors) that focus radiation on a receiver located on the top of a tower where the heat is stored, using a receiver fluid (liquid sodium or molten salts, for example), for subsequent use in any kind of power cycle. The main advantage is that higher temperatures can be reached than with parabolic troughs.
- The solar dish configuration is similar to the central receiver design but on a smaller scale, in which each module has its own “dishes” or parabolic disks and its own receiver. These plants, whose chief advantage is their modularity, generally use Stirling or gas turbine cycles.

Biomass generation, which involves obtaining energy from biological sources, such as energy crops (also called biomass feedstock), livestock waste or forestry residue, uses resources available in nearly any habitat. That may explain why it is gaining popularity in developing countries and others such as India. This technology adopts one of two basic approaches.

- A direct combustion in specific furnaces to produce steam subsequently used in a turbine cycle, as in conventional steam power stations. This approach is often used in a CHP configuration (see next paragraph).
- An indirect process including a conversion that could be thermal or biological, transforming the organic matter into a combustible gas or liquid, generally used to feed an internal combustion engine or gas turbine coupled to an electric generator.

Finally, CHP (combined heat and power) or co-generation technology is based on the fact that many industrial plants have process heating requirements. The basic principle is to make industrial use of the surplus heat produced by some type of steam generation system instead of allowing it to simply dissipate as the return fluid cools.

The generation mix

The existence of such a wide variety of technologies in most countries can be justified in a number of ways.

First, there is a purely economic justification resulting from the shape of the load curve. The range of fixed investment costs involved in building a power station and the operating costs of generating electricity vary widely from one technology to another. Nuclear power plants, for instance, call for very high investment, but have comparatively low operating costs, due to the low price of the fuel, in this case uranium, per unit of energy output. This factor makes nuclear power an attractive technology, from the standpoint discussed here, for the segment of the demand curve that reflects needs for 8760 h in the year.

The other extreme is gas turbine technology, which has the highest operating but lowest investment costs, as well as a high level of operational flexibility, making it a very appealing type of generation to cover demand peaks, i.e. a relatively small number of hours per year, and to provide operating reserves to cope with imbalances of supply and demand in real time.

Conventional steam stations fall in between these two extremes. Obviously, the assumptions on which economic analyses are based and which justify the co-existence of different technologies always involve some degree of uncertainty in connection with factors such as the shape of the future demand curve, fuel costs, the specific operation of each generating station, capital costs, regulatory decisions and market prices (as appropriate).

Not only economic but also political and environmental considerations weigh heavily in the reasons for deploying a technology mix in electricity generation. Ensuring a supply of fuel as independent as possible from political and economic crises, be they international, such as the oil price crisis, or domestic, such as a

miners' strike, entails the implementation of a diversification strategy. Moreover, the internalisation of environmental costs and medium- and long-term environmental sustainability go hand-in-hand with regulatory measures to encourage the use of production technologies with a lower environmental impact. This is the case of generation from renewable energy sources, such as hydro, wind, solar, biomass, geothermal or a diversity of marine technologies. Most of these technologies have high investment costs but low or zero variable costs. Wind and solar plants with no storage have a highly variable and unpredictable output and they are frequently termed "intermittent generation".

Today, most electricity generation takes place in large production centres scattered around a country, often at long distances from the major consumption centres. Stations are logically built close to the source of fuel (mines and ports for coal, refineries for fuel oil, regasification plants and pipelines for gas-fired stations, rivers with a heavy flow or head for hydroelectric stations, sites endowed with high wind or solar resources) and on the coast or river banks, since water is a vital coolant in large steam plants. Efforts are usually made to site large power stations at a substantial distance from densely populated areas, due to such issues as pollution or the adverse social reaction to nuclear power plants.

The transmission grid is responsible for carrying the electricity generated in large power plants to the substations that feed the major consumption centres. The enormous size of modern electric power plants is a result of the lower unit costs obtained by increasing their dimensions. This effect, known as economies of scale, has been a factor in the decision to build nuclear plants with an installed capacity of up to 1500 MW, or 500 MW or even larger coal- or oil-fired steam stations, since they are more competitive than smaller plants using the same technologies. The advent of CCGT technology has reversed this trend; as such plants are much more modular, they can be a great deal smaller and still be competitive. Good wind and solar resources are frequently located far from the major load centres, therefore requiring important network investments. On the other hand, the next few decades may see a dramatic rise in distributed generation, with generation units located much closer to consumers, even at their very premises, supported by regulatory measures encouraging diversification, energy savings, such as in CHP, and reduction of environmental impact. Traditional demand nodes in the transmission network may become generation nodes—permanently or occasionally—when the embedded distributed generation in the corresponding local area happens to exceed the local demand.

The generation of electric power in large plants calls for particularly large investment, amortised over the very long term (25 or 30 years) after several years of construction (five, ten or even more in the case of nuclear plants or large-scale hydroelectric stations). The high financial risk that this entails can be assumed by State-owned entities or private initiative if the government provides sufficient guarantees to ensure the recovery of investment and operating costs through regulated tariffs. CCGT technology has changed the economic context substantially by significantly reducing risk: these stations are more flexible, modular and competitive, smaller and therefore quicker to build. All of the foregoing has

greatly facilitated private investment, in the wake of the regulatory changes to introduce free competition in the electricity industry that have taken place worldwide since the early 1990s.

1.3.3 Transmission

The transmission grid connects large and geographically scattered production centres to demand hubs, generally located near cities and industrial areas, maintaining the electric power system fully interconnected and in synchronous operation.

The long-distance transmission of huge amounts of power involves operating at high voltages to cut down the current intensities, therefore reducing cable sizes, voltage drops and ohmic losses in power lines.

The transmission grid is the backbone of the electric power system, interconnecting all its hubs. Its key role in the dynamic equilibrium between production and consumption determines its typically web-like structure, in which every station on the grid is backed up by all the others, to prevent the consequences of possible failures. Ideally, the system should operate as though all generation and all demand were connected to a single bus or network node. It is fitted with sophisticated measurement, protection and control equipment so overall system operation is not endangered by faults, i.e. short-circuits, lightning, dispatch errors or equipment failure.

The transmission grid has acquired particular importance in the new regulatory context that encourages competition, since it is the wholesale market facilitator, the meeting point for market players, as discussed below. The growth of transmission grid capacity, together with the development of connectivity between transmission grids, both within and across national borders, has paved the way for regional- or international-scale electricity markets.

Power lines

Transmission grid lines consist of steel-cored aluminium conductors supported by towers. Line design is based both on mechanical and electrical considerations. The towers must be sturdy enough to bear the weight of the conductors while maintaining the minimum safety distance between them—and from them to the tower and to the ground as well—imposed by the voltage. The conductors are attached to the towers by a very conspicuous assembly of insulators. Since each insulator can accommodate a voltage ranging from 12 to 18 kV, a typical 400 kV line may need from 20 to 25 such links in the insulation chain. Sometimes two lines run along a parallel route, sharing the same towers; this is known as a double circuit, an example of which is illustrated in Fig. 1.7.

Electrically, the cross-section of the conductors determines the maximum current they can transmit and therefore the line transmission capacity. The greater the current, the greater are the line energy losses due to the Joule effect. Higher

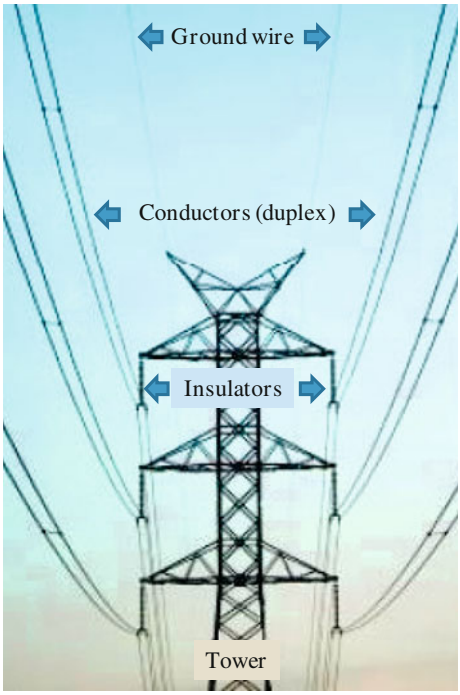


Fig. 1.7 Double circuit line

conductor temperature leads to increased conductor expansion and lengthening, resulting in a shorter distance to the ground and greater risk of discharge.

In order to reduce so-called corona discharge (rupture of the insulation capacity of the air around the conductors due to high electrical fields, occasioning line losses and electromagnetic disturbance that may cause interference in communications systems) each phase of the line is generally divided into two, three or more conductors, giving rise to duplex or triplex configurations.

One of the most important line parameters, inductance, depends largely on the relative geometric position of the three phases on the tower. Moreover, the lines induce a capacitive effect with the earth and among the conductors that establishes their capacitance to ground. The inductive effect predominates in lines under heavy loads, which consume reactive energy, whereas the capacitive effect prevails during light load periods, typically at night, when the lines generate reactive energy.

Some transmission lines run underground, mostly in urban areas where the operating voltage is lower and only rarely in the case of very high-voltage circuits. High-voltage underground systems involve the deployment of rather expensive cable technology, since the very short distance between the line and the ground requires the installation of heavy-duty insulators. These lines have a much more pronounced capacitive effect than overhead lines. Submarine cables connect island power systems to the mainland and also other power systems separated by moderate water distances.

In a meshed system like the transmission grid, energy flows are distributed across the lines according to their impedance, pursuant to Kirchhoff's laws. The long distances and the large scale of the power transmitted may reduce the grid's ability to maintain system operation, favouring the appearance of instability, detrimental to the dynamic equilibrium between generation and demand. This may reduce line transmission capacity to less than its natural thermal limit.

In some cases DC is used instead of AC technology. DC line technology requires expensive converter stations—based on power electronic technology—at both ends of the line to connect it to the prevalent AC system (any required voltage level transformation will also require expensive power electronic devices since the classical transformer device cannot operate in DC). However, this technology presents some advantages—no reactive power flows, higher transmission capacity, lower losses and lower voltage drops for the same voltage and size of the conductors, controllability of the flow, no frequency dependence, reduced stability problems—that has turned it into the chosen alternative in some specific applications such as:

- Long distance lines. The resulting costs may be reduced using DC technology since the price of the converter stations may be largely compensated by the higher transmission capacity of a DC line and the need of additional devices required to control the stability problems that may appear for such long distance in AC lines. The longer the distance the smaller the proportion of the costs of the converter stations referred to the total cost of the line.
- Submarine cables and medium-distance underground cables. The configuration of submarine cables—the conductors should be isolated from the water surrounding them—and in a lower extent the underground cables, induce an extremely high capacity effect that result in a high reactive power generation and a high associated current, that significantly reduces the available transmission capacity of the line. DC technology eliminates reactive power associated problems. Only short distances (like the Strait of Gibraltar) make sense for AC submarine cables.
- Interconnections between power systems with large different security standards. AC interconnections force a joint exposure to disturbances at both ends. DC interconnections allow uncoupling both systems from a security point of view.
- Interconnections between power systems with different frequency values. An AC interconnection is not technically feasible between two systems that operate with different frequency values—for instance Uruguay (50Hz) and Brazil (60Hz).

As noted above, environmental considerations have made it increasingly difficult to expand and reinforce the transmission system, resulting in a growing need to make optimum use of existing facilities. This represents an important challenge, since it entails narrowing security margins and perfecting protection, measurement and control logic. With the development of power electronics, new devices have become available that are designed to increase actual line capacity and steer

current flow toward the lines with the smallest load flows. Such devices are known as flexible alternating current transmission systems or FACTS.

Substations

Substations, which constitute the second fundamental component of the transmission grid, serve three chief purposes. They are line interconnection buses, transformation nodes that feed the distribution grids that reach consumers and centres where system measurement, protection, interruption and dispatch equipment are sited.

Typically, several high-voltage lines feed into the substation, which steps the voltage down and sends the resulting current over the outgoing lower voltage transmission or distribution lines. Conversely, generation substations receive the input power from power plants, raise its voltage and inject it to the transmission grid.

Materially, the substation is structured around thick bars to which the various lines connect. Circuit-breakers—opening and closing devices—ensure the connection and disconnection operations needed for dispatch, configuration changes or the isolation of failed lines or other elements.

A wide variety of substation configurations is used. The number and arrangement of busbars (single, split, double or triple-bar or ring-shaped substations, with or without transfer bars) and the number of circuit-breaker and dispatch devices per outgoing or incoming line determine the configuration type. Increasing the number of such devices raises substation costs but enhances safety, preventing such anomalies as momentary downstream outages due to simple connection operations.

The most representative facility in substations is the power transformer, which raises or lowers the voltage. Transformation is performed electromagnetically through two sets of coils, one at high and the other at low voltage, wound around a ferromagnetic core. The entire assembly is immersed in a vat of oil to ensure optimum conductor insulation. These are very large-scale, expensive, heavy, high-performance facilities, with a very low failure rate.

Many power transformers are involved in system voltage control. In these, the windings are fitted with taps that allow for slight modifications in the turns ratio and therefore voltage step-up or step-down. In some transformers, regulation can be performed during operation, while in others it may not. The photograph in Fig. 1.8 depicts several substation transformers.

Other substation components include line breaker and switching devices. As noted earlier, substations are the interconnection buses on the grid, where the connections between the various elements are made or severed. This function, which is natural and foreseeable in normal operation, is crucial in the event of failures. Indeed, the system must necessarily be protected from short-circuits occurring in lines or substation bars, since they trigger the circulation of very strong currents that could damage cables and equipment. A fault must then be cleared, i.e. the overcurrent cancelled, as soon as possible and isolated to repair the damaged component; otherwise, the system as a whole may be endangered.

The most sophisticated line breakers are automatic circuit-breakers, which are able to open a circuit when overcurrents occur. The protection devices detect

Fig. 1.8 Substation voltage transformers

overcurrents and, applying appropriate logic, decide which lines must be opened to clear the fault. Many types of such breakers have been designed, ranging from compressed air (or pneumatic) or magnetic blow-out breakers, for low power and voltage, to oil-immersed circuit-breakers or sulphur hexafluoride (SF_6) devices, for systems with very high voltage and capacity. One special feature of these mechanisms is their ability to open twice in immediate succession. Since many faults have a very short duration, because the cause of the outage disappears spontaneously (where due to a faulty contact or a contact that is burnt out by the current flow, for instance), the system usually attempts to reconnect the circuit-breaker automatically, in case the cause of the fault has in fact been eliminated. If not, the breaker will re-open. Breaker design is such, however, that their position (open or closed) cannot be ascertained by mere visual inspection.

Once the fault is cleared and identified, the damaged area must be electrically isolated to reconnect the rest of the elements initially shut down by the circuit-breaker. This is done with local disconnectors, used to open or close a line when the current is negligible. Their function, therefore, is not to cut off the current, but simply to visibly isolate a section of line or a device, machine, substation bar or any other element so it can be handled for repair or maintenance in full confidence that it is not live. The operator opens the circuit manually after confirming that the circuit-breaker has worked properly and removed voltage from the entire area in question. Several types of disconnectors can be used: rotary, sliding, column rotary and pantographic.

Lastly, the line breakers used in grid dispatching have a break capacity of the order of the nominal intensity of the current in the circuit or line they are designed to open or close. Consequently, they do not open in the event of a short-circuit. Air-break switches, automatic air switches, automatic gas circuit-breakers, magnetic blow-out switches and oil or hexafluoride switches are some of the protection devices used for this purpose.

Today some substations are entirely immersed in hexafluoride. Although more expensive, this arrangement considerably shortens the distance between bars, conductors and cables, and this is particularly attractive for urban environments where square footage is costly. Such substations are, moreover, extremely safe. Most are sited underground.

1.3.4 Distribution

Lower voltage networks branch off the high-voltage grid from the substations in multiple directions to carry electric power to even the most secluded areas. The structure of this system, generically called the distribution grid or network, is very different from the transmission grid structure.

The upper or regional level, which actually forms part of the transmission grid, has an open web or loop configuration and operates at somewhat lower but still very high voltages, typically 132, 66 and 45 kV. The substations fed by this part of the grid step the voltage down to 20, 15 or 6.6 kV, splitting power off into the distribution grid per se, which is the part of the system that supplies power to the end consumer.

The structure of this medium-voltage network may vary, but its operation is always radial. Substations normally house circuit-breakers that protect the so-called feeders, i.e. lines running to other transformer stations where the voltage is stepped down again to supply low-voltage power. The low-voltage lines, which may be 400, 230, 130 or 110 V, depending on the country, supply residential customers, wholesalers and retailers and the like. Consumers connect into the system at the voltage level best suited to the scale of their consumption.

Distribution grids in urban areas, where high load densities are concentrated in small areas, generally run underground. While it is more costly to lay and repair underground cables, the distances involved are much shorter than in rural networks. Urban systems usually have a meshed structure for greater reliability, but by and large these networks operate radially, with circuit-breakers normally open for ease of operation and to reduce eventual short-circuit currents. The reliability level is typically high, due to the short distances, the undergrounding and the meshed network configuration that provides redundancy.

In rural areas, distribution networks are generally radial and consist of less expensive overhead lines because load density is not high and space is not an issue. There is less redundancy in the supply to any given load and the reliability levels are typically lower than in urban networks. One problem encountered is that reliability declines as the distance from the substation increases. For this reason, rural networks are sometimes designed to provide downstream emergency supply in the event of failures. The voltage drop problems that also arise in these networks are solved by placing taps on the transformers and installing capacitor banks that supply reactive power.

Distribution networks, which comprise thousands of kilometres of wires, are subject to more frequent failure than the transmission grid and their structure is less redundant; this means that most of the supply outages that affect the end consumer originate in the distribution network. In terms of investment plus operation and maintenance costs, distribution accounts for a large share of total system costs, which normally is several times higher than the share of the transmission grid.

1.4 Electric Power System Management

Electric power system management is an extremely complex undertaking, due to the breadth of the task, which covers technical, economic, regulatory, social, business and environmental factors. It broadly encompasses investment planning and system operation, both closely conditioned by the technological characteristics of the power system.

The planning and operation of electric power systems are the result of a complex chain of decisions. The first link comprises long-term provisioning (capacity expansion, long-term fuel contracts); the second, medium-term planning (hydroelectric management, facility maintenance scheduling, shorter term fuel contracts); the third, short-term unit commitment (generating unit connection, reserve operating capacity); and the fourth, actual system operation (generating unit dispatching, frequency regulation, response to possible emergency situations). This decision-making process is supported by computer models, fed by highly sophisticated data acquisition and communications systems. Today's computational resources make it possible, for instance, to precisely calculate the marginal cost of meeting demand (i.e. the cost of one additional kWh) at any given point in the system at a given time, taking into account the chain of decisions outlined above.

Decisions affecting electric power system expansion and operation should be guided by criteria of economic efficiency in order to minimise the cost of delivering an acceptable quality of supply to end consumers. Nonetheless, constant account must be taken of technical considerations to ensure the material feasibility of supplying electric power, which is arguably much more vital to this industry than to any other one, given its specific characteristics. As discussed below, the importance of such considerations grows as the time lapsing between decision-making and implementation narrows down to real time, when the distinction between economic and technical factors blurs and no sharp line can be drawn between them.

Given the size, dimension and complexity of the problem, the entire decision chain must be rationalised and organised. This is achieved by ranking expansion and operating functions chronologically. In longer term decisions, for example, where future uncertainty and economic criteria carry considerable weight, a rough approximation of system technological behaviour suffices. Such decisions successively guide shorter term decision-making, in which the technical specifics

carry much greater weight, culminating in real-time operation, where system dynamics must be analysed in full detail, millisecond by millisecond.[3]

The criterion underlying the entire decision-making process is the maximisation of social welfare in the production and consumption of electric power. This involves two fundamental concerns. The first is the effort to minimise the entire chain of costs incurred in providing the service, including both investment and operating costs. The attainment of an inexpensive service is not the only factor used to measure social utility, however. The quality of supply must also be satisfactory. Social utility would be low for both industrial and residential consumers if service were cheap but plagued by constant outages. Due to a number of uncertainties, including rainfall, real demand growth, generating, transmission and distribution equipment failures or lack of wind or solar input, outage-free service cannot be guaranteed in future scenarios. In other words, some likelihood will always exist of failure to service all of the demand at all times, which is a measure of system reliability. It is equally clear, however, that such likelihood of failure can be reduced by investing in more facilities and operating more conservatively. Increased reliability entails higher costs.

For this reason, the first criterion referred to above, cost minimisation, must be qualified to accommodate the second criterion, which reflects system reliability. This can be built into the decision-making process in a variety of ways. One is to set a minimum reliability threshold based on past experience and social perception of the concept, measured in terms of the likelihood of an interruption of electric power supply or some other similar indicator. A more soundly based method consists of attempting to quantify the financial harm caused by service interruptions on the basis of consumer utility. Building this factor into the cost minimisation process as one more expense to be taken into account is actually a way of maximising the social utility of the service. The inherent difficulty in this second approach lies in quantifying "utility", which may vary from one consumer or individual to another, and which cannot be clearly measured, although efforts have been made in this respect by conducting systematic consumer surveys specifically designed for this purpose.

Reliability is a factor that involves the entire decision-making process, from long to short term. Service interruptions may be due to investment-related issues, for example, if system installed capacity is insufficient to meet demand, perhaps because demand growth has been unexpectedly steep, hydrological conditions particularly adverse, transmission capacity lacking or the implementation of new investment delayed. They may also be caused by operating inefficiencies, such as poor reservoir management, lack of immediate response to failures in units or lines due to a shortage of reserve capacity, or by real-time system stability problems. Therefore, in nearly all the decision-making scenarios reviewed below, cost and

[3] The anticipated—sometimes already existing—massive presence of intermittent generation, demand response, electric vehicles or distributed storage, creates a significant impact of the short-term behaviour of the power system on capacity expansion planning. On-going research tries to introduce short-term modelling capabilities into long-term capacity expansion or strategic analysis.

reliability factors must be brought into balance in the decision reached. The term adequacy is generally used to describe long-term reliability, while security refers to short-term operations.

There is no standard way to organise electric power system planning and operation. The solution to such a complex problem nearly always involves breaking it down into simpler parts and all the approaches taken to date involve, in one way or another, time-scaled hierarchical decomposition. Decision-making is ranked by time frame, and the respective functions are scaled accordingly. The highest level comprises the longest term decisions, which tend to address strategic problems. The solutions adopted are then passed on to the lower levels, delimiting their scope of action. At successively lower levels of the scale, as functions approach real time, the optimum solution must be sought within the restrictions imposed by the problem and by the guidelines handed down from the higher levels. This structure, abstractly defined in this section, is described in greater detail in the items below.[4]

1.4.1 Two Power System Management Paradigms: Centralised Planning Versus Decentralised Market-Based

All the previously introduced functions must be addressed within the regulatory and legal context prevailing in each country: what has implications on how these activities are conducted and who the decision-makers are. With the profound regulatory changes undergone in the power industry in most of the world, any attempt to describe the economic environment should cover both the traditional scenario, still in place in many countries, as well as more liberalised situations. Each country handles the organisation of its power system according to its own characteristics, but, in general terms, two major regulatory approaches to maximise the social welfare can be distinguished. They are described with a variety of terms: traditional versus liberalised, centralised versus decentralised, centrally planned versus free market-oriented or regulated versus deregulated,[5] among others.

In order to provide a first pass at how decision-making happens in a power system and how it depends on the underlying regulatory regime, the discussion in the next sections will be first based on the general assumption of a single

[4] As mentioned before, this classical viewpoint is becoming increasingly challenged by the new developments in power systems: distributed generation, active demand response, distributed storage or electric vehicles.

[5] The term "deregulated" is a misnomer, since restructured and liberalised power systems are typically subject to a very detailed and complex regulation. The regulatory change that has taken place in the power sector of many countries during the last two decades is more a restructuring and re-regulation than a deregulation. The term "deregulation" is used to mean a departure from the heavily regulated monopolies under traditional regulation.

decision-maker in a vertically integrated utility under the traditional regulatory framework of a regulated monopoly with cost of service remuneration: the traditional centrally managed approach. Then, at the end of each section, the most salient differences that restructuring and liberalisation have introduced in the decision-making process of electricity companies will be succinctly described: the market-based decentralised approach.

Let us start by briefly describing the main characteristics of both approaches.

The traditional centrally managed approach

Under the traditional regulation scheme, electric utilities are awarded territorial franchises where they supply electricity with a regulated monopoly business model. The utilities are vertically integrated, i.e. they own and operate all the generation and network assets in their franchised territories, they have the obligation to supply any requested demand by the consumers in their respective areas; and they plan and implement the expansion of production and network capacity under the guidelines and authorisation of the regulatory authority. Centralised planning of operation and investment must seek the minimisation of costs—which in principle must include the consumer costs of loss of electricity supply—and a satisfactory reliability level (sometimes capacity expansion planners or system operators impose reliability-based constraints) and environmental impact. These utilities may reach voluntary agreements to exchange electricity with the neighbouring ones. In many countries a single publicly owned utility has supplied the entire demand within the national territory.[6]

Under traditional regulation the consumers are mandatorily assigned to the utility serving the area where they are located. Since the consumers have no option to choose the supplier, they are protected by the regulator, which may establish minimum standards of quality of service and also determines the remuneration of the utilities. The remuneration is based on the incurred cost of service, which includes a reasonably attractive rate of return on the invested capital. Diverse efficiency incentives have been used in some cases.

The market-based decentralised approach

Electricity industry restructuring has brought profound change to electric power system operating and planning. Industry liberalisation has gone hand-in-hand with the dramatic decentralisation of these two tasks. The electricity industry's new structure is based on the introduction of competition and the institution of electricity markets [1, 9, 10]. Wholesale markets for electric power open to all generators, both incumbents and new entrants, as well as to all consumer entities, have been established in countries adopting the new model. The core of this wholesale market is typically a spot market for electricity, which serves as a reference for

[6] In Europe, for instance, prior to the 1990s' liberalisation process, most major countries (except for Germany and Spain) had just a single publicly owned electric utility.

medium- and long-term contracts of different types, and even for organised markets for electricity derivatives. The agents trading on such markets are generators, eligible consumers and different categories of supplier companies, acting on behalf of non-eligible consumer groups or eligible consumers, or simply as strict intermediaries. Consumers are clients who are free to choose the supplier on the basis of the available commercial offers.

Account must be taken of the fact that the supply of electric power under competitive arrangements is subject to the existence of certain activities associated essentially with transmission and distribution grids, whose control entails absolute power over the electricity market. Moreover, these networks do not have the technical and economic characteristics that would permit them to provide their services under a market-based regulatory regime. Consequently, these grid-associated activities must be wholly independent of competitive businesses, namely generation and supply or retailing. For these reasons, industry organisation and its ownership structure nearly always have to be modified (unbundling process) before competition mechanisms can be implemented.[7]

Under this decentralised approach, system operation and expansion are the result of individual company decisions based on the maximisation of business earnings, either under organised tendering or through private electric power supply contracts. Economic and financial risk and anticipated returns on investment, instead of the traditional cost minimisation criteria, drive decision-making. The challenge for administrative and regulatory authorities is to design liberalised market rules that ensure that the strictly entrepreneurial behaviour of each market player leads to maximisation of social welfare—i.e. overall minimisation of system costs while respecting reliability and environmental objectives and constraints—reflected in the prices charged to final consumers.

Real-time system operation, however, continues to be a centralised task. A central operator, usually called the System Operator, oversees system safety. Ensuring real-time technical viability of the system calls for highly sophisticated co-ordination of all the available resources. This in turn requires absolute independence from the various actors' individual interests.

In this scenario electric power system operation is viewed, therefore, from a wholly different vantage point. New functions, responsibilities and ways to address the decision-making process have arisen, along with changes in the roles played by each of the agents involved.

[7] In practice, in most liberalised systems transmission grid activities—but not the distribution ones—have been wholly unbundled. In the latter, the companies involved have been legally separated but may be owned by the same holding company as generation and retail firms.

1.4.2 Planning and Investment

Traditional centrally managed approach

The first-level decision-making takes a long-term approach, projecting anywhere from 2 or 3 to 10, 15 or more years into the future, to define generating plant and transmission/distribution grid investments. The process involves determining the type, dimensions and timing of new generation and network facilities to be installed, based on several factors. These include demand growth forecasts, technical alternatives and costs, estimated fuel availability and price trends, reliability criteria adopted, environmental impact constraints, diversification policies and objectives relating to dependence on the foreign sector. Such distant horizons are necessary because the (very large) investments involved are justified on the grounds of the earnings over the service life of such facilities, which may be from 25 to 40 years in the case of steam power stations and much longer for hydroelectric plants.

With such distant horizons, uncertainty is obviously a key determining factor. A whole suite of scenarios must be addressed, the respective probabilistic assessments conducted and the most suitable criteria adopted, such as minimisation of expected average costs, minimisation of regret or minimisation of risk, taken as the variance of the cost distribution function.

For the same reason, it makes no sense in these types of studies to evaluate the technical behaviour of system operation in detail, since it is neither feasible nor sensible to seek accuracy in the assessment of operating costs when the process involves much greater levels of financial uncertainty.

One of the chief requisites for the process is a good database, which must contain information such as updated data on technologies as well as historical series on demand, hydrology (rainfall), equipment failure rates and so on. The long-term demand forecast built from these data adopts the form of a probability curve that determines system expansion needs. As indicated above, the hourly demand profile is as important as total demand in this regard, since the choice of technologies largely depends on this information. The next step is to determine the generating plant expansion required to meet demand under terms that seek to minimise the anticipated costs over the entire period considered while taking the aforementioned strategic criteria into consideration. Such costs include the fixed investment outlays plus operating costs throughout the entire period, which obviously depend on the type of investments made. Simulation and optimisation models are often deployed as an aid in such estimates. Because of the scale of the problem posed, decomposition analysis techniques are generally used that run iteratively and alternatively between two modules, one specialising in expansion cost calculations and the other in operating costs, exchanging information between the two as necessary until the results converge.

Decisions to enlarge the transmission grid have traditionally depended on new generation plant investment needs and demand growth. This is because considerably less investment and time were needed to build grid facilities than generating stations. The situation has reversed now, with the short installation times of wind,

solar and CCGT plants, while transmission facilities have to face increasing siting difficulties. Ideally, once the new plant sites and the consumption and production growth rates are estimated, grid expansion is determined, comparing the necessary investment costs with the benefits accruing to the system: lower operating costs, smaller system losses and greater reliability in meeting demand. The decision process is also affected by security criteria that ensure that supply will not be subject to interruptions due to grid reinforcements or other technical aspects, such as voltage or stability issues. In reality transmission planners sometimes have to initiate the construction process based on their best estimates of generation and demand growth. Expansion decisions are, naturally, dynamic over time, since they must be periodically adjusted when real data on demand growth, technological innovations or fuel purchase terms modify the assumptions underlying the initial expansion plans.

Distribution network planning has the primary objective of connecting all consumers and distributed generators to the grid, while meeting prescribed reliability targets. The remuneration of this network activity is typically determined by the regulator, on the basis of the estimated cost of service, plus incentives related to efficient performance, reduction of network losses and achievement of quality of service objectives. Both the planning and regulation tasks will become more difficult with the increasing presence of distributed generation, local distributed storage, electric vehicles and demand response.

Market-based decentralised approach

Restructuring and liberalisation have completely revolutionised generation expansion planning, since it is now left up to each potential investor the evaluation and the final decision on the advisability of investing in new generating facilities, based on individual cost–benefit analysis.

Each new investment is studied from the perspective of its estimated net profit over a period of several years, which strongly depends on the behaviour of the wholesale electricity market: fuel prices, demand levels, other new investments, market prices and the like. Financial risk assessment plays a crucial role in such reviews because it is the key to obtaining adequate financing. Risk management and the formulation of contracts to supply power and to hedge against uncertainties, along with access to futures markets and options, are elements of great importance in this new context.

Whilst transmission grid planning continues to be centralised, the perspective has changed and today's decisions are subject to much greater uncertainty. The basic planning criterion, namely optimisation of the social utility of electricity production and consumption, remains unvaried. But such utility no longer necessarily adopts the form of minimisation of production costs, but is based, rather, on the maximisation of the aggregated benefits (surpluses) of all the individual actors: the utility of electricity consumption less acquisition cost, for consumers, and the revenues from the sale of electricity less generating costs, for producers.

Moreover, the uncertainty surrounding the transmission planner's decisions has risen dramatically. Traditionally, the grid was planned on the basis of prior results

of the expansion of generating capacity, information which in an environment of free competition is not compiled centrally or a priori, but rather is the result of business decisions made individually by the different players at any given time.

Consequently, neither the amount nor the location of new generating units is known for certain when the grid is being planned. Furthermore, the time that may lapse from when the decision to build a line is made until it becomes operational is growing longer, due in particular to difficulties in permitting regarding environmental impacts and territorial organisation. Together, these two factors render transmission network investment planning very complex. In addition, new generation plants should be carefully sited, taking current network capacity into consideration because construction lead times are often longer for the transmission grid than for the power stations themselves.

Transmission investment has an impact on both generation plant competitiveness and market power. The former complicates decision-making. Because of the trend to create supra-national markets and the appearance of large-scale renewable production facilities, transmission planning has become a critical issue whose geographic scope has been substantially enlarged.

Distributed generation, electric vehicles, storage and demand response may take place within both the traditional and liberalised frameworks. Their implication in the functioning of the power system may be very different depending on the choice of regulatory paradigm.

1.4.3 Operation Planning

Traditional centrally managed approach

Once the installed capacity of generation and network is defined, the medium and long-term operating plans for the production facilities must be established. For a one- to three-year horizon, depending on the system in question, such planning involves determining the best unit and grid maintenance cycle programme, the most beneficial fuel purchase policy and the most efficient use of power plants, subject to primary energy limitations—hydroelectric stations in particular—to the availability of wind and solar resources or to yearly production restrictions for environmental reasons.

Electricity generating plants are sophisticated systems with thousands of components that must be checked periodically to prevent major and sometimes hazardous failure and ensure plant efficiency from a technical standpoint. Operation of conventional steam plants is usually interrupted around 20 days a year for this purpose. Nuclear plants need to have their fuel (uranium bars) recharged once every year and a half approximately, so maintenance tasks are programmed to concur with such plant shut-downs. Electric power lines and transmission and distribution grid components located in substations also need upkeep, such as the replacement of faulty insulators or their cleaning to prevent loss of insulation power. Although the technology to perform these tasks on live lines is becoming

more widely available, most of these operations are conducted on de-energised facilities for obvious safety reasons, which involves disconnecting particular substation lines or areas. This, in turn, requires careful maintenance planning to interfere as little as possible with system operation.

Fuel management also calls for careful planning, sometimes far in advance. Once input needs are defined, fuel purchases, often on international markets, must be planned to buy at the most advantageous price, make shipping and storage arrangements and take any necessary logistical steps to ensure that stations do not run out of fuel.

Finally, the use of water in hydroelectric plants must also be planned as if it, too, were a fuel. In fact, water can be regarded as a cost-free fuel in limited supply. Therefore, its use must be scheduled in the most beneficial manner. Run-of-the-river stations require no planning, but decisions are imperative when there is an option to either generate energy or store water for production at a later time; therefore, while ideally free, water actually has an opportunity cost. Depending on the size of the reservoir, such decisions may cover time frames ranging anywhere from a single day to several weeks, months or even years for the largest reservoirs. The operation of stations whose management and regulation extend over several months should be planned on a yearly or multi-yearly basis. Since the logical aim of such planning is to attempt to replace the most expensive thermal production, this type of planning is usually termed hydro-thermal co-ordination. A similar approach is taken for any other technology subject to restrictions on use that limit production in a given time frame, typically seasonal or annual. One example would be the existence of mandatory national fuel consumption quotas or yearly pollution limits.

Market-based decentralised approach

Under market conditions the objective of the plant's owner is to maximise the margin of revenues over production costs. Not surprisingly, the two considered regulatory frameworks ideally lead to the same operation schedule for the generating plants. The various agents—producers and consumers—participating in the electricity market must attempt, under the liberalised scheme, to optimise their medium- and short-term individual operational decisions. Generators, in particular, must perform the following tasks:

- To formulate medium-term economic forecasts: revenue projections and yearly budgets.
- To provide support for the aforementioned long-term functions: contract management and determination of long-term business strategies.
- To provide support for short-term functions, in particular the formulation of bids in the daily and ancillary services markets, by making explicit the boundary conditions for these short-term decisions: available hydro resources or fuel quotas for the considered period, environmental restrictions, fuel costs and scheduled availability status of conventional plants.

The decision support models for generators competing in electricity markets must seek the optimisation of the net profit of each agent. This requires making use of theoretical microeconomic- and game theory-based concepts into the model. In the former case, markets are seen as dynamic elements that reach certain points of equilibrium characterised by the various agents' production structures. In the latter, markets are viewed as the result of a game in which the established rules ultimately impose a strategy to be followed by each agent in response to the reactions of all the others.

1.4.4 Unit Commitment and Dispatch

Traditional centrally managed approach

Short term decision-making typically refers to a weekly horizon, at most, i.e. from one or a few days up to one week. It involves determining the production plan for hydroelectric and steam power stations on an hourly basis for each day of the week. This plan must also abide by the instructions received from the immediately higher decision level described in the preceding section, in connection with factors such as maintenance work, weekly hydroelectric management, emissions plans or fuel quota management.

At this level, system details are extremely important, and account must be taken of aspects such as steam plant generating unit start-up and shut-down processes and costs, wind and solar forecasts, hydrological constraints in river basins, stations in tandem arrangement, demand chronology profiles (which call for accurate production monitoring) and generating capacity to be held in reserve to respond immediately to fortuitous equipment failure.

The ability to vary steam power station output is limited by the technical characteristics of the respective generating units. It takes an inactive station a certain minimum amount of time to recover operational status, which is primarily determined by the time needed to heat the boiler to a suitable temperature. Therefore, this minimum lead time depends on how cold the boiler is, or in other words, the amount of time it has been turned off. Conventional coal fuelled steam power plants may need from 8 to 10 h if the boiler is completely cold. Gas fuelled plants are more flexible, with lead times of a few hours for CCGTs or even only a few minutes for single cycle gas turbines. In addition to this, one must include the impact on the maintenance costs of each start-up process, a value that is typically specified in the maintenance contracts for each individual plant. As a result, the cost of starting up a steam power station may be significant and can be quantified as the price of the fuel that must be unproductively burnt to heat the boiler to the appropriate temperature plus any associated effect on the maintenance costs. For this reason, even if demand declines substantially, it may not be cost-effective to disconnect certain steam stations at night, but rather to maintain a minimum production level. This level, known as the plant's minimum load, is generally relatively high due to boiler combustion stability requirements, on the order of

30–40 % of the station's maximum output. Depending on the findings of cost-effectiveness studies, a decision must be made on whether it is more economically sound to start and stop the station every day (daily start-up cycle) or to shut it down on weekends only (weekly cycle), or simply never shut it down, as in the case of nuclear stations. Sometimes it may be more efficient to keep the boiler hot but unproductive. Boiler thermal constants and their limits likewise impose constraints on how quickly steam plant output rates can be modified, which are known as upward or downward ramp constraints. Therefore, careful planning of unit start-ups and shut-downs is required, a problem known as unit commitment.

This decision is also strongly influenced by daily or weekly hydroelectric management, as well as system reserve capacity requirements. Hydroelectric stations are much more flexible than thermal power plants, with practically zero lead time, no significant start-up costs and virtually no real limits on modulating generating capacity. The optimal scheduling of hydroelectric production to meet variations in demand takes a number of considerations into account: higher level decisions on the amount of water resources to be used in the day or week, the most cost-effective hourly distribution (weekly or monthly hydro-thermal coordination) and technical constraints on steam plant units. Hydroelectric generation and its use of water in reservoirs must likewise comply with possible restrictions imposed by water management for other purposes, including irrigation, fauna, and minimum reservoir and river flow levels, as well as other conditioning factors typical of water works and their configuration: canals, pipelines, reservoir limits and reservoirs operated in tandem.

Here also, reliability criteria play a role in decision-making. Provision must be made for the immediate replacement of any plant in the system that may reasonably be expected to fail or for the ability to respond to transmission grid failures. This translates into the start-up and connection of new units; although these may be unnecessary under normal conditions, the system would not otherwise be able to generate power for several hours if needed to cover emergencies.

Market-based decentralised approach

Although electricity markets may be organised in a variety of ways, a series of short-term organised markets usually co-exist with private bilateral agreements between producers and consumers. Generating stations' short-term output is determined from both market clearing and bilateral arrangements. Organised markets, typically including a day-ahead market, several intraday markets, a balancing market and an ancillary service market, play a central role as they provide a transparent and publicly available price reference for energy and reserve services.

The day-ahead market is usually the most important market in terms of trading volume. Consequently, a substantial share of a company's short-term operations tends to revolve around the preparation of daily bids. This process is governed by higher level and longer term strategic decisions. As such, it is informed by output guidelines deriving from analyses covering longer terms and its role consists of setting market prices on a day-to-day basis. In some power systems (those using what is called "complex bids", see Chap. 7) all costs are declared in the bids.

Other markets use "simple bids", just specifying quantity and price; in this case the generators need to estimate the future short-term prices and the dispatch pattern of their plants, as a guide for decisions on the internalisation of steam generating unit costs and the utilisation of their hydro resources. These are complex decisions that are generally supported by sophisticated market simulation and optimisation models.

1.4.5 Real-Time Operation

Traditional centrally managed approach

Security criteria acquire an increasing importance, even dominating economic criteria, in the design and implementation of real-time operating functions. The economic component of these processes is defined by higher level decisions, where the economic aspects of reliability should always be kept in mind. Supervision, control and monitoring ensure the technical viability of the immense and dynamic electric power system, as described above. The concerned functions and facilities are organised by levels or layers.

On the first level, the components that comprise the system backbone, i.e. generating stations, high-voltage grids and large substations, are centrally monitored and controlled from a control centre that supervises system status in real time; the generating plant, line flows, voltage levels, system frequency and the like are checked by remotely transmitted and duly processed measurements.

This supervisory and control system goes by the name of SCADA, acronym for supervisory control and data acquisition. These control centres, of which there may be one for the entire country, or several, scaled by order of importance and coordinated, are intended to ensure system security and may transmit instructions to generating stations to produce real or reactive power, order grid dispatching operations, change transformer taps or connect capacitor banks. Such instructions are based on system data, interpreted by operators on the grounds of their experience or with the support of sophisticated models that analyse operating conditions and determine line flows or bus voltages under various hypothetical system contingencies.

The control systems installed in production plants constitute the second level of operation. The two most important such systems are speed and voltage regulators.

Speed regulators maintain the instantaneous balance between generation and consumption in the system as a whole. The generating plant must respond immediately to any increase or decrease in demand. Similarly, the accidental tripping of a unit in operation at any given time (where nuclear power is involved, this may mean up to 1500 MW) induces an instantaneous imbalance between power generated and consumed that must be compensated for by immediately replacing the failed unit. When the power generated differs from the system load, any surplus power is stored or any shortage is drawn from the kinetic energy stored

in rotating machines. Speeding up or slowing down these facilities changes the revolutions per minute in the AC generators or the frequency of the generated voltage wave. Such parameter changes automatically activate the respective steam-, water- or gas-driven valve to modify plant generation accordingly. This is called *primary regulation* or *load–frequency control*.

A power shortage caused by a power station failure, for instance, prompts a joint response across the entire inter-connected system, i.e. all the independently dispatched power systems, synchronously connected to the power system where the shortage occurred, are involved. This arrangement prevents system frequency from falling further, but is unable to re-establish it exactly to the nominal value. Nor are the predetermined values of power exchanges with neighbouring systems sustained, due to the flows required to maintain the frequency. A second control loop, known as *AGC or automatic generation control*, re-establishes frequency to the nominal value and the exchange flows to their initial levels. This constitutes what is known as *secondary regulation*, which is also usually automatic, and does not involve all generators, in particular none located in neighbouring systems. The extra generation required is redistributed among the stations chosen for this purpose according to pre-computed factors resulting from economic considerations. This also regenerates the primary reserve capacity, to ensure continued operation and prevent system standstill resulting from units reaching their capacity limit.

Lastly, *tertiary regulation* may also be implemented. At this level, in which supervision is not automatic, the control centre may change long-term dispatching instructions to enhance economic efficiency and restore the so-called secondary reserve capacity, in much the same way that secondary control restores the primary reserve capacity. Secondary and tertiary regulation form part of the higher control level referred to earlier, but they have been described here for greater clarity[8].

In a centralised environment, the economic aspects at this level are mainly taken into account by economic dispatch according to the variable costs of the scheduled units and recomputed every few minutes. The system operator, taking into account the technical and economic characteristics of the available plants, schedules the provision of operating reserves to maintain system security. No special payments are necessary to obtain these services from generators under a cost-of-service remuneration that covers their fixed and variable costs.

Power stations are equipped with a second control loop relating to system voltage. System voltage must be kept within certain allowable margins to ensure system security and guarantee that the power delivered is of reasonable quality. The voltage level of an electric power system is closely related to the balance of reactive power. High reactive consumption, either by live lines or inductive motors, tends to depress system voltages, whereas a supply of reactive power, from

[8] The terminology that is used to classify frequency-related reserves varies much among power systems because of legacy reasons and also because of the diversity of generation mixes and reserve requirements. See Chap. 7 for a larger description.

non-live lines or capacitor banks, for instance, tends to raise system voltages. For these reasons, power stations, which are able to produce or consume reactive power at will with their (synchronous) AC generators, are ideal candidates to monitor and correct dangerous voltage fluctuations. The voltage regulator measures voltages at generator terminals or selected points of the system, compares the measurement to a reference value and adjusts the AC generator excitation current accordingly, which controls the reactive power supplied or absorbed by the unit.

Power stations are naturally fitted with protection systems that prevent potential damage. The AC generator, pumps, turbines and all other vital components are equipped with the respective measuring systems, tripping relays and alarms. The approach is as discussed above for substations: the protection relays must detect and locate faults, the automatic circuit-breakers clear them and the disconnectors isolate the failure so that service can be re-established in the rest of the system while the fault is repaired. Protection relays must be sensitive enough to detect the fault, selective to minimise the impact of clearance, able to respond quickly for protection to be effective and reliable, i.e. neither tripping facilities unnecessarily nor failing to act in critical situations. They must also be robust, since they operate under widely varying adverse circumstances, and able to operate independently and automatically, even in the absence of electricity.

Market-based decentralised approach

As under the traditional organisational approach, real-time operation is strongly influenced by security considerations and has a similar structure, although considerable effort is generally made to differentiate and clearly value the various types of so-called ancillary services provided by each agent in this respect.

Economic dispatch is covered by the outcome of the day-ahead market, backed by shorter term markets (intra-day and balance markets). Deregulation has also caused several ancillary services to surface. These services are needed to ensure security, quality and efficiency of supply and were traditionally rendered by generation, but involve more than mere power production. Under competitive market conditions, ancillary services can be only provided if they are paid for or mandated. Whenever possible, the System Operator, as the entity responsible for ensuring system security, uses market mechanisms to competitively decide who provides what service and at what price. Substantial efforts have been made to define the ancillary services and the corresponding markets. Typical ancillary services include load-frequency regulation, i.e. secondary and tertiary regulation and load shedding, but also voltage control and even black-start capability (from blackouts). Primary load-frequency control is normally considered a basic mandatory service, at the disposal of the System Operator as a key element to guarantee system security.

In this competitive framework, companies must offer their services on the grounds of the costs incurred in their power stations to provide them, as well as of other considerations such as market opportunities.

1.5 Environmental Impact

Environmental concerns have become a factor of equivalent weight as security and economic considerations in the planning and operation processes of the electric power industry. There is widespread belief that one of the major challenges facing humanity today is the design of a model for sustainable development, defined as development that meets the needs of the present without compromising the ability of future generations to meet their own needs. Besides such weighty issues as the enormous social and economic inequalities between peoples or the existence of a growth model that can hardly be extended to the entire world population, other issues directly related to electric power systems, such as the intense use of known energy resources and their adverse impact on the environment, also come under the umbrella of sustainable development. For these reasons, environmental impact is a factor of increasing importance that conditions the present operation and development of these systems and will indisputably have an even more pronounced effect on the industry in the future.

Generation is arguably the line of business in electric power systems that produces the greatest environmental impact, in particular with regard to thermal plant emissions and the production of moderately and highly radioactive waste. As far as combustion is concerned, coal- and oil-fired steam plants compete with the transport industry for first place in the emission of carbon dioxide (CO_2) associated with greenhouse gas-induced climate change, nitrous oxides (NO_x) and sulphur dioxide (SO_2), the former related to the formation of tropospheric ozone and both responsible for acid rain.

Carbon dioxide is an inevitable by-product of the combustion of organic material. NO_x comes from the nitrogen in the air and SO_2 from the sulphur in coal and oil. Other environmental effects of conventional steam power stations include the emission of particles and heavy metals, the generation of solid waste such as fly ash and slag, the heating of river, reservoir or sea water to meet the cooling needs of the steam cycle and, indirectly, the impact of mining. As no combustion is involved in the operation of nuclear power plants, they produce no CO_2 emissions and are not directly contributing to global warming.[9] Leaving aside the possibility of an accidental catastrophe, the inevitable accumulation of radioactive waste is, presently, an unsolved problem that affects coming generations so severely that nuclear power as it is known today cannot be regarded to be a sustainable source of energy.

In any event, even generation facilities that use renewable energy and are considered to be the most environmentally friendly technologies, have an adverse impact, in particular when the entire life cycle of the plants is contemplated. The most numerous of such facilities, namely hydroelectric power plants, which have

[9] For nuclear, as well as for any other technology, the entire life cycle must be considered, and not only the operation phase. Under this correct perspective no technology is free from emitting greenhouse gases or other environmental negative impacts, although significant differences exist among them.

existed ever since electric power was first industrialised, change the surroundings radically, flooding vast areas, altering rainfall, disturbing habitats or even transforming microclimates, not to mention the risk of accidents that can spell vast ecological and human disaster.

Other more recent technologies also have adverse consequences. Wind energy involves the disturbance of natural habitats and noise; solar energy, land occupancy and the pollution inherent in the manufacture of the components required for the cells, and more specifically the heavy metals present in their waste products; and biomass, the use of the land for this purpose and, potentially, an adverse CO_2 emissions life-cycle balance, depending on the biomass management approach. In fact, all electricity generation activities have one feature in common: visual impact due to land occupation, but the area involved and the (not necessarily proportional) extent of social rejection vary considerably with the technology and specific local conditions.

In a similar vein, the huge overhead lines that carry electric power across plains, mountain ranges, valleys and coasts and circle large cities have at least a visual impact on the environment, which is being taken more and more seriously. Less visible but indubitably present are the electromagnetic fields whose potential effects on people are still being evaluated. Such considerations have important consequences, since environmental permits and rights of way constitute strong constraints on the expansion of the transmission grid. As a result, the grid is operating more and more closely to its maximum capacity, occasioning new technical problems, for instance with respect to its dynamic behaviour, which logically have economic and reliability impacts. In some cases alternative solutions are available, albeit at a higher cost, such as running underground lines in densely populated areas.

The question, however, is not solely the establishment of the magnitude of the environmental impact of the electricity industry or of the awareness that minimising this impact generally entails increased system costs. The question, rather, is whether or not this impact should be considered when deciding how to best allocate society's scant resources. In a free market, the tool for resource allocation is product price, in this case, of the various power options. Nonetheless, the general opinion, of both the public at large and governmental authorities at the various levels, is that energy prices do not cover all the types of impact discussed above. This is what is known as a market failure or externality, defined as the consequences of production or consumption processes for other economic agents that are not fully accounted for by their costs.

The existence of such externalities, also called external costs, therefore leads to an improper resource distribution in the economy, preventing the market from suitable and efficient allocation of the resources on the basis of their price. Indeed, since account is not taken of these external costs, the price of energy is lower, and therefore consumption and environmental impact are higher than they would be if the true costs of resources were efficiently allocated. The existence of externalities, if not taken into consideration, also leads to the choice of more highly polluting power technologies than if allocation were optimum. In order to correct this

market failure and achieve optimum allocation, such costs must be internalised and built into the price in such a way that the economic agents can include them in their decision-making and ensure the best possible outcome for society as a whole. In the meantime, command-and-control type of regulatory measures can be adopted to create a more level playing field among more and less clean technologies.

1.6 Challenges and Prospects

This section is not intended to be a lengthy or detailed discussion of the prospects for change in the electricity industry, presently one of the most dynamic sectors of the economy. Rather, it contains a necessarily incomplete but, it is hoped, representative annotated list of developments that are expected to acquire importance in the electricity industry in the years to come. Some of these may change the industry radically in the future. For most, the enormous potential for change lies in the interactions among technology, economics and regulation.

- *Distributed generation*: the promotion of the use of renewable energy and combined heat and power for environmental reasons, together with technological advances leading to lower costs for wind and solar photovoltaic generation, micro-turbines or fuel cells, are prompting spectacular growth in this kind of generation. This, in turn, may entail profound change in the functions, planning and operation of transmission grids and distribution networks and the economic management of electric power systems.
- *Renewables*: wind and solar generation are two important assets in the quest for a sustainable energy model. Nevertheless, their large-scale connection to the grid could be problematic for several reasons: the variability and limited predictability of their production and their mid- and long-term unpredictability about their future deployment growth and their geographic distribution. New developments are needed to cope with the challenge to achieve more flexible demand-side management mechanisms able to induce rapid consumer response, to develop economically viable storage technologies and to increase the operational flexibility of the remaining generation sources.
- *Off-grid rural electrification*: the most appropriate solutions for supplying electric power to the over 1.4 billion people who presently lack access to electricity may often be small, isolated grids or individual systems, using suitable distributed generation technologies.
- *Environmental and strategic considerations*: environmental restrictions and the progressive internalisation of costs arising from environmental impact, together with long-term approaches for security of supply, will have a gradual and significant effect on future investment in new production resources. Specific markets for “green electricity” and pollutant emissions will tend to sprout up everywhere.

- *Multi-utilities*: as some electric utilities did in the past, today's companies are also offering additional services (gas, water or telecommunications distribution) in a single package, to take advantage of the synergies among these various lines of business.
- *FACTS*: the difficulties encountered in enlarging transmission grid capacity and the technical and economic problems caused by loop flows, especially in multinational markets, will further the development of electronic devices to control grid flows in order to optimise the use of the individual transmission capacity of each grid facility. Their universal use will change the traditional approach to transmission grid supervision and control.
- *Technical and economic management of regional markets*: multinational electricity markets are arising, consisting of the coordinated operation of organised competitive energy markets or exchanges in parallel with countless bilateral transactions. At the same time, technical management of security of operation will be left in the hands of a coalition of independent system operators, posing complex organisational problems that must be resolved if both system security and economic efficiency are to be adequately guaranteed. A paradigmatic example of such a problem is the coordinated management of grid congestion. New organisational schemes and communications and information systems, together with coordination models and algorithms, must be further developed to respond to this challenge.
- *Electricity trading in the digital economy*: the liberalisation of an industry as important as electric power, together with the development of e-commerce, is already leading to the spectacular rise of electricity trading on the Internet, with products ranging from long-term contracts to online purchases, and also including risk insurance.
- *Electric vehicles*: the environmental advantages of electric compared to conventional petrol vehicles point to a significant increase in their number in the future. The first effect of the fact that these vehicles must be plugged into the grid will be significant growth in demand in multiple locations; the truly promising feature of this technology, however, is its storage capacity. Such vehicles will not merely constitute a load on the grid, but storage devices able to return energy to the network as distributed micro-generation, thereby flattening the load profile and compensating for intermittent renewable generation if necessary.
- *Superconductors*: although the use of these components in electric power is still limited to the manufacture of large electromagnets and experimental facilities, superconductivity may change the future design of large transmission equipment, particularly in and around large cities.
- Further the integration of information and communication technologies (ICT) in the electric grids. The electricity network may be used for telecommunication purposes and information technologies may enhance network functions. Recent technical developments have made it possible to use local distribution grids for the high-speed transmission of information. Therefore, electric utilities might also provide competitive Internet services, while available cheap information

technologies are making possible remote metering of electricity consumption, advanced demand-side management schemes and enhanced distribution automation.

- The broad concept of so-called smart grids is the final objective of this integration of ICT and electric grids. Smart grids will be enhanced automated networks with the electrical, control and communications technologies necessary to achieve greater management flexibility. They will combine several of the ideas discussed above: high-scale integration of embedded generation, new distributed storage devices (including electric vehicles) and demand control, down to the household appliance level. This will result in new elements such as micro-grids (low-voltage networks connecting a set of distributed energy resources, which act as a single body, where appropriately controlled) or virtual plants (virtual aggregations of resources with complementary technologies operating on the market as a single agent). Thus, a wide range of alternatives is unfolding, indeed, a new paradigm is being created, essential to achieving a sustainable energy model.

References

1. Hunt S, Shuttleworth G (1996) Competition and choice in electricity. John Wiley and Sons edition, Chichester, UK
2. International Energy Agency (2010) World energy outlook 2010. OECD/IEA, Paris
3. Pérez-Arriaga IJ, Rudnick H, Rivier M (2009) Electric Energy Systems—An Overview. In: Gómez-Expósito A, Conejo A, Cañizares C (eds) Electric energy systems analysis and operation. CRC Press, Boca Raton, FL pp 1–50

References for In-depth Study of Specific Subjects are Listed Below

4. Bergen AR, Vittal V (1999) Power system analysis. Prentice Hall, Upper Saddle River, NJ
5. Schavemaker P, van der Sluis L (2008) Electrical power system Essentials. John Wiley & Sons, Chichester, UK
6. Sioshansi FP, Pfaffenberger W (eds) (2006) Electricity market reform An international perspective. Elsevier, Oxford, UK
7. Smil V (2001) Energy at the crossroads: Global perspectives and uncertainties. MIT Press, Cambridge, MA
8. Wood AJ, Wollenberg BF (1996) Power generation, operation and control, 2nd edn. John Wiley & Sons, New York

Already Classic, Pioneer Texts on Power System Restructuring and Liberalization

9. Joskow P, Schmalensee R (1983) Markets for Power: An Analysis of Electric Utility Deregulation. MIT Press, Cambridge, MA
10. Schweppe FC, Caramanis MC, Tabors RD, Bohn RE (1988) Spot pricing of electricity. Kluwer Academic Publishers, Boston, MA

Chapter 2
Power System Economics

Mariano Ventosa, Pedro Linares and Ignacio J. Pérez-Arriaga

Why the electricity industry underwent profound change worldwide in the 1990s is a question that cannot be fully answered without an understanding of the economic fundamentals.

This chapter introduces the economic principles that help understand why power systems are structured the way they are: why different electricity generation technologies are needed; why the various components of the power industry (generation, transmission, distribution and retailing) are structured as competitive markets or monopolies; and why they are regulated in different ways in different countries.

For many years the electricity industry was organised around vertically integrated, regulated monopolies. The fundamental economic reasons for this choice, which are studied in this chapter, include the existence of economies of scale, the nature of electricity as a non-storable product despite fluctuating demand and the consideration of electricity as a public service essential for society.

However, electricity markets, particularly for wholesale but also for retail trading, have been instituted in a significant number of countries around the world over since the early 1990s. Under the conventional economic paradigm, providing certain conditions are met, competitive markets are the preferred solution to achieve economic efficiency in any productive sector. This chapter deals with the fundamentals of electricity economics and explains why, under ideal conditions (not easily satisfied), the law of supply and demand is the best mechanism to allocate production resources and to set prices for goods or services.

However, in practice, competitive markets fail to ensure efficient output when: not all players have full information about the market; high transaction costs deter the entry of new players; or the commodities traded generate externalities (costs and benefits not acknowledged by the market). When such failures occur, markets must be regulated with suitable intervention and policy instruments.

Moreover, real markets may exhibit other types of imperfections that also require regulatory action and intervention. This chapter also reviews structural

M. Ventosa (✉) · P. Linares · I. J. Pérez-Arriaga
Universidad Pontificia Comillas, Instituto de Investigación Tecnológica,
Alberto Aguilera 25, 28015 Madrid, Spain

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_2,

imperfections in which some players' relatively large market share enables them to consistently set prices at other than competitive levels. This is known as market power. Bidding strategies and collusive behaviour in concentrated markets, known as oligopolies, with the intention of exercising market power, are discussed below.

The economic fundamentals introduced in this chapter are crucial to understanding the basics of the organisational models that have been in place throughout the history of the electricity industry, from monopolies to markets.

This chapter is organised as follows. Section 2.1 defines the basic economic principles and the main cost items involved in the generation and transmission of electricity, from investment to operation. Section 2.2 explains market fundamentals, including consumer and supply behaviour. Section 2.3 addresses perfect competition and the conditions for economic efficiency conditions. The existence of economies of scale and the need in this case to set a regulated monopoly as the single provider are explained in Sect. 2.4. Sections 2.5 and 2.6 discuss other market imperfections such as concentration in oligopolistic markets, strategic and collusive behaviour to exercise market power. Finally, Sect. 2.7 deals with market failures associated with environmental externalities.

2.1 The Role of Economics in Power Systems

Economics has been defined as the discipline that deals with the allocation of scarce resources to satisfy human needs. In this case, the resources in question are needed to produce and distribute electricity: typically capital, labour and fuel. Even distributed generation from a renewable source such as small-scale rooftop solar PV panels that require no fuel or distribution system calls for resources: basically, the capital needed to buy PV panels and the labour to install them.

The human need in this case would be the supply of electricity. That does not mean, however, that electricity should be considered a uniform commodity: the need for electric power may differ in terms of the time when it is consumed (because electricity is not easily storable), the quality of supply, or the type of appliance powered (air conditioning, heating, lighting or others). Therefore, different types of electricity can be distinguished: base load electricity, peak hour electricity, high quality electricity, low quality electricity; and electricity for heating, lighting and industrial use or transport.

The need for power also varies in extent: in some cases it may be an absolute necessity and in others a luxury. This generates differences in demand from one service or type of power to another, as well as in the willingness to pay for it. The cost of providing these services likewise varies: it may be very expensive under some circumstances, and nearly cost-free in others. The balance between these variations in demand and costs determines how this scarce resource is allocated.

Therefore, the economics of power systems may be concluded to be the discipline that studies the allocation of scarce capital, labour and raw materials to satisfy a range of electricity services. As in other areas of the discipline,

a distinction may be drawn between positive and normative economics. Positive economics describes the way goods are allocated in the real world; normative economics the way they should ideally be allocated. What is meant here by "ideal" must be defined before proceeding to the study of the economics of power systems. Sometimes "ideal" is "affordable"; sometimes it is "affordable and reliable"; today it is usually assumed to mean "welfare maximizing", where welfare includes not only monetary but also environmental and social considerations.

The three elements required to allocate resources to provide electricity-related services have now been identified: the characteristics of demand (the need), the costs of supply, and the allocation objectives and mechanisms. Each is characterised in the items below.

2.1.1 Electricity Demand

Electricity is a consumer good, but not the "typical" uniform good portrayed in economic manuals. The two main elements that characterise electricity demand are: first, it varies with time. Although it is consumed in a continuous flow, this flow is time-dependent. Second, electricity is non-storable.[1] This means that supply (which may be constrained in the short term) must match demand at all times, instantaneously.

These characteristics are not exclusive of electricity: road transport is similar in these respects. They do make its management a complex affair, however. Indeed, electricity can be more accurately analysed if it is regarded as a service rather than as a good.

Furthermore, the utility derived from electricity changes with time, quality and type of use (or user), giving rise to different services, rather than a single uniform one. These types cannot be freely interchanged, however, for flexibility in that regard is limited.

Therefore, demand for electricity is not all the same but varies with each type of electricity. The willingness to pay and the elasticity of demand (the variation in electricity consumption when prices change[2]) differs for these services.

Another feature of electricity is that it is generally considered, at least in developed countries, as a basic service that has to be provided universally, reliably and at an affordable price to facilitate on-going development. In fact, the supply of electricity is considered to be one of the key factors for enhancing welfare and it can be argued that access to electricity should be considered a basic human right.

---

[1] This is not absolutely true: electricity may be stored in batteries, flywheels and other devices. Presently the scale of such storage is small compared to the size of power systems. This may change in the future.

[2] Formally, the price elasticity of demand is defined as the percentage change in quantity demanded over the percentage change in price.

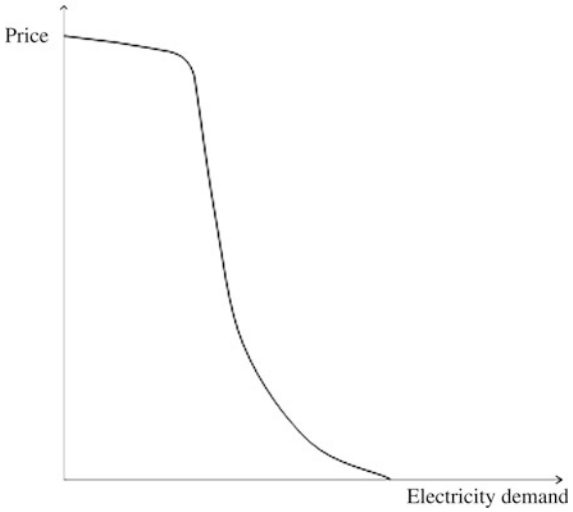


Fig. 2.1 Typical demand curve for electricity

All these elements—plus the fact that most electricity consumers are only exposed to uniform prices that only change about once a year or longer, see below—combine to make demand for electricity highly inelastic (i.e., changes in prices have only a minor effect on demand), particularly in the short term. Indeed, the usual assumption is for elasticity to be zero or very small in practice in the short term. It is an irreplaceable basic service, its share of household spending or industry expense is generally small, and most customers do not receive the appropriate price signals (in most countries the tariffs are still flat). Figure 2.1 shows a typical (short-term) demand curve for electricity.[3] In the long term, consumers may seek alternatives and thus widen elasticity. The typical short-term elasticity values, which range from 0.05 to 0.25 for households and from 0.1 to 0.9 for industries, may quadruple in the longer term.

As indicated before, another reason for such narrow elasticity is that (particularly for residential and small business) consumers are not offered different prices at different times of day; rather, electricity tariffs are usually flat. If consumers are unaware of prices or if prices are constant regardless of demand, the result is inelasticity.

This might change if consumers were charged different prices for different electricity services. This is the goal of the demand response programmes that have been in place in several countries for many years. Such programmes have become more popular with the advent of advanced metering, primarily because time of use pricing is only possible where time-of-use metres are installed. When prices vary,[4] the implicit elasticity is revealed, as shown in Table 2.1, which presents estimates

[3] The particular shape of the demand function in Fig. 2.1 will be explained later in Sect. 2.2.1.

[4] Variations may be structured around fixed time-of-use (TOU), critical peak pricing (CPP) or dynamic, real time pricing (RTP).

Table 2.1 Summary of price elasticity estimates (adapted from USDOE [9])

Target customers	Type of programme	Own price elasticity	Elasticity of substitution	Region
Residential (and small commercial)	TOU		0.07 to 0.21 (0.14 average)	US
	TOU/CPP	−0.1 to −0.8 (−0.3 average)		US-international
	CPP		0.04 to 0.13 (0.09 average)	California
	RTP	−0.05 to −0.12 (average −0.08)		Illinois
Medium or large commercial and industrial		−0.01 to −0.28		Georgia
		−0.01 to −0.27		UK
		<−0.01 to −0.38		N–S Carolina
			0.10 to 0.27	Southwest US
			0.02 to 0.16 (0.11 average)	New York

of the elasticity of demand for electricity for different types of uses and time frames.

The table shows two types of elasticity: own-price elasticity expresses the variation in demand in a given period for a 1 % price rise in that period, whereas elasticity of substitution expresses the demand that shifts from peak to off-peak periods given a 1 % higher peak than off-peak price.

More details about consumers' behaviour are discussed in Sect. 2.2.

2.1.2 The Costs of Producing Electricity

A clear idea of the costs of producing electricity is a requisite for the study of power systems. As shown below, the type and magnitude of the costs involved have a heavy impact on the structure of power generation, market or monopoly design, and system regulation.

This section contains a description of the general cost categories, followed by a definition of different types of costs. Finally, present electricity generation, transmission and distribution costs are reviewed.

2.1.2.1 Production Costs

The inputs for the supply or production of any good or service generally include: (i) natural resources or raw materials (for electricity the major such component is fuel, such as oil or gas), (ii) human physical and mental skills or labour, and (iii) capital, consisting of the manufactured goods, such as factories and machinery, used as production resources.

The total cost of producing a certain output, q, can therefore be expressed as a combination of the cost of labour, the cost of capital and the cost of natural resources. The production function, f, relates labour, capital and resources to the production of q units.

For any time interval the total cost can be expressed as the sum of variable and fixed costs:

$$\begin{aligned} q &= f(L, K, T) \\ \mathrm{TC}(q) &= w \cdot L \; + \; r \cdot K \; + \; e \cdot T \end{aligned} \tag{2.1}$$

where

T is total production cost
w is the unit labour cost
L is the amount of labour
r is the unit capital cost
K is the amount of capital
e is the unit cost of natural resources
T is the amount of natural resources

Note that the only costs generally included in the production function refer to inputs that have a price. Electricity generation, however, entails other resources that have no price and therefore are not included in the production function. These are "external costs", real costs imposed on society but not traded on markets. One example is the harmful effects of pollutants released from power plants on human health. Such harm clearly has an economic impact in the form of medical costs and earlier deaths, but it is not usually included in the production function because it is not traded on markets and therefore has no price. A possible solution to this is the creation of a market for such harm (or for emissions as a proxy), such as the EU Emissions Trading Scheme, which has put a price on carbon emissions, or to put a price directly through a tax. This issue is dealt with in greater detail in Sect. 2.7 and Chap. 11.

2.1.2.2 Fixed and Variable Costs

The notions of variable and fixed costs are associated with the basic inputs: labour, capital and natural resources. We shall refer in this section to the different components of the cost of electricity production. During a certain time period—a day, for instance—some of these costs do not change with the level of the total demand to be met (e.g. the salaries of the workers or the fraction of the investment costs corresponding to that day for the existing power plants), while the costs of the fuel to operate the plants during the day do change much with the demand level. We shall term "fixed costs" to the costs that remain unchanged for the considered period of time. While a cost might be fixed for a certain period, such as a month,

it may vary in the longer term, such as a year or a decade. In a time span of minutes all normally considered production costs are fixed. When intervals of hours, days, months or up to a year are considered, some costs (usually termed "variable costs") change with the demand level, while others (the so-called "fixed costs") are unaltered during this time. For longer time spans of several years, all costs of production may change with the demand level, even the investment costs, since there is time to retire old generation plants and network infrastructures and install new ones to meet the future demand. The per-unit costs of labour, capital and resources may also change during this time. The available technologies, which could be considered stable during intervals of several years, will also change over longer time periods. In the economic treatment of power systems the complex reality of operation and planning of power systems is commonly made to fit into just two time intervals. The short term typically ranges from several minutes to 1 year, when the installed capacity of generation plants and networks can be considered to be fixed, but their operating conditions and production costs depend on the demand level. The long term reaches several years, up to one or more decades, when investment decisions for new electricity infrastructures are materialised responding to the expected demand growth, among other factors.

$$\mathrm{TC} = \mathrm{VC} + \mathrm{FC} \tag{2.2}$$

where TC is total cost, VC is variable cost and FC is fixed cost.

2.1.2.3 Average and Marginal Costs

The average cost is the total cost divided by the output:

$$\mathrm{AC} = \frac{\mathrm{TC}}{q} = \frac{\mathrm{VC}}{q} + \frac{\mathrm{FC}}{q} = \mathrm{AVC} + \mathrm{AFC} \tag{2.3}$$

where AC is average total cost, AVC is average variable cost and AFC is average fixed cost.

Marginal cost is equal to the change in total cost when output rises or declines by one unit. It can therefore be expressed as the derivative of the total cost with respect to output:

$$\mathrm{MC} = \frac{\partial(\mathrm{TC})}{\partial q} \tag{2.4}$$

where MC denotes marginal cost.

Since in the short run capital costs are fixed, the short-run marginal cost is found as the derivative of the variable costs. Given that in electricity, fuel costs are the main variable cost, marginal costs are found as the derivative of fuel costs with respect to output, i.e. the amount of electricity produced. Typically the short-run

marginal cost of electricity production at power system level is determined by the variable cost of the marginal generator, i.e. the one responding to changes in demand at a given time.[5] In the long run, no cost, not even capital cost, is fixed. Consequently, the long-run marginal cost is found as the derivative of total costs, including investment, with respect to output. Short and long-run marginal costs may differ.[6]

A usual assumption in the electricity sector is that, in the long term, marginal costs are flat, that is, they do not depend on the level of output. This is based in turn on the assumption that there are no economies of scale to be exploited, or that the availability of the marginal technology or fuel is unlimited. This will be addressed in more detail later.

The importance of marginal costs in competitive markets and how they relate to average costs are discussed in Sects. 2.3 and 2.4 of this chapter, respectively.

2.1.2.4 Generation Costs

Electricity is generated in power plants, whose three basic cost elements are listed below.

- The construction of a power plant requires capital, labour, and raw materials. Taken together, these inputs constitute what is known as investment cost, which is incurred during the construction time—from a few months to up to 10 years or more, depending on the technology—and paid for (depreciation plus interest on borrowed capital) during the economic life of the facility, typically from 25 to 40 years or more. These costs are usually expressed in monetary units per kW of power installed. When the costs incurred throughout the construction period are "aggregated" in a single year, they are termed overnight costs.
- Operation and maintenance (O&M) covers the cost of the plant operating staff as well as maintenance operations and repairs (which include labour, capital and natural resources). Although fuel is obviously necessary for operation, fuel costs are generally dealt with separately. Some O&M costs are fixed during periods of 1 year or inferior (they are constant, regardless of the amount of electricity produced, such as personnel overheads), and therefore are typically measured in monetary units per year. Others depend on how and how much electricity is produced (such as turbine maintenance, which depends on operating hours and the frequency and absolute value of start-ups), and are measured in monetary units per kWh.

[5] This statement is essentially correct, but in practice other factors may affect the marginal production cost of a system, such as coupling among time periods, non linear production costs or network effects.

[6] In reality it is difficult to establish a base of comparison, since in the long run one usually refers to the total aggregated annual demand, while in the short run costs and demands refer to much shorter time intervals, such as hours.

- Fuel costs are an important element for fossil fuel and biomass units, much less so for nuclear plants, and zero for wind, solar, geothermal of marine power plants. Fuel costs can be measured at two points in the process: before entering the power plant, or as a fraction of the cost of electricity. In the former they are measured in a variety of units, depending on the fuel and the region (coal in units per tonne, gas in units per cubic metre, oil in units per barrel…). In the latter they are measured in units per kWh of electricity produced (taking into consideration the efficiency with which the raw fuel is converted into electricity). Fuel costs are neither necessarily constant nor linear with output: in thermal power plants, for example, they start from a certain output level (the technical minimum, below which the plant cannot function) and then evolve nonlinearly with the level of output.

These costs vary depending on the generation technology and fuel used. Some technologies require very large investments per installed kW, whereas others do not. Some rely on expensive fuels, whereas in others fuel is cost-free (solar or wind). Future electricity generation costs are, moreover, highly uncertain, since they depend on the variation in fuel prices as well as on technological developments. This explains the wide variations sometimes observed in future cost estimates.

Table 2.2 gives the estimates for future generation costs in year 2017 by the US Energy Information Administration in the Annual Energy Outlook 2011, released in January 2012. These are levelised costs,[7] which may be used to compare technologies, although with some caveats. Other sources of data are the International Energy Agency and the SETIS website[8] of the European Commission. The table does not include the start-up costs of thermal units, which comprise the fuel costs to get the plant ready to produce and the incurred additional maintenance costs.

The table shows the large inter-technology differences in investment and fuel costs. Some technologies, such as nuclear and hydro, have large investment but low operating costs. Others, such as combined cycle gas turbines, have lower capital but higher fuel costs. Still, peaking units—typically open cycle gas turbines—have lowest investment costs, although highest variable costs. As electricity cannot be

[7] The Levelized Cost of Electricity (LCOE) is the price at which electricity from a specific generation source must be remunerated to break even over the lifetime of the project. It is an economic assessment of the cost of the electricity-generating system including all the costs over its lifetime: initial investment, operations and maintenance, cost of fuel and cost of capital, and is useful in calculating and comparing the costs of generation from different sources. To evaluate the total cost of production of electricity, the streams of costs are converted to a net present value using the time value of money. These costs are all brought together using discounting cash flow. Typically LCOEs are calculated over 20 to 40 year lifetimes, and are given in the units of currency per MWh, for example €/MWh. It must be noticed that the LCOEs values are very dependent on assumptions such as the capacity factors, economic lifetimes or discount rates. *Source* Wikipedia.

[8] Strategic Energy Technologies Information System, http://setis.ec.europa.eu.

Table 2.2 Generation costs

Plant type	Capacity factor (%)	U.S. Average levelised costs (2010 $/megawatthour) for plants entering service in 2017				
		Levelized capital cost	Fixed O&M	Variable O&M (including fuel)	Transmission investment	Total system levelised cost
Dispatchable technologies						
Conventional coal	85	64.9	4.0	27.5	1.2	97.7
Advanced coal	85	74.1	6.6	29.1	1.2	110.9
Advanced coal with CCS	85	91.8	9.3	36.4	1.2	138.8
Natural gas-fired						
Conventional combined cycle	87	17.2	1.9	45.8	1.2	66.1
Advanced combined cycle	87	17.5	1.9	42.4	1.2	63.1
Advanced CC with CCS	87	34.3	4.0	50.6	1.2	90.1
Conventional Combustion Turbine	30	45.3	2.7	76.4	3.6	127.9
Advanced combustion turbine	30	31.0	2.6	64.7	3.6	101.8
Advanced nuclear	90	87.5	11.3	11.6	1.1	111.4
Geothermal	91	75.1	11.9	9.6	1.5	98.2
Biomass	83	56.0	13.8	44.3	1.3	115.4
Non-dispatchable technologies						
Wind	33	82.5	9.8	0.0	3.8	96.0
Solar PV[a]	25	140.7	7.7	0.0	4.3	152.7
Solar Thermal	20	195.6	40.1	0.0	6.3	242.0
Hydro[b]	53	76.9	4.0	6.0	2.1	88.9

[a] Costs are expressed in terms of net AC power available to the grid for the installed capacity
[b] Hydro is assumed to have seasonal storage so that it can be dispatched within a season, but overall operation is limited by resources available by site and season

readily stored, generation and consumption must be instantaneously matched. Consequently, during peak hours when electricity demand is high, most of the generation units in the system should be producing, while during off-peak hours when demand is low, some plants should be shut down because they are not needed. The economic optimisation of the total cost of electricity production calls for a mix of generation technologies to meet demand that fluctuates seasonally and daily.

Therefore, the optimal generation mix results in a combination of base power plants that produce electricity during many hours in the year (at both peak and off-peak times), mid-range or intermediate plants that follow the load changes and whose production varies between their rated capacities and their technical minima, occasionally shutting down during off-peak times, and peaking power plants that only produce at peak hours. Base units are characterised by high fixed and low variable costs and peaking units by the contrary, high variable and low fixed costs, with intermediate plants in the middle position. Nuclear, some coal units, and run-of-the-river hydro power plants are considered base load technologies, more flexible coal plants and combined cycle gas turbines (CCGT) are intermediate

plants, and single cycle gas turbines are the peaking technologies. This may change depending on the structure of the power system and the variation in the costs corresponding to each technology. Renewable-based generation such as wind and solar are also characterised by high fixed investment costs and very low variable costs, although in this case the hours in which electricity is generated depend on the availability of the resource.

Other important features of electricity generation technologies are the minimum, as well as the optimal, power plant sizes (and therefore total and per unit costs). Because of technical constraints and economic reasons, there is roughly a minimum achievable size for each technology, so that it would be technically difficult and economically very expensive to build a smaller generation unit, and there is roughly also an optimal size for each technology (e.g. no more than 1000–1500 MW for nuclear reactors of the most common technologies, around 500 MW for coal generation units or about 2 MW for on-shore wind generators, nowadays), so that, if more installed capacity is needed, it is better to add more units rather than to build a larger one. This explains the previously mentioned characteristic of electric power systems that long-term marginal generation costs, for any given technology, do not depend on the level of output, except in small and isolated power systems. Again, we refer readers to Sect. 2.4 where this issue is dealt with in more detail.

Figure 2.2 shows the daily generation profile for mainland Spain's electricity system for the week of November 8, 2010. An examination of the performance of the different generation technologies during the complete year 2010 allows us to reach the following conclusions:

- Some technologies, such as nuclear energy, furnish a constant (base) supply of power.

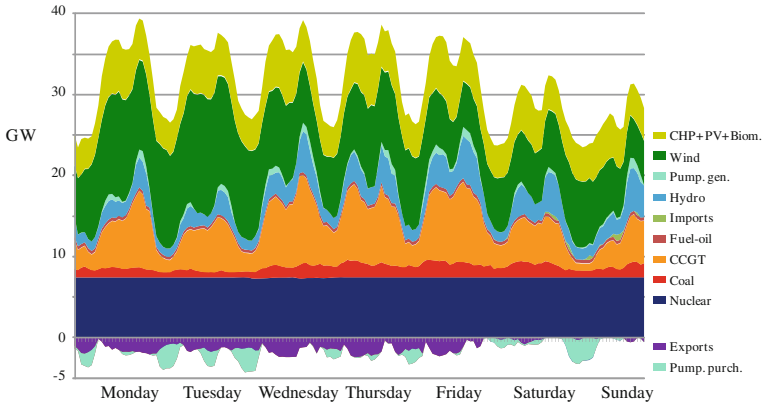


Fig. 2.2 Electricity supply in the Spanish system from Nov. 8–14, 2010 (Data taken from the system operator's information system (SIOS), www.esios.ree.es.)

- Others, such as coal and especially fuel–oil/gas, produce more at demand peak times than off-peak hours. In fact, many of these plants shut down during off-peak times.
- Hydroelectric plants operate primarily at peak times.
- Wind plants generate power basically irrespective of demand.
- Demand may be lower than supply in systems with pumped storage, if pumping is not considered an electrical demand.

The rationale behind a particular mix of technologies stems from the joint consideration of the fixed and variable costs and the technological capabilities of each one (storage hydro, for instance) and how competitive each one of them is to meet the different demand levels—base, intermediate or peak—, since each level requires to operate for a very different number of hours.

In principle, hydroelectric plants with storage should be regarded in short and medium-term operation optimisation as the most expensive variable cost plants, and operate only at peak times. This may seem paradoxical, since rainfall is cost-free. The reason is that only a limited amount of water is available in the reservoir for electricity production: only a certain number of MWh can be generated hydroelectrically per year. At the same time, since water is stored in reservoirs, the choice as to when it should be used to generate electricity can be made relatively freely.

The use of water should logically happen when such operation is the most cost-efficient, i.e. when the plant with the highest variable cost would have to be used instead. This, as noted above, is during peak times. Expressed in economic terms, the opportunity cost of water is the cost of the conventional plant with the highest variable cost.

Other types of generation, essentially wind, solar and reservoir-free hydroelectric generation, known as run-of-the river plants (a turbine installed in a riverbed, for instance), are essentially uncontrollable.

The cheaper power generated during off-peak times can be profitably used to pump water to storage reservoirs and later dump it onto turbines during peak times, instead of using plants with high variable costs for this purpose. Just how profitable this is depends, among others, on the efficiency of the pumping cycle.

2.1.2.5 Transmission and Distribution Costs

The major cost involved in electricity transmission and distribution is the investment needed to build grid facilities, i.e. power lines and substations and the subsequent operating and maintenance costs, which happen to be roughly proportional to the volume of network facilities and, therefore, to the investment costs. In the short run these costs do not change with the amount of electricity transported.

Ohmic energy losses happen in electricity networks, and depend quadratically on the amount of electricity transmitted. However, these are not network costs, but generation costs, since the cost of producing the electricity that is transformed in

Table 2.3 Overhead transmission line investment costs

Voltage (kV)	Number of circuits	Power rating MVA	Cost k€/km	Cost per unit k€/km/MVA
230	Single circuit	400	410	1025
230	Double circuit	800	660	825
345	Single circuit	750	580	773
345	Double circuit	1500	930	620
500	Single circuit	1500	830	553
500	Double circuit	3000	1325	433

Source Western renewable energy zone (WREZ) initiative [14]

heat in the wires took place in some generation plant. Unless network regulation somehow makes the owners of transmission or distribution facilities responsible for the losses that take place in these networks, the network owners do not incur into any additional cost that depends on how cold or hot the wires are. Total network[9] losses in well-developed networks could amount to 9 % of the total transmitted energy and they take place during the entire life of these installations, usually well beyond 40 years. Network loss reduction is therefore a significant issue in the design and operation of electricity networks.

Table 2.3 gives indicative figures on investment costs for different types of transmission facilities.[10] Annual fixed operating costs, usually expressed as a percentage of the total investment cost, range from 1 to 2 % for transmission lines up to 10 % or even higher for distribution facilities.

As shown in the table, network costs per unit of transport capacity strongly decrease with the capacity of the line, which is closely related to its voltage, among other factors. This fact has two important consequences, whose regulatory implications will be shown in the chapters devoted to distribution and transmission networks. The first one is that, in general terms, an electricity network should be built by a single company, since splitting the network into two or more competing networks will increase the total cost. Second, for the volumes of electricity that typically have to be transported between any two nodes and the distances involved, a single line of the right capacity can do the job. Even more, since investments in lines must necessarily be discrete, the best economically justified investment decision results in a line with much surplus capacity for the job to be done. This is very different from the case of generation in large power systems, where several generating units of the same technology have to be jointly used to meet the total demand and there is no need or justification for significant surplus capacity, beyond security margins. Strong economies of scale exist, thus, in electricity networks, with relevant implications in how the transmission and distribution

[9] Including transmission and distribution.

[10] Another interesting source of data on transmission costs is the EU research Project RealiseGrid, see http://realisegrid.erse-web.it/default.asp.

activities should be regulated. This effect is much more noticeable in transmission networks.

2.1.2.6 Other Costs of Power Systems

Readers running a quick calculation of the costs involved in electricity generation, transmission and distribution, including returns on investment, and comparing them to their electricity bill, may conclude that they are being swindled. This, of course, is not generally the case. The main reason for this divergence is that power systems incur more than just generation and network costs (this will be explained in detail in Chap. 8). These additional costs vary depending on the system but generally include:

- Compensations transitorily due to electric utilities derived from regulatory changes, such as the competition transition charges or stranded costs associated to many restructuring and liberalisation processes.
- Subsidies to domestic non-competitive fuels.
- Programs for the promotion of diverse types of renewable generation sources.
- Energy efficiency and savings programs and demand management programs.
- Support to energy-related R&D programs.
- Support to smart grid innovation activities.
- The costs of running a regulatory agency for the power system.
- The costs of running a power exchange for wholesale markets.

These costs may be quite significant in some systems, reaching for instance up to 25 % of the tariff or even more.

2.1.3 Economies of Scale. From Monopolies to Markets

Competitive markets based on the law of supply and demand are an efficient way to organise trading among producers and consumers. For many years, however, electricity supply and consumption have not been organised around markets where firms compete to supply consumers, but as utility monopolies supplying electricity to all the consumers located in a given area. The major reason for this was the presence of strong economies of scale in the power industry, so that the average production cost decreased with the volume of output until the vertically integrated utility reached a fairly large size.

Economies of scale are a determinant in industry structure, opening the door to competitive markets after exhaustion, or favouring the existence of natural monopolies otherwise. The exhaustion of economies of scale constitutes a key factor in the satisfactory development and operation of competitive markets.

Economies of scale are defined to be the decline in average production costs as output increases. This decline in average costs is due to the higher efficiency that can be obtained with larger-scale means of production or with a better coordination of these means of production. This has been the case for many years with generation of electricity, and it is still the case in small and weakly connected or isolated power systems, where installation of large generation plants is out of question. Network costs per unit of energy transported strongly decrease with the capacity of the line. Economies of scale may also appear when negotiating fuel or electricity supply contracts, or when negotiating financing terms for investments. After a certain output level (the break-even point) is reached, greater size may no longer afford higher efficiency, e.g. due to the need to deploy more expensive resources, the increase in financing costs, the managerial problems faced by very large companies and the risks associated with fluctuations in demand.

Monopolies

Where economies of scale exist, average costs decline with rising production plant size. Under such circumstances, only companies with large-scale production centres are cost-effective. The firm with the largest output can produce at the lowest cost, driving competitors out of the market if the economies of scale are not exhausted. Consequently, when the average costs of a given industry's entire output can be reduced by increasing output, the market ultimately tends toward a monopoly, as smaller and less efficient companies gradually exit the industry. In this case, cost subadditivity makes one producer more efficient than two or more. This is a natural monopoly situation. It is "natural" because of (1) the underlying characteristics of the production process and (2) the size of the market. Once competitors have been eliminated, a "natural" monopolist can exploit its position to raise prices above production costs.

Historically, electric utilities operated as monopolies in a given region and the businesses involved (generation, transmission and distribution) were vertically integrated to capitalise on both the economies of scale and coordination. In electric power distribution still the most efficient solution (lowest average cost of supply) is attained when only one company engages in the respective business, due to the economies of scale involved. In electricity transmission, also because of economies of scale, the network must be unique and centrally operated, although the individual assets may have different owners.

Monopolies are discussed in Sect. 2.4.

Markets

On the other hand, in large electricity systems, the economies of scale associated with generation have often been exhausted. For that reason, liberalising reforms have focused on organising generation around wholesale markets where producers sell their electric power, which is purchased by distributors (in systems where the retail market is not liberalised) or retailers for end customers. Lastly, under centralised operation full advantage can be taken of the significant economies of scale associated with the coordination between the means of electricity production and

Table 2.4 Type of markets by number of participants

Demand → Supply ↓	Many buyers	Few buyers	One buyer
One seller	Monopoly	Partial monopoly	Bilateral monopoly
Few sellers	Oligopoly	Bilateral oligopoly	Partial monopsony
Many sellers	Perfect competition	Oligopsony	Monopsony

the transmission grid (ancillary services). The mechanisms used to coordinate ancillary services are analysed in detail in Chap. 7.

Since the early 1990s, a series of technical and economic changes in production technologies and market sizes, as well as in the prevalent ideological concerns, have driven the change in the economic paradigm governing the organisational structure of the electricity sector. Regulators in several countries have opted for electricity markets as an alternative to traditional monopolies and introduced competition in generation and retailing activities. Electricity transmission and distribution, where strong economies of scale still prevail, continue to be regulated monopolies for the most part, however.

While ideal markets are optimal vehicles for attaining economic efficiency, the imperfections that characterise real markets may affect expected outcome and performance. In practice, then, markets may require regulatory intervention to ensure fair competition, the free entry of new players, and the emission of efficient price signals to all agents.

Table 2.4 summarises the definitions of market structure by the number of supply—and demand-side market players.

Competitive markets are discussed in Sect. 2.3. Prior to this, the basic concepts of the functioning of markets are presented next.

2.2 Market Fundamentals

The question central to microeconomics is: what forces determine how much of a given good or service is produced and the price at which it is bought and sold? This branch of economics uses two essential tools to provide an adequate reply to this question: the demand curve and the supply curve. The present section discusses the so-called law of supply and demand, showing that the demand for a given good or service is based on consumer preference and the supply on the costs of this supply. These two forces of any market economy, supply and demand, are balanced by price. We apply these ideas to the electricity generation market, given that, as mentioned before, transmission and distribution have different economic properties.

2.2.1 Consumer Behaviour and the Demand Curve

What forces govern consumer decisions when exchanging money for products in the marketplace? Consumers should logically decide to buy a product when its possession affords greater satisfaction than the price paid for it. Consequently, the criterion followed by all consumers, from grocery shoppers buying fruit or fish to a generating plant trader buying fuel on international energy markets, is to buy goods only if the price is less than their expected utility, i.e. the degree of satisfaction or profit obtained by the consumer. Increase in the level of satisfaction is measured in terms of marginal utility, where marginal means the additional utility obtained by consuming one additional unit of a given product or service.

Formally, consumers seek to maximise the difference between the satisfaction obtained when purchasing a certain amount of product, q, which is defined as total utility, $\mathrm{TU}(q)$, and the sum paid for that amount of product, calculated as the product of price, p, times the amount purchased, q. The following optimisation problem represents consumers' rational behaviour:

$$\frac{\partial(\mathrm{TU}(q) - p \cdot q)}{\partial q} = \mathrm{MU}(q) - p = 0 \tag{2.5}$$

In other words, consumers continue to purchase a good for as long as their marginal utility, which measures the increase in their satisfaction for consuming one additional unit of product, is higher than the price that they must pay for each unit. Similarly, when marginal utility is lower than the price, the rational decision is to refrain from purchasing the product. The optimal amount is therefore the amount at which marginal utility equals price. Consequently, Eq. (2.5) relates the optimal quantity that a consumer is willing to acquire to the purchase price.

Aggregating the amounts that all consumers would be willing to purchase at each price (horizontal sum of all consumers' marginal utility) yields the demand curve shown in Fig. 2.1, which relates the demand for a product and its price. This curve shows that the demand of all the buyers for a product increases when its price drops, as more consumers are willing to buy it. Therefore, this curve should have a negative slope, as illustrated in the figure.

This diminishing marginal utility for consumers is common to all markets and is simple to understand. For instance, since the first megawatt-hour (MWh) of electricity consumed by a factory is used to power the facilities deployed in its most productive activities, its utility is very high, whereas the final MWh consumed by the same factory might be used to power the car park lighting or some similar low priority application.

When consumers have no way of experiencing the price of electricity in real time (this is the case of all consumers with traditional meters), the curve of demand of electricity is vertical, since the demand is completely inelastic. Since demand is unresponsive to price, a regulatory intervention is needed to set the market price in case supply is not able to meet demand. In most systems the regulator simply establishes a price cap, which tries to represent in a single number

the value of the loss of utility for an average consumer of giving up electricity consumption. This is obviously a crude approximation, since the "value of lost load" depends on the nature of the consumer, the time of the day, the amount of forgone consumption or the anticipation in the notification of the curtailment. The shape of the upper section of the demand curve in Fig. 2.1 and the non-completely vertical middle section try to convey the idea of the response of a mix of diverse consumers with different levels of knowledge about the real-time prices. The lower section shows how demand saturates and cannot grow beyond a certain value, even if the electricity price becomes zero.

To summarise, optimal consumer behaviour can be explained by the demand curve, which plots marginal product utility versus amount demanded. Consumers buy a product whenever their satisfaction, measured in terms of marginal utility, is greater than the price: this is the portion of the demand located to the left of the equilibrium point of the market in Fig. 2.4. Conversely, when the price is higher than the marginal utility, consumers decide not to buy; this is the portion of the demand located to the right of market equilibrium in Fig. 2.4.

2.2.2 Producer Behaviour and the Supply Curve

Just as the demand curve defines consumer conduct, the supply curve explains producers' market behaviour. This function expresses the relationship between the quantities of a product that firms are willing to supply and the market selling price. Producer behaviour and the supply curve depend heavily on two factors: production costs and the number and size of companies competing on the same market.

An understanding of the supply curve calls for an analysis of optimum producer behaviour. When companies offer their products on the market, they seek the highest possible profit, which means that they must compare the cost of producing one additional unit to company revenue for the sale of that unit. The following sections discuss how companies may be expected to behave in two extreme situations: under perfect competition (Sect. 2.3) and as a monopoly (Sect. 2.4).

The aggregate supply curve for all producers relates selling price to the quantity of product supplied by all producers and is used to obtain the market equilibrium between supply and demand.

Figure 2.3 shows the typical shape of the generation supply curve to meet demand in the short term, for a particular moment of time.[11] The shape is due to the simultaneous existence of a variety of generation technologies:

Must-run plants, with a minimum output level because of a variety of reasons: run-of-the-river hydro plants, ecological minimum river flows, technical

[11] However, things get more complicated when multiple consecutive levels of demand are considered. For example, the hourly change in the output of each generation unit is limited by the ramp rates, and start-up costs should be also considered.

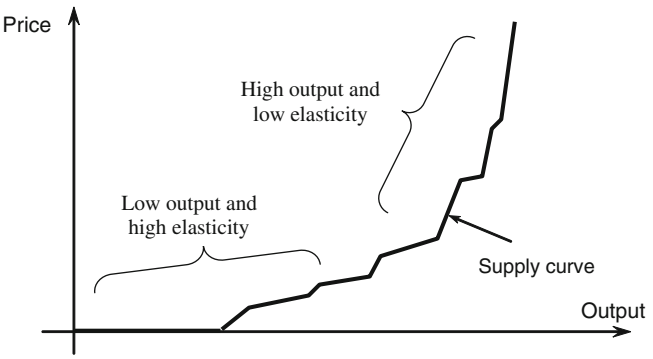


Fig. 2.3 Typical electricity supply curve in the short term

constraints in the transmission network, technically constrained steam generation, including nuclear plants that cannot afford to shut down for short intervals or technical minima of thermal plants that must be in operation because of security reasons, like meeting a minimum level of operating reserves, expected ramps of demand or wind output uncertainty.

Base load power plants, with the lowest variable costs (zero as wind or solar power, or low such as nuclear and efficient coal). This area of the supply curve is typically highly elastic (i.e. small changes in price result in large changes in production).

Mid-range plants (such as less efficient coal and CCGTs, although this depends on fuel and, if this is the case, carbon prices) corresponding to the middle section of the supply curve.

Peaking plants, the facilities in the high output level area, include the units with the highest variable costs (single cycle gas turbines, oil and the dispatchable fraction of hydro plants with storage, which therefore have the opportunity of producing at times with the highest market prices), which come on stream during peak hours only. The vertical shape of this part of the curve denotes low elasticity of the volume of supply to prices.

2.2.3 The Law of Supply and Demand

The law of supply and demand is the reply to the question posed at the beginning of this section. Market equilibrium is reached at a point where the quantity supplied equals demand, in other words, the point where the supply and demand curves intersect. At this point and at this point only, price balances supply and demand, since consumers buy all the units with greater utility than this price and reject the purchase of units whose utility is lower. This same reasoning can be applied to producers.

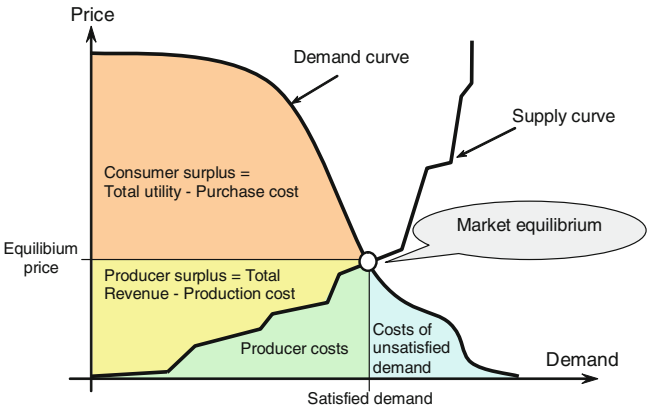


Fig. 2.4 Equilibrium between electricity supply and demand in the short term

Market equilibrium in the short term is illustrated schematically in Fig. 2.4, which is also useful to determine the consumer surplus. This surplus is defined to be the total utility less the total price paid, i.e. the area located between the demand curve and the equilibrium price. Similarly, the area to the right of the equilibrium point underneath the demand curve is the aggregate consumer utility that is not satisfied because marginal utility is lower than the equilibrium price.

Where producer behaviour conforms to perfect competition as described in the section below, the supply curve coincides with the marginal cost curve in the short term. From this it may be deduced that the area under the supply curve represents producer costs, and therefore producer surplus is the area to the left of the equilibrium point above the supply curve, i.e. total revenue minus total costs.

In most existing electricity markets, a fraction of the energy—ranging from one hundred to a few per cents—is traded on short-term (i.e., spot) markets (so-called electricity pools or power exchanges, see Chap. 7), typically day-ahead markets, in which the market equilibrium point in the short-term is reached by auction (or an optimisation algorithm based on market agents' bids and offers). The product being negotiated in such auctions is normally the price of electric power for a given hour on the following day. Generally speaking, all demand-side bids for a price above the equilibrium price are accepted and the demand they represent is met by all the supply-side bids offering a price lower than the equilibrium price. On the contrary, offers to purchase power at a price lower than the clearing price and sales offerings at a price higher than the clearing price are rejected. In some wholesale markets each hour is auctioned separately, in others the bids for electric power are submitted for whole days. This procedure for establishing the price and determining the amount of power to be generated by specific units in accordance with demand is what is known as market clearing. More details about these issues are given in Chap. 7.

2.3 Perfect Competition

The perfect market principle provides the theoretical economic grounds for the social and political system known as the free market economy. In practice, however, the conditions guaranteeing the existence of a perfect market are difficult to meet. Broadly speaking, perfect competition is said to exist when no producer is able to exert individual influence on price: in other words, when all suppliers are "price takers", i.e. they must sell their output at the price provided by the market.

The conditions for perfect competition are met when a large number of small firms produce a homogeneous[12] good or service in quantities small enough so that none of the firms may have an influence on the market price. Since in a power system there are power plants of different technologies that have to meet the multiple levels of demand, as shown in Fig. 2.3, this general microeconomic statement translates into the power sector as the following two sub-conditions: (a) the power system is large enough, so that there is a sufficient number of similar power plants of each technology and there are no significant economies of scale; (b) the ownership of the plants of each technology is well distributed among the several competing firms, so that none of them has a large percentage of the installed capacity of any given technology. Still it has to be added that the total installed generation capacity that is available in the power system at any given moment must exceed the demand at that moment in time by at least a required minimum margin. In this way, a generation company can only increase the marginal price in the wholesale market by withdrawing several plants, therefore being easily detected.

An additional condition for perfect competition is that all the actors must have perfect or full information on the going price, which is tantamount to saying that the market is transparent, with all buyers and sellers in possession of the same complete information for trading.

2.3.1 Perfectly Competitive Markets in the Short Term

As explained earlier, market equilibrium is the point where the supply and demand curves intersect. Given a demand curve in the short term, all it takes to find the market's equilibrium point is to obtain the short-term supply curve assuming perfect competition. To determine the short-term supply curve (no long-term considerations, such as investment decisions, are taken into account), each player

[12] In wholesale markets, electricity is a homogeneous good in each time interval, regardless of the producer or the source of primary energy used, but perfect competition requires the presence of multiple production firms competing at each demand level.

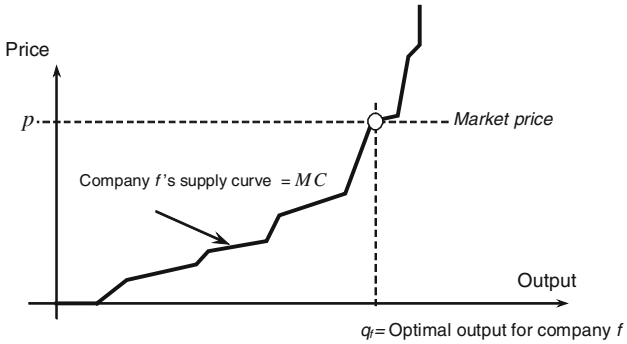


Fig. 2.5 Short-term supply curve for a company operating on a perfect market

on a competitive market is assumed to behave as a profit maximiser. The profit, $\prod_f$, of each market firm,[13] f, is calculated as:

$$\prod_f = p \cdot q_f - \mathrm{TC}_f \quad \forall f \tag{2.6}$$

Where p is the market price and TC_f is company f's total cost to produce q_f.

The production level providing the highest profit is calculated by setting the value of the derivative of the profit with respect to output, q_f, to zero. The definition of marginal cost as given in Eq. (2.4) is also included in expression (2.7):

$$\frac{\partial \prod_f}{\partial q_f} = p + q \cdot \frac{\partial p}{\partial q_f} - \frac{\partial \mathrm{TC}_f}{\partial q_f} = p - \mathrm{MC}_f(q_f) = 0, \quad \forall f \tag{2.7}$$

The above equation assumes that the derivative of the market price with respect to company output is zero, due to the previous definition of a perfect market, where each company only controls a tiny fraction of each part (in turn corresponding to one technology) of the supply curve.

The interpretation of Eq. (2.7) is straightforward. Each company f's entire output should be produced with facilities that can operate at a marginal cost MC lower than p, since these units ensure profits; by the same token, none of its facilities whose MC is greater than p should be operated, since they would generate losses.

Figure 2.5 shows the response of each company's production function to the market equilibrium price.

In conclusion, on a perfectly competitive electricity market, each producer's optimum output and price are defined by its marginal cost curve, and, in the aggregate, this ensures that demand is met by the most efficient units, see Fig. 2.6.

[13] The optimisation problem for each company is posed in simplified terms. In other words, it is assumed that cost is defined by a monotone non-decreasing function, and no limitations or restrictions to production are considered.

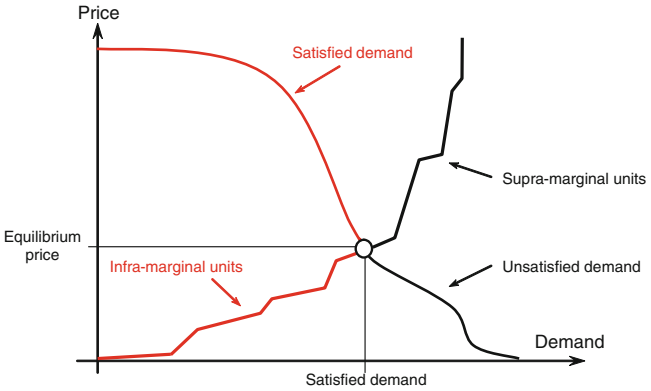


Fig. 2.6 Short-term equilibrium between supply and demand on a perfect market

The interpretation of Figs. 2.5 and 2.6 is also fairly straightforward from the standpoint of a generator or electric power producer. Generators whose units can produce power at a short-term marginal cost lower than the market price (supply to the left of the short-term equilibrium point) will be dispatched, whereas generators whose marginal cost is greater than the market price (supply to the right of the short-term market equilibrium) will not.

Similarly, from the standpoint of demand, buyers whose short-term marginal utility is greater than the price will decide to buy because this purchase affords a satisfaction greater than zero (demand to the left of equilibrium point), whereas buyers whose short-term marginal utility is lower than the price will decide not to buy because the "satisfaction" stemming from any such purchase would be negative (demand to the right of the equilibrium point).

The conditions guaranteeing perfect competition cannot be fully met in existing electricity markets, where normally only a few large producers operate plus, perhaps, a few other small ones. Moreover, consumer behaviour is very price inelastic in the short term, i.e. demand is almost vertical because consumers use electricity with no heed to price (or they do not receive any information on the value of the short-term price of electricity). Nonetheless, a study of market behaviour in perfectly competitive electricity markets is insightful for subsequent comparison to the conditions prevailing on imperfect markets.

2.3.2 Economic Efficiency in the Short Term

Microeconomic theory contends that both a perfectly competitive market (when conditions for such a market exist) and a centralised economic management of the operation of the power system with perfect information ensure the optimal allocation of scarce resources. Such efficiency can be perceived more or less intuitively from Fig. 2.6; in the short-term goods are produced with the most efficient (lowest

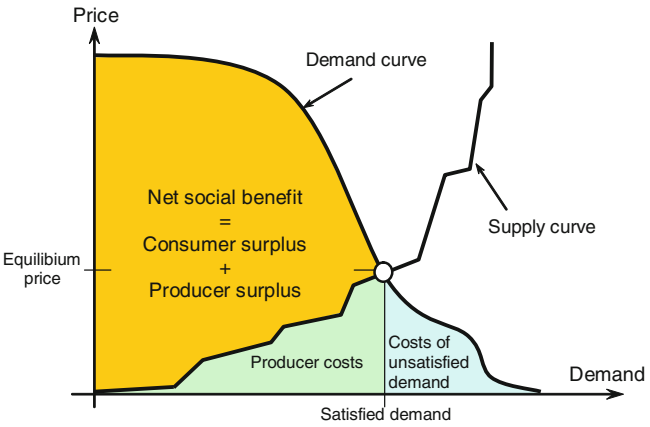


Fig. 2.7 Consumer and producer surplus at equilibrium in the short term in a perfectly competitive market

cost) means of production, located to the left of the equilibrium price on the supply curve, whilst the consumers most keen on buying (for whom utility is greatest), i.e. those located to the left of the price on the demand curve, see their needs satisfied.

The following is a more formal discussion of the above idea. In the short-term market equilibrium, Fig. 2.7, two features merit specific attention: consumer and generator surplus in perfectly competitive markets.

Firstly, economic efficiency under centralised short-term management of supply and demand must be defined and compared to the results obtained on a perfectly competitive market. Economic efficiency is attained when the welfare of society as a whole is maximised. This welfare is defined as net social benefit (*NSB*), calculated as the sum of consumers' and generators' surplus.[14] In the figure above, *NSB* is the area bound by the price axis and the demand and supply curves.

An ideal centralised decision maker with complete access to the information of every consumer and generation would find the maximum *NSB* from the following optimization problem:

$$\max_{q} \ \mathrm{NSB}(q) = \max_{q}[(\mathrm{TU}(q) - p \cdot q) + (p \cdot q - \mathrm{TC}(q))] \tag{2.8}$$

In this equation for maximising NSB, the decision variable is the output, or q, traded on the short-term market. The first parenthesis represents the consumer surplus, computed as total utility, TU, minus the purchase cost (p times the output, q, purchased), whilst the second parenthesis represents the generator's surplus, computed as revenues (p times the output, q, sold) minus total production costs, TC. Cancelling out the two identical but oppositely signed terms, $p \cdot q$, and setting

[14] Social welfare may include other concepts, besides the aggregated surpluses, such as technology diversification, security of supply or reduction of environmental impact.

to zero the derivative with respect to the decision variable q, yields the condition of optimality for (2.8) above:

$$\frac{\partial \mathrm{TU}(q)}{\partial q} = \frac{\partial \mathrm{TC}(q)}{\partial q} \Rightarrow \mathrm{MU}(q) = \mathrm{MC}(q) \tag{2.9}$$

The above result shows that the most efficient solution is reached when consumer marginal utility equals producer marginal utility in the short term. Intuitively, that means that under optimal conditions, if production is increased by one unit, the additional or marginal cost of producing that unit is equal to the additional increase in satisfaction of demand, measuring such additional satisfaction as marginal utility.

The following analysis of how consumers and producers behave in the short term on a perfect market is subsequently compared to the condition for economic efficiency expressed mathematically in (2.9).

The demand side attempts to maximise its surplus separately. Finding the derivative of its optimisation expression yields a familiar result: consumers decide to buy as long as prices are less than or equal to marginal utility.

$$\max_q \ \mathrm{TU}(q) - p \cdot q \Rightarrow \frac{\partial \mathrm{TU}(q)}{\partial q} - p = 0 \Rightarrow \mathrm{MU}(q) = p \tag{2.10}$$

The supply side, in turn, attempts to maximise its own surplus separately. Finding the derivative of its optimisation expression also yields a familiar result: producers operating on a perfectly competitive market decide to sell as long as the price they can command is greater than or equal to their marginal costs.

$$\max_q \ p \cdot q - \mathrm{TC}(q) \Rightarrow p - \frac{\partial \mathrm{TC}(q)}{\partial q} = 0 \Rightarrow p = \mathrm{MC}(q) \tag{2.11}$$

If the above results for consumer (2.10) and generator (2.11) behaviour are combined, the conclusion drawn is that on a perfect market $\mathrm{MU}(q) = p = \mathrm{MC}(q)$, as sustained by Eq. (2.9), which is exactly the proof sought: both a perfectly competitive market and centrally managed power system operation with perfect information yield the same outcome and ensure the optimal allocation of scarce resources.

By way of conclusion, when consumers and producers operating in the short term on a perfectly competitive market negotiate who sells, who buys and the transaction price, the result is economically efficient, since total economic welfare is maximised. Furthermore, the price is the signal that tells each agent whether or not to effect a transaction: buyers compare the price with their marginal utility and sellers with their marginal cost. That maximum net social benefit can be attained when each consumer and producer acts independently in pursuit of his own best interest is what Adam Smith called the "invisible hand" of competition.

Based on this premise, short-term economic efficiency can be mathematically shown to be attained both by a perfect market and by the ideal centralised decision maker in possession of perfect information. Nonetheless, in practice, these two approaches are not equivalent due to many reasons, the most relevant being that

perfectly competitive markets do not exist, and that no actual centralised decision maker would ever possess full information and could decide on behalf of each one of the consumers. The choice of a pro-market approach must take into account its potential shortcomings: market failures, information failures, regulatory failures or lack of complete markets. However, in general, it is considered that markets are better able to solve the information failure, and may therefore be more efficient, as long as market failures such as lack of competition or externalities are properly addressed.

2.3.3 The Transition Between the Short-Term and the Long-Term

The analysis of economic efficiency in the short term considers the optimal allocation of supply resources to meet price-responsive demand at a given instant in time. Things get more complicated when a more complete perspective is adopted. On one hand, daily, weekly and seasonal variations in electricity demand can cause the optimal mix of generation technologies and the corresponding electricity price to change at every time instant. On the other hand, the mix of installed generation capacities (which cannot change in the short term) must evolve in time to adapt to factors such as demand growth, changes in fuel prices and changes in investment costs. In these contexts, electricity prices must allow full recovery of investment and operation costs.

Several distinct features of electric power systems will be important here. First, due to the multiple levels of demand that the installed generation capacity in a power system has to cover, a mix of different technologies (as opposed to a single dominant technology) happens to be the optimal investment choice that maximises total social welfare. Second, in power systems of a certain size, there will be several similar generation units of each adopted technology (base, intermediate or peak). Therefore, the long-term marginal cost for each technology can be assumed to be independent of that technology's output level; i.e., increments in demand at each level will be met by just adding more units of the same technology, and there will be no economies of scale. Third, the lack of significant storage capability in present power systems amplifies the division between the short and the long terms: market prices that change every hour or less (every 5 minutes in some power systems) are the basis of the remuneration for diverse portfolios of power plants that have taken years to build and that have very significant investment costs. Finally, the operation and capacity expansion planning of power systems are often subject to relevant constraints that distort elegant, general-purpose microeconomic principles. It is necessary to understand these distortions to develop and apply meaningful regulation to the power sector.

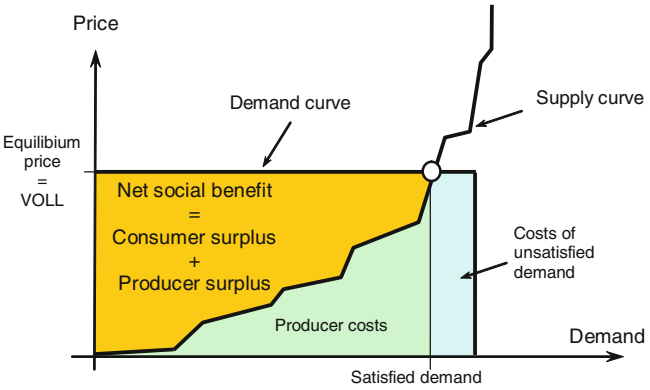


Fig. 2.8 Consumer and producer surplus at equilibrium in the short term in a perfectly competitive market, with a simplified representation of the relationship between demand and price

A stylized optimization model of the capacity expansion and generation processes in a simple power system will allow us to show the main features of, and relationships between, the short- and the long-term decision making processes.

2.3.3.1 A Simple Example

Let us assume a power system with just three generating technologies, base load (B), intermediate (I) and peak (P), and the annual load duration curve represented in Fig. 2.9. As described in Sect. 2.1.2, base load plants have high fixed costs and low variable costs; intermediate plants have intermediate fixed and variable costs; and peakers have low fixed costs, but high variable costs.

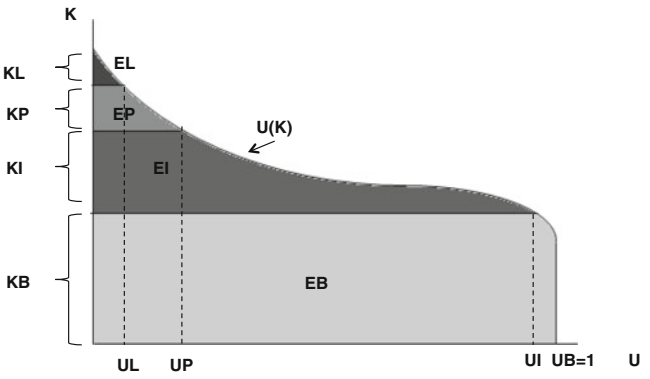


Fig. 2.9 Graphical solution for the centralised planner capacity expansion problem

In this simple setting with just three generation technologies, ignoring any network aspects and other features of actual power systems that would unnecessarily obscure the simple message to be spelled out here, we shall examine: (a) how a fully centralised decision maker would operate and expand a power system to optimally maximise social welfare; (b) how a monopolistic utility should be regulated to achieve the same result; and (c) the similarities and differences with the outcome that could be achieved if a competitive market were created in this power system.

To arrive in the most direct way at the fundamental results that we are after, some characteristic traits of actual power systems have not been included in the model: no network effects of losses or operation constraints; chronology is lost as we use the annual load duration curve rather than the actual hourly demand shape; continuous variables, rather than discrete ones, are used to represent the generation investments; technical minima of power plants and start-up costs are ignored; and the model is deterministic (i.e., the different sources of uncertainty are represented by their average or more characteristic values). As shown later, power sector models that are more detailed than the simple example presented here ultimately yield the same conclusions.

The Centralised Planning Model

In the centralised model, the decision maker has complete control over decisions about the operation of power plants and demand response in the short term, as well as decisions about investing in new generation capacity to meet demand growth or to replace retiring plants. The objective of the centralised decision maker is to maximise total social welfare, which is defined as:

$$\text{consumerSurplus} + \text{supplierSurplus}$$

Consumer surplus is measured in monetary units, such as dollars or Euros, and represents the net difference between the economic value (utility) of electricity to consumers and the cost of acquiring that electricity. Producer surplus is also measured in monetary units and defined as the difference between electricity sales revenues and supply costs (both investment and operation costs). Combining these two definitions, the central planner maximises

$$(\text{consumerUtility} - \text{electricityCost}) + (\text{salesRevenue} - \text{supplyCost})$$

Because electricityCost and salesRevenue are the same (because the producer receives what the consumer pays), these two quantities cancel and result in the following definition for social welfare:

$$\text{totalSocialWelfare} = \text{consumerUtility} - \text{supplyCost}$$

Demand elasticity is simply modelled here by assuming that demand does not respond to price until it reaches a threshold. At this threshold, the assumed per unit cost of non-served energy for any consumer is the "value of lost load" (VOLL). When the price of electricity exceeds this level, demand prefers to quit, since the

acquisition cost becomes larger than the utility apportioned by the electricity consumption. Under this assumption, the demand—and therefore, the utility of electricity to consumers—does not depend on supply or electricity prices except for a partial load curtailment when the price of electricity reaches the per unit cost of non-served energy. This simplification is frequently used in all sorts of electricity models and also in the rules of competitive markets as a default representation of "passive demand", which does not have the opportunity to respond to real-time prices. Figure 2.7 would become Fig. 2.8 under this assumption. Note that both figures correspond to a model of demand as a single value. However, even in this simple example of a power system, demand of electricity has multiple values during the considered interval of time, as shown by the load duration curve in Fig. 2.9. We shall assume here that the response of demand to price in Fig. 2.8 applies to all levels of demand in Fig. 2.9.

Examination of Figs. 2.7 and 2.8 shows that the central planner's problem of maximising total social welfare (the net social benefit in the figures) for any of the multiple demand levels is equivalent to the problem of minimizing the total supply cost including the cost of any non served energy (nseCost)[15]:

$$\max(\text{totalSocialWelfare}) \approx \min(\text{supplyCost} + \text{nseCost})$$

Our centralised planning model contains other significant assumptions, in addition to the simplified representation of demand response to price, The optimization of operation and capacity expansion is performed greenfield (i.e., starting from scratch) and for just one single future year, ignoring the trajectory of investments that would otherwise take the system from the present to any other future year. The investment variables that represent the total installed capacity of a given generation technology are continuous. The fixed and variable per unit costs for each technology are constant, regardless of the volume of investment (i.e., no economies of scale) and the production level (i.e., nonlinearities in the efficiency) of each plant. Technical minimums and start-up costs of generation plants are ignored. There is no uncertainty in the model; therefore, the demand is perfectly known and generator failures are not considered.[16]

The optimization model under centralised planning consists of minimising the total investment and production cost required to meet the annual demand using the three available generation technologies:

[15] The reader is invited to play with Fig. 2.7, by assuming that the market price that determines the level of demand and the remuneration of generators is higher or lower than the equilibrium price in the Figures. The reader will discover that the combined new surpluses of consumers and producers (or, alternatively, the combined new producer costs and costs of unsatisfied demand) will be lower (alternatively, larger) than when the equilibrium price is applied. Therefore, the net social benefit is maximised when both the producers and the consumers experience the equilibrium price.

[16] As it will be shown later, these assumptions can be removed without changing the conclusions obtained from this simplified model.

$$\min_{k_b, k_i, k_p, e_b, e_i, e_p, e_l} \text{totalCost} = \text{TC} = k_b f_b + k_i f_i + k_p f_p + v_b e_b + v_i e_i + v_p e_p + v_l e_l \tag{2.12}$$

where, for each technology (base load b, intermediate i and peak p), k represents the installed capacity (MW), f represents the per unit fixed cost (€/MW), v represents the per unit variable cost (€/MWh), and e represents the production level (MWh). The last term is the loss of consumer utility because of non-served energy in the system, where the l subscript refers to lost demand and v_l is the per unit value of lost load (termed VOLL frequently in the technical literature). The decision variables in this optimization problem are the capacity investment for each technology, the energy produced by each technology and the total amount of non-served energy. As we shall see shortly, these variables are not all independent.

Let us assume that the values k_b, k_i, and k_p of the three installed capacities are known. Because the per unit variable cost of the base load technology is lowest, we should use it as much as possible. Graphically, maximising the use of the base load technology is analogous to filling the lower part of the load duration curve as shown in Fig. 2.9. Note that this also determines the amount of energy e_b produced with the base load technology. Next, the capacity of the intermediate generation technology will be used as much as possible, yielding the energy production e_i. Finally, the installed capacity of the peaking technology will produce e_p. If the values of the three installed capacities are not enough to meet the entire demand, there will be some non-served energy e_l.

Once the capacity values k_b, k_i, and k_p are known, the values for the corresponding energies that are produced by each technology and the value of lost load can be determined graphically in Fig. 2.9. The area inside of each partition represents the total production level, e, for that technology. The x-axis values that k_b, k_i and k_p intercept the load duration curve at represent the capacity factor, u, for each technology.

Therefore, the optimization problem has only three independent variables, the three installed capacities k_b, k_i and k_p. Next, we shall solve the optimization problem analytically. As it is well known, the optimality conditions that the three independent variables have to satisfy are that the partial derivatives of the objective function to be minimized—i.e. the total cost of investment plus operation TC—, with respect to the installed capacities of each technology are zero (a standard condition for finding global minimums and maximums).

$$\frac{\partial \text{TC}}{\partial k_b} = \frac{\partial \text{TC}}{\partial k_i} = \frac{\partial \text{TC}}{\partial k_p} = 0 \tag{2.13}$$

Let us start examining the first optimality condition for the installed base load capacity k_b. In the first line of the equation below it can be observed that only a few terms of the total cost are affected by a change in k_b and thus need to be examined. The impact of a unit increment of k_b on the investment cost is f_b, obviously. The derivatives with respect to the four operation costs in the parenthesis can be easily obtained by simple inspection of Fig. 2.9. When k_b is

incremented by one unit (i.e. 1 MW) this additional capacity displaces a horizontal layer of demand 1 MW high and of length u_i that was previously supplied by intermediate plants. Because the base load plant has a lower variable cost, this increment in base load capacity saves $u_i \cdot (v_i - v_b)$ in operation costs. However, the unit increment in k_b also impacts the remaining production costs in Fig. 2.9 because it shifts the rest of the technologies upward. Specifically, there is a layer of length u_p where production by peakers is replaced by production by intermediate plants, with a savings of $u_p \cdot (v_p - v_i)$. There is also a layer of length u_l where former non-served energy is now supplied by peaking plants, with a savings of $u_l \cdot (v_l - v_p)$. This is the first optimality condition:

$$\begin{aligned}\frac{\partial \mathrm{TC}}{\partial k_b} = 0 &= f_b + \frac{\partial}{\partial k_b}\left(v_b e_b + v_i e_i + v_p e_p + v_l e_l\right) \\ &= f_b - u_i(v_i - v_b) - u_p\left(v_p - v_i\right) - u_l\left(v_l - v_p\right)\end{aligned} \tag{2.14}$$

The optimality condition for the installed capacity of intermediate plants k_i is even easier because it only impacts more expensive technologies and leaves the base load production unaltered. Proceeding in the same fashion, we obtain:

$$\frac{\partial \mathrm{TC}}{\partial k_i} = 0 = f_i + \frac{\partial}{\partial k_i}\left(v_i e_i + v_p e_p + v_l e_l\right) = f_i - u_p\left(v_p - v_i\right) - u_l\left(v_l - v_p\right) \tag{2.15}$$

Finally, the optimality condition for the installed capacity of peaking plants k_p can also be obtained as:

$$\frac{\partial \mathrm{TC}}{\partial k_p} = 0 = f_p + \frac{\partial}{\partial k_p}\left(v_p e_p + v_l e_l\right) = f_p - u_l\left(v_l - v_p\right) \tag{2.16}$$

In summary, the three optimality conditions are:

$$\begin{aligned} f_b - u_i(v_i - v_b) - u_p\left(v_p - v_i\right) - u_l\left(v_l - v_p\right) &= 0 \\ f_i - u_p\left(v_p - v_i\right) - u_l\left(v_l - v_p\right) &= 0 \\ f_p - u_l\left(v_l - v_p\right) &= 0 \end{aligned} \tag{2.17}$$

By straightforward manipulation, these optimality conditions give rise to several pair-wise equalities that define the optimal capacity factors for each technology.

$$\begin{aligned} f_b + v_b u_i &= f_i + v_i u_i \\ f_i + v_i u_p &= f_p + v_p u_p \\ f_p + v_p u_l &= v_l u_l \end{aligned} \tag{2.18}$$

The pair-wise equalities define the breakeven number of hours (represented as a capacity factor u, the number of hours of operation per year divided by 8760 h in a year) after which one technology displaces another as the least-cost technology. For instance, for a utilisation factor lower than u_i it is less expensive to use intermediate plants than base load plants. The breakeven point between

intermediate and peaking plants is u_p. Lastly, it is less costly to leave some non-served energy than to invest in more peakers when the number of hours with some unmet demand is lower than u_l.

These three equalities are represented in Fig. 2.10. Their intersections yield the optimal values of the four utilisation factors u_b, u_i, u_p and u_l. The graphical solution provides an intuitive cost explanation for why the values of the installed capacities are optimal. The total cost in this model is the sum of all fixed investment and variable production costs. For each technology, the fixed cost f takes on a constant value and the variable cost is the product of the unit variable cost v and the number of hours of generation. In Fig. 2.10, the number of hours of generation is represented as a capacity factor. The cost function for each generation technology is linear with y-intercept f and slope v. A plot of the cost functions for all technologies versus their capacity factors reveals that no single technology is economically efficient for all hours of the year and all demand levels. Instead, technologies with high fixed costs and low variable costs are economically efficient at high capacity factors, and technologies with low fixed costs and high variable costs are economically efficient at lower capacity factors. In the cost function plot, the costs for each technology intersect when one technology displaces another (because the technology with greater fixed costs is able to distribute those costs across more hours of generation). These intersection points directly map to the load duration curve as illustrated, identifying the optimal installed capacities for the three technologies. This graphical solution to the centralised

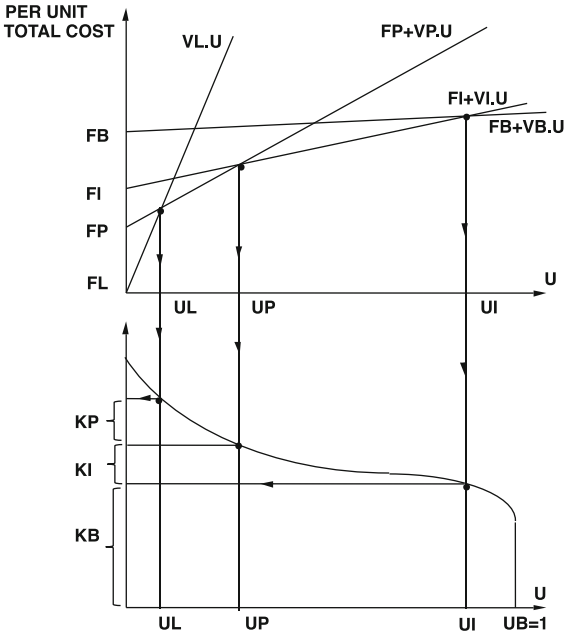


Fig. 2.10 Graphical solution to the centralised optimal investment problem for generation

capacity expansion and operation problem is commonly referred to as "screening curves" and was first developed by [6].

The monopolistic generation company

Assume now that all generators of the three technologies belong to the same monopolistic company. No other competitors can enter this business in a franchised territory to challenge the monopolist. We shall see later in this chapter that, in this situation, the monopolist can raise the price of electricity and, despite some expected reduction in demand, increase its benefits (margin of revenues over costs). What is the appropriate regulatory action that will prevent the monopolist from abusing its market power and achieve the same optimal level of social welfare as the level obtained under centralised planning of capacity expansion and operation?

It suffices with mandating the monopolistic generation company to perform so that the same objective function

$$\max(\text{totalSocialWelfare}) \approx \min(\text{supplyCost} + \text{nseCost})$$

will be minimised. In other words, the regulated generation company will be paid the cost of service (including a reasonable return on the invested capital) if it is able to prove to the regulatory authorities that the investments and the incurred operation costs have been necessary to supply electricity, with the objective of minimising the cost of supply plus the cost of the loss of utility to consumers because of curtailment.

Short- and long-run marginal costs

Decisions in power systems are made at different time scales, as described in Chap. 1. For the sake of simplicity, here we shall only consider two of them: (a) the *long run or long term*, with sufficient lead time to adapt the power system to changing conditions (like demand growth) by investing in new generation facilities and by dispatching these plants to minimize the operation cost; and (b) the *short run or short term*, when there is not enough time to add new investments and the system can only respond to changes in demand or other factors by dispatching the existing units in the most economic way.

When the decision-making process is reduced to just these two time ranges, Fig. 2.11 can help to understand how the short- and long-time ranges relate to each other. Given a demand D to be met, a centralised decision maker has to find a set of investments K in generation capacity that—operated optimally—result in the lowest cost of meeting the demand D. Given D and any set of investments K, the short-term operation problem consists of using the installed capacities in the best possible manner to minimise the operation cost $\text{VC}^*(K, D)$. In this problem, fuel F is the major cost component. The loss of utility to consumers because of non-served demand must also be considered an operation cost. The system planner must find the optimal set of investments K^* such that, when operated optimally, results in the minimum possible total cost of investment plus operation $\text{FC}(K^*) + \text{VC}^*(K^*, D)$. Capacity expansion and system operation are, therefore, two interrelated tasks hierarchically organised in time: K is optimised first,

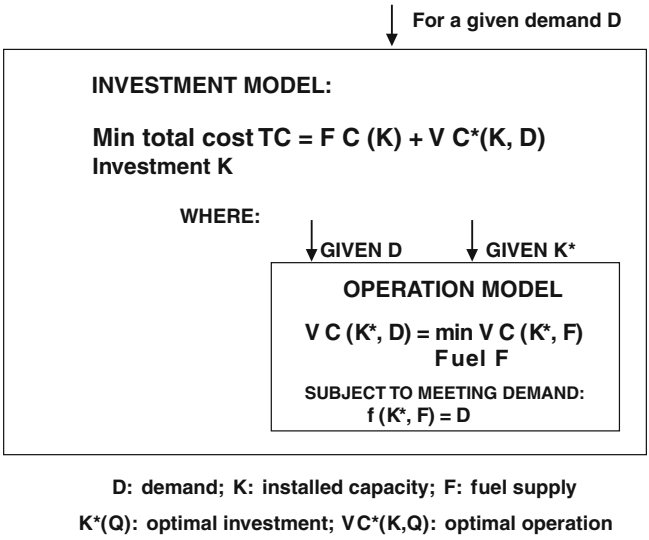


Fig. 2.11 Decision-making in the short and the long time ranges

based on estimates of the future operation of these assets. Once K has been built and enters into operation, then all available assets should be operated as economically as possible.

Assume now that the monopolistic generation firm with the three generation technologies is "perfectly adapted," meaning that it supplies demand at minimum cost, and therefore meets the three optimality conditions that were obtained previously. This will be possible if it is regulated according to the rule for generation monopolies that was established before. Now we shall define and compute the long-run and short-run marginal costs for this system.

The long-run marginal cost of generation is the increment in the production cost caused by a unit (one more kWh) increment of demand in the long-run context, i.e. when there is time to adapt the installed generation capacity to meet any new conditions—e.g. demand growth—in the most efficient way. An increment in demand in the long term can be specified in many different ways. Here, we choose a simple and meaningful definition: a uniform increment of 1 MW over the entire considered period of one year, as shown in Fig. 2.12, so that the entire load duration curve is shifted upwards uniformly.[17] The response of the power system in the long run must be consistent with the optimality conditions derived before (refer to Fig. 2.10): because the system has time to adapt the generation mix optimally, the utilisation factors will remain the same as before the increment took place. As Fig. 2.12 shows, all installed capacities will remain the same except for the base

[17] In this long term context a meaningful increment of demand would be any sustained raise over an extended period of time. The conclusions of this section do not depend on the adopted pattern of increase in demand.

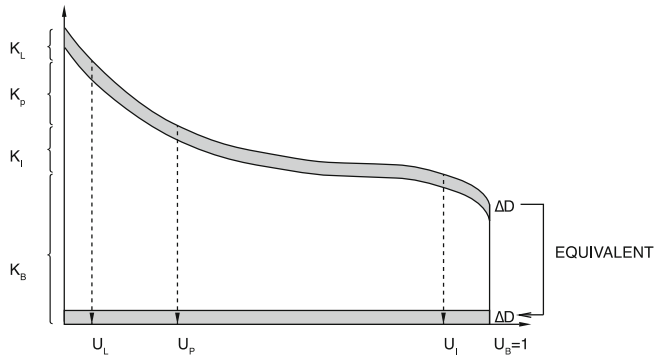


Fig. 2.12 Meeting a uniform increment of demand in the long run

load capacity, which increases by 1 MW. Therefore, the long run-marginal cost, i.e. the increment in total cost caused by the unit increment in demand, is:

$$\text{Long run marginal cost: } \frac{\partial \text{TC}}{\partial D} = f_b + (1) \cdot (v_b)$$

where the value of the utilisation factor, 1, represents the total duration of the considered period (e.g., 1 year).

The short-run marginal cost of generation at any given time is the increment in the production cost required to respond to a unit increment (one more kWh) of demand. Contrary to the long-term scenario, where new investments can modify the installed capacity of each technology, in the short term the system must produce electricity with its existing mix of generation resources and no time to adapt the generation mix to new conditions. The short-run marginal cost of generation changes with time as demand varies, and different plants with their respective variable costs become responsible for responding to these changes. In this example, because we are dealing with the entire load duration curve for 1 year, we will compute the aggregate short-run marginal costs for all hours in the year by considering a uniform demand increment of 1 MW. As Fig. 2.13 shows, each generation technology—base load, intermediate or peak—must respond by increasing production to meet the increment in demand, resulting in an overall increase in the operation cost:

$$\text{Short run marginal cost: } \frac{\partial \text{TC}}{\partial C} = u_l \cdot v_l + (u_p - u_l) \cdot v_p + (u_i - u_p) \cdot v_i + (1 - u_i) \cdot v_b$$

The expressions for the long- and short-run marginal costs appear to be quite different. However, if the generation mix is well adapted to demand to start with (i.e., if the three optimality conditions hold), then the expressions for the short-run and long-run marginal costs are equivalent. This equivalency can be easily proved (the manipulation of the equations only involves simple substitutions using the three optimality conditions):

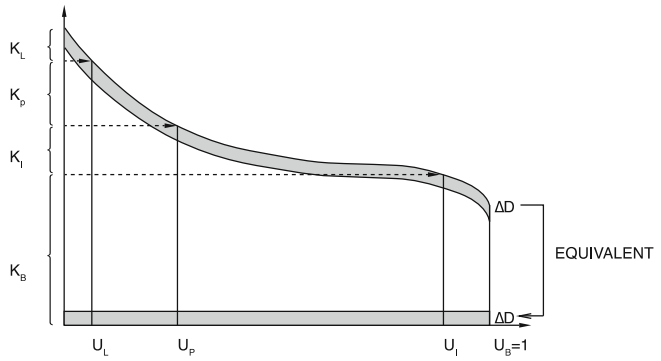


Fig. 2.13 Meeting a uniform increment of demand in the short run

$$\begin{aligned}\text{Long run marginal cost} &= f_b + v_b \\ &= f_i + (v_i - v_b) \cdot u_i + v_b \\ &= f_p + (v_p - v_i) \cdot u_p \cdot + (v_i - v_b) \cdot u_i + v_b \\ &= (v_l - v_p) u_l + (v_p - v_i) \cdot u_p \cdot + (v_i - v_b) \cdot u_i + v_b \\ &= u_l v_l + (u_p - u_l) \cdot v_p + (u_i - u_p) \cdot v_i + (1 - u_i) \cdot v_b \\ &= \text{Short run marginal cost}\end{aligned}$$

From our simple example, it can be concluded that for a monopolistic generation firm with centralised planning of operation and capacity expansion, if the firm's generation mix is well adapted and the firm itself is adequately regulated so that its private interest coincides with the maximisation of the global social welfare, then the long-term marginal cost of electricity is equal to the short-term marginal cost of electricity.

The competitive generation market

Assume now that the power system operates under a competitive market regulatory framework. Many generation firms exist, and each firm makes its own decisions about what technologies to invest in and how much capacity to build. Competition exists for all levels of demand (peak, intermediate and base load) and investors make their own decisions about what new generation capacity to build. The model assumes that no economies of scale exist, implying that the per-unit fixed supply cost does not depend on the volume of new investment (this is a reasonable assumption for power systems with more than several GW of demand). In this ideal model of quasi-perfect competition, many small supply firms (relative to the size of the entire market) sell electricity, and no firm has market power (i.e. they do not have the capability of modifying the market price to their benefit). No entry cost or other barriers to entry exist. Because each firm rationally attempts to maximise its profit, each firm will produce electricity when its variable cost of production is less than the price that it can sell electricity for. At any point in time, the price of electricity is set by the intersection between the supply and demand

curves, as we have seen previously. In this competitive environment, market prices that are too low will not attract new investment, while high market prices will bring new entrants until all extra profits[18] are exhausted.

In this model, generators will bid their variable costs and receive the market price that is set by the marginal generator at every moment in time. Let us consider investment in peaking plants. From inspection of Fig. 2.9, it can be inferred that the revenues, costs and net profits for the owner of a peaker plant with capacity A_p are as follows (note that the peaker plant only makes a profit when there is non-served energy and the market price is v_l; at all other times, the market price is equal to the variable cost of the plant and the plant does not make any net profit, so we shall ignore all times except when there is non-served energy):

$$\text{Revenues: } v_l \cdot u_l \cdot A_p$$

$$\text{Costs: } f_p \cdot A_p + v_l \cdot u_l \cdot A_p$$

$$\text{Net profit: } (v_l - v_p) \cdot u_l \cdot A_p - f_p \cdot A_p$$

Under the assumption of perfect competition, the net profit for each generation firm must be zero (i.e., no "extra profits" should exist beyond recovering all costs, including an adequate return on investment; otherwise, new competitors would enter the market and sell electricity for a slightly lower price). Note that perfect competition implies that the different types of plants have to recover their total costs from market prices: over-recovery attracts new investment until it disappears, while under-recovery deters investment until market prices rise enough to attract new investors. The perfect-competition equilibrium for a peaker plant of capacity A_p is characterised by the following equation:

$$(v_l - v_p) \cdot u_l \cdot A_p - f_p \cdot A_p = 0$$

This equation that characterises perfect competition in peaking plants is exactly equal to the optimality condition for investment in peaking plants in the centralised planning model. As the reader can easily check, the same relationship holds for intermediate and base load plants. Consequently, under the ideal conditions assumed in this simple model, the investment decisions for a power system under centralised planning and perfectly competitive market conditions are the same, and they should lead to the same generation mix.

In summary, under perfect competition the decisions of individual investors result in the same generation mix that would be achieved under centralised system planning. Additionally, under perfect competition, the short-term market prices

[18] It is important to understand that when we talk about the total production cost of a generator, this includes both the fixed and variable costs. Note that the fixed costs consist of both the annual depreciation (amortisation) costs plus a return on the invested capital, therefore providing an adequate remuneration to the investment. The phrases "extra profits" or "generator surpluses" refer to additional revenues beyond the reasonable remuneration to be expected for the capital invested in the plant. The existence of "extra profits" in a power system attracts new investment.

obtained from the intersection of supply and demand at every moment in time will pay for the complete fixed and variable costs of all generation plants.

The non-regulated monopoly

Finally, we shall examine what happens if a generation monopoly is left unregulated. Assume that, because of some technological, economic or regulatory constraints, only one company has the possibility of investing in peaking units. This constraint results in a de facto monopolistic situation for this niche of electricity production, and a single company owns the entire investment capacity k_p in peaking plants. What is the best strategy for this unregulated company so that its profits are maximised?

The monopolistic company is free to decide how much to invest so that its profit is maximised.[19] We have just determined that the net profit for a capacity k_p of peaking plants is:

$$\text{Net profit: } \mathrm{NP} = (v_l - v_p) \cdot u_l \cdot k_p - f_p \cdot k_p$$

The condition of optimality for the monopolistic generation firm is[20]:

$$\frac{\partial \mathrm{NP}}{\partial k_p} = 0 = (v_l - v_p) \cdot u_{\mathrm{ln}\,ew} + (v_l - v_p) \cdot k_p \cdot \frac{\partial u_l}{\partial k_p} - f_p$$

$$f_p = (v_l - v_p) \cdot u_{\mathrm{ln}\,ew} + (v_l - v_p) \cdot k_p \cdot \frac{\partial u_l}{\partial k_p}$$

This equation must now be compared to the optimality condition under centralised planning and perfect competition:

$$f_p = (v_l - v_p) \cdot u_{\mathrm{l}}$$

Because the last term in the equilibrium equation for the monopoly is negative (u_{l} decreases when k_p increases), then $u_{\mathrm{ln\,ew}} > u_l$ implying that the optimal investment in k_p for the monopolistic firm will be less than the investment that is optimal from a global social welfare viewpoint, leading to more hours with non-served energy and extra profits for the monopolist. A more complete analysis of monopolies and their regulation will be presented later in this chapter.

[19] The monopolistic generation company can try to maximise its profit by choosing the total amount of investment in peaking plants k_p and also by bidding strategically (above the variable cost) in the short-term market. It is assumed here that the short-term behaviour of the monopolist in the market is carefully monitored by the regulator, so that it has to bid precisely its variable cost. Thus the only decision variable left to the company is the level of investment k_p.

[20] Note that u_l now depends on the value of the independent variable k_p.

2.3.3.2 The General Theory

Classical microeconomic theory [11], as presented in most textbooks and as it has been probably learned by the readers of this book with some background in economics, does a poor job in explaining the relationship between operation and planning, costs and prices in the short and long terms, in a power systems context. The major shortcomings are:

- The distinction between the decision-making processes in the short and long terms is often blurred.
- The same thing happens with marginal costs and prices. As shown in the simple model, in power systems it is critical to show the role of short-run (spot) prices in the remuneration of the generation plants.
- The multiple levels of demand within a considered time period, which is of essence to determine the generation mix, are not properly captured since a single demand value is assumed for the operation period.
- The multiple existing constraints in operation and planning are not represented and their impact on generation dispatch, prices and cost recovery is missed.

All this does not mean that microeconomic theory cannot be applied to the power sector or that the general microeconomic principles do not hold for electricity. They do and it is possible to develop a version of microeconomic theory well adapted to the actual characteristics of power systems. The groundwork for this was laid out by Fred Schweppe and his colleagues at MIT during the 1980s [1, 8] and it has been extended for the specific comparison between centralised and market oriented approaches to electricity generation management in [4] and [5].

A full mathematical presentation of the latter documents would be too detailed for this book. The simple example in the preceding section has introduced all the basic concepts using just three generation technologies and a simplified version of the capacity expansion and the decision-making processes at operation level. It is important to understand well this example before starting the more detailed discussion in this section, since the simple example allows picturing in concrete terms what is really going on. In this section a complete discussion of all relevant aspects will be provided, but avoiding the mathematical formulation, which could be found in the references mentioned before. The standard general-purpose microeconomic presentation of the same principles can be found in Annex A to this chapter. As it can be easily verified, the basic general principles are the same, but many critical issues are missed in the general-purpose exposition.

Objective and method of this analysis

The objective of the more ambitious power system formulation in [4] and [5] is to examine how the outcome of the economically rational behaviour of generators and consumers in a *perfectly competitive market*, both in the short and the long terms, may differ from the decisions that an ideal all powerful and benevolent central planner would adopt in the pursuit of maximising the global social welfare of all market agents, i.e. producers and consumers of electricity. In the same way,

the behaviour of all generation being treated as a *regulated monopoly* will be compared to the ideal centralised management model.

In the first case (perfectly competitive market) the method used in these references consists of looking for the optimal economic signals that generators and consumers in a perfectly competitive market must receive so that their behaviour, even in the presence of planning and operation constraints for the generators, is consistent with the goal of a correct regulatory policy: the maximisation of global net social benefit. This exercise results in optimal prices for consumers and generators: not a single price in general, but different prices for energy, operating reserves and contribution to system adequacy. Then these prices are checked to identify and understand mismatches between generators' revenues and costs, and their relationship with the planning and operation constraints. Finally, the traditional concepts of short- and long-term marginal costs, STMC and LTMC, are reviewed and related to the optimal prices for consumers.

In the second case (regulated monopoly) the objective is to find the rules that have to be imposed to the regulated generators so that the outcome also coincides with that of the ideal centralised model.

The common power system model

A streamlined conceptual mathematical model of the planning and operation functions in a power system has been used for this purpose, as described in [4] and [5]. The model considers three (extending the two levels in the simple model) nested time ranges for decision making: capacity expansion planning (time horizon: up to 10 or 15 years, for instance; this is the long term), operation planning (time horizon: 1 week up to1 or 2 years) and dispatch (hourly up to 1 day or even 1 week; this is short term), as well as different types of generation and planning constraints, which cover the major prototypical cases encountered in reality. Note that, for each generation technology, the model distinguishes between *installed capacity* (decision variable in the long term), *connected capacity* that is ready to function to produce energy or to provide operating reserves (decision variable in operation planning) and actual *generation output* (decision variable in the short term).

All network aspects have been ignored in the model. Operation reserves are roughly represented by the distinction between connected capacity and actual generation output. Decision variables are assumed to be continuous. Non-served energy is modelled as one more generation technology, with very high variable cost and no fixed cost: if the existing generation could not meet all the demand, the rest would be covered by non-served energy, at a variable cost equal to the estimated cost of non-served energy.

Uncertainty in demand, plant unavailability and other external factors (like temperature) that may affect the consumers' utility function are conceptually modelled by considering any number of scenarios that cover all possibilities. It is important to understand the sequence in the nested decision-making process and what can and cannot be decided at each time range. In the short run both the installed and the connected capacities are given, there is certainty concerning plant availability and demand level and only the output of the connected generators can

be modified. The operation planning decision of connecting generation units is made before the uncertain availability of the plants or the actual demand level is known. Connected generation cannot exceed the amount of installed capacity. Only in the long run may the installed capacities be modified and a unique decision to install new capacity is made for the complete set of uncertain scenarios.

This power system model has been separately formulated for three distinct contexts: (a) the ideal case of an all-powerful centralised manager as the single decision maker, both for supply and demand, (b) a fully competitive generation market and (c) the traditional regulated monopoly framework. The optimal behaviour of the decision makers in each context has been modelled as a mathematical optimization problem.[21] Comparison of the optimality conditions for the three cases yields: (a) the expressions of the optimal prices for consumers, (b) the formulation of the planning and operation problems under traditional regulation and (c) the expressions of the optimal prices for generators in a competitive market.

There are some differences in the constraints that have to be included in the model for each one of the three cases. Some constraints are obviously common to all of them, like, for instance, the minimum technical outputs of generators, or the maximum available resources for future investment in any given technology, hydro, for instance. However, other constraints reflect specific objectives of the centralised manager and they only should appear in its model, such as a mandatory minimum consumption of a given fuel; or a reliability target, for instance in the form of a mandatory margin of installed generation capacity over peak demand in the centralised management case. Or an operational security target, in the form of a minimum volume of operating reserves at all times. Since in a competitive market the generators are free agents that are not subject to these mandates, these constraints are not included in the competitive market model, and the same fuel consumption, reliability or security targets have to be achieved by means of some economic incentive, like an extra component of the electricity price.

The ideal centralised management model

The objective of the ideal centralised manager is to maximise the global net social benefit associated to the supply and consumption of electricity, over a multiyear horizon. In this ideal model it is assumed that the regulator directly controls all the decision variables: investment, connection and production of the generators and investment and consumption of demand. The model explicitly includes; (a) a

[21] Mathematical details: For each one of the three cases, the Lagrangian function (where all constraints are incorporated to the original objective function by means of the respective Lagrangian multipliers) is formulated. The optimality conditions for the decision makers within each model are directly obtained by taking the corresponding derivatives of the Lagrangian function with respect to each one of the decision variables. For the sake of mathematical simplicity all decision variables have been assumed to be continuous, what does not affect the conceptual results of the model. The results do not apply to very small power systems, where the discreteness of the generation investments and typical high horizontal concentration levels violate the assumptions of this analysis. This issue has been examined in [7].

reliability constraint as a minimum requirement on the surplus of installed capacity over demand in the long run and (b) a security constraint as a minimum requirement of the volume of connected capacity over demand at operation level. Here the objective function to be maximised is the aggregated surplus of consumers and generators.

Note that the reliability and the security constraints may be considered to be unnecessary in the model, since the probability of incurring in costs of non-served energy should result by itself in adequate levels of installed capacity and operating reserves. The reason to include these constraints in the model is the universal practice of centralised planners, regulators and system operators: they typically prefer to mandate generous security and reliability margins, regardless of the value of those that would be strictly justified economically. Lack of electricity supply is a too politically sensitive matter, with more implications than the purely economic ones, and this is reflected in the actual requirements normally established by system operators and regulators.

The centralised model may include other constraints, which are meant to achieve specific objectives in generation investment or production. Characteristic examples are targets of deployment of installed capacity or production of generation with renewable energy sources, or minimum levels of consumption of indigenous resources such as local peat or coal, as well as limits on emissions of CO2, NOx or SO2.

The perfectly competitive market model

Now the model is a collection of individual optimization problems, one for each consumer and generator. Each agent independently invests in new assets and operates the existing ones at any moment in time, with the purpose of maximising its individual economic surplus. The optimality conditions are the collection of the ones corresponding to each one of the agents' models.

The reliability and the security constraints do not appear in the formulation, as the regulator cannot impose this kind of conditions to the agents in a market. Instead, the same objectives will have to be achieved by suitable economic signals: some remuneration for the provision of reserves and for the contribution to system reliability of the installed capacity.

Note that the market by itself will provide some amount of reserves and margin of installed capacity over peak demand, since there are business opportunities for any excess generation to be called to operate when other more competitive generation units fail. The issue here is whether the generation margins that the market provides by itself are considered to be sufficient by the regulator.

Other economic signals will be needed to remedy the fact that markets may result in undesirable levels of investment and production, in excess or below of what would be justified economically under the viewpoint of social welfare optimization. As shown later in this chapter, this may happen because there are market failures; a characteristic example is to ignore certain externalities of electricity generation, such as the economic impact of carbon emissions or other environmental impacts, thus creating a tilted paying field against cleaner

technologies. In the absence of additional economic signals, markets will ignore policy objectives of utilisation of indigenous energy resources or penetration of renewables.

The traditional regulated monopoly model

Now it is assumed that a single company owns all generation in the system. This model is very similar to the ideal centralised management model, but it only includes the supply side and the firm's objective is to try to maximise its surplus (revenues from selling electricity minus its production costs), not the global social welfare. Therefore, the regulator may want to impose some conditions to align the legitimate objectives of the company with global social welfare maximisation. The formulation of the consumers' behaviour is the same one as in the competitive market's model, but now consumers react to regulated charges instead of market prices.

In the absence of regulation, a generation monopoly would behave as predicted by the classical microeconomic theory (see Sect. 2.4 below and the application of the simple model in 2.3.3.1 to a monopolistic situation): underinvestment in generation capacity in the long run and withholding of production in the short run, so that prices will increase and generation surplus (revenues of selling electricity minus production costs) is maximised.

Finding: Conditions for the same generation mix under competitive market and ideal centralised conditions. Optimal market prices

Comparison of the sets of optimality conditions for the ideal centralised management model and the competitive market model shows that they are identical, and therefore result in the same solution of generation expansion (the generation mix) and system operation (generation outputs and operating reserves) and achieve the same optimal social welfare IFF (i.e. if and only if) some specific prices are used to remunerate the energy output, the operating reserves and the available installed capacity that is provided by each generator. Next, the nature of these prices is examined.

In the first place we have the *market price for energy*, which happens to be equal to the variable cost of the "last generation unit in the economic merit order" (the "marginal unit") of energy production.[22] All energy produced at any given instant of time must receive this marginal price, which must be paid by the consumers of this electricity. Infra-marginal plants, i.e. all generators except for

[22] The model does not include some non-linearities that occur in the production function of actual generators: start-up costs, minimum production levels and changes in efficiency with the output of the plants. There is no generalised consensus on how to account for these nonlinear factors and different market designs have adopted a diversity of schemes to compute the energy market prices. For the purposes of this chapter, and because of the simplifying assumptions of the model, it suffices to know that, at a given moment in time, all produced energy must be remunerated at the variable production cost of the generator at the margin. A review of the main implementation approaches to address this issue in present wholesale electricity markets can be found in Sect. 7.4.3.2.

the one(s) at the margin, receive remuneration above their variable production costs. The difference of revenue minus cost is termed "infra-marginal rent" and is necessary to recover the incurred investment costs.

This energy market price is all that is needed to remunerate the generators if the regulator is not seeking any specific objectives, i.e. those that in the centralised management model have to be introduced as regulatory constraints, such as those on security, reliability or environmental concerns.

When the system operator wants a certain mandatory level of operating reserves that corresponds to the one achieved by centralised planning under the security constraint, the market price must include an additional component. This component has the format of a "*reserve capacity payment*", per MW of connected capacity, to any generator g with a capacity K_g of whatever technology. Actual wholesale electricity markets frequently have ad hoc markets for operating reserves, where the price of these services is competitively determined.

When the regulator wants the market to have a reliability level that corresponds to the one achieved by centralised planning under the reliability constraint, the market price must include an additional component. This component has the format of a "*capacity payment*", per MW of installed capacity, to any generator g with a capacity K_g of whatever technology, with the following value:

$$\left(\frac{\mathrm{dReliability}}{\mathrm{dKg}} \Big/ \frac{\mathrm{dReliability}}{\mathrm{dKpeak}}\right) \cdot \left(\frac{\mathrm{dICpeak}}{\mathrm{dKpeak}} + \frac{\mathrm{dOPex}}{\mathrm{dKpeak}}\right) \tag{2.19}$$

where Kpeak is the peaking technology in the considered system (the one with lowest capital cost per installed MW), ICpeak is the capital cost of the peaking technology and OPex is the generation operation cost of the system. Note that the derivative of OPex with respect to Kpeak is negative, as more installed capacity should result in lower operation costs.

To understand this expression, we start by applying it to the peaking technology. Then the first factor is just 1.0, and the second is equal to the net cost (economic loss) for the system of 1 MW of additional investment in peaking plants, i.e., the cost of investment of 1 MW of peaking capacity minus the corresponding savings in operation costs. When the reliability constraint is active, there is more installed generation capacity than what should be optimally installed in the absence of the constraint, and one more MW of installed capacity brings more costs than benefits.

In a situation where the regulator deliberately seeks some "extra installed generation capacity", the "capacity payment" for a peaking unit seeks to compensate the deficit in the remuneration of this kind of plants under market conditions. This is necessary to encourage investment so that an excess of installed capacity over the strict economic optimum can be achieved. In the power system model that has been adopted here, the impact of 1 MW of peaking plant investment on the operation cost OPex of the system is equal to the reduction in the hours of non served energy (HNSE) multiplied by the estimated cost of non served energy (CNSE) and by the average availability of the peaking unit ($1 - \mathrm{FOPpeak}$),

where FOPpeak is the forced outage probability of the plant, i.e. the probability that the unit is out of order. Therefore, Eq. (2.13) of the capacity payment for a peaking unit becomes:

$$\text{ICpeak}_{\text{perMW}} - (\text{CNSE} - \text{VCpeak}) \cdot \text{HNSE} \cdot (1 - \text{FOPpeak}) \tag{2.20}$$

Note that HNSE represents the regulated target reliability level, e.g. 10 h per year when some electricity demand cannot be served because of lack of available generation capacity. The regulator wants generation investment to increase until this reliability level is reached.

Note also that the second term in Eq. (2.20) is also the expected surplus (revenue minus variable costs) for a peaking unit under market conditions. In the absence of any extra "capacity payment", investors in peaking units will add new capacity only while the expected surplus exceeds the €/MW investment cost $\text{ICpeak}_{\text{perMW}}$. This corresponds to a reliability level of

$$\text{HNSE} = \text{ICpeak}_{\text{perMW}} / (\text{CNSE} - \text{VCpeak}) \cdot (1 - \text{FOPpeak}) \tag{2.21}$$

It the regulator wants a target reliability level that is stricter (i.e. lower HNSE) than what is economically justified, then a capacity payment to the generator will be necessary, according to Eq. (2.20).

When applied to any technology g other than a peaking unit, the second factor in Eq. (2.13)—which is more easily computed as (2.20)—remains the same. The first factor in (2.13) measures the relative contribution to the reliability of the system of the considered technology g with respect to the peaking technology. In practice, and for conventional plants, the first factor in (2.13) is equal to

$$(1 - \text{FOP}g)/(1 - \text{FOPpeak}) \tag{2.22}$$

Therefore, for any technology g, the expression (2.22) can be computed in practice as the product of the two elements in Eqs. (2.16) and (2.20).

The implications of the implementation of a capacity payment in a competitive market are multiple. It obviously implies an extra cost for consumers, but also the expectation of more installed generation capacity and, therefore, better reliability. If investors behave with economic rationality there will be more generation investment and the desired reliability target will be achieved. The impact on cost recovery for the individual generators will be discussed later in this section, but we can advance that each generator is expected to recover its total costs, no more or less.

Other constraints are possible in the ideal centralised management model, corresponding to other regulatory objectives or actual hard constraints. Details can be found in [4] and [5]. In all cases these objectives can be met by a suitable component of the price for energy, capacity or reserves. Examples of these policy objectives or constraints are: exhaustion of some generation resources, for instance adequate hydro sites; global cap on carbon emissions for the entire power sector; minimum production quota of an indigenous fuel; or minimum level of deployment of renewable generation, in terms of capacity or production. Failure of the

regulator to take these results into account in actual electricity markets will typically result in windfall profits or in economic losses for the generators involved.

In summary, the essential finding of this section should not be lost in the details: a perfectly competitive electricity market can deliver optimal global social welfare if the generators are subject to a set of optimal prices.[23] The most important of these prices is the energy market price, which is equal to the variable cost of the marginal generator. Other prices that modify the basic energy price or remunerate connected or installed generation capacity are also necessary for the market to satisfy additional objectives—operational security, investment adequacy, environmental targets, promotion of certain fuels or technologies—that the regulator wants to achieve.

Finding: Conditions for the same generation mix under regulated monopoly and ideal centralised conditions. Optimal marginal costs.

Comparison of the optimality conditions of the regulated monopoly model and the ideal centralised management model reveals what is missing in the former one so that both yield the same results in operation and capacity expansion. Both models become identical if the following two conditions are met:

- The monopolistic generation company must include the cost of non-served energy to consumers as an additional cost item within its supply cost function. Obviously, compliance with this condition has to be subject to regulatory supervision, since the generation companies will not willingly adopt it by themselves. This requirement may be implemented in a multiplicity of ways by the regulator. The most common one has been to mandate minimum reliability levels of electricity supply and to subject the reimbursement of the costs of capacity expansion and operation to regulatory oversight.
- The consumers are subject to electricity charges that reflect the short-term marginal cost of electricity generation. This short-term marginal cost (see the definition below) coincides with the short-term price of a perfectly competitive market that has the same technology mix.[24]

[23] There are still some other complexities that have been ignored in this discussion and that may create a divergence between the competitive and the centralised approaches in practice. Under centralised conditions, once a generation investment has been made, its cost is sunk and only the variable costs will matter in future decisions; therefore the existing asset will continue in use, even if new more competitive technologies appear during its lifetime. On the other hand, in a competitive regime investors will deploy any new more competitive technology as soon as it is available, without waiting for the existing one to finish its economic life. Thanks are given to Dr. Carlos Vázquez for this clever observation.

[24] In a competitive market the consumers are subject to the wholesale electricity prices plus any other regulated charges derived from the use of the networks and other concepts. In a regulated monopoly the regulator determines the consumer tariffs based on the costs of generation plus any other regulated charges (in principle the same ones as in the market situation). The comparative analysis of optimality conditions of the ideal centralised and regulated monopoly models shows that the costs of generation that the regulator has to include in the charges to the consumers of the regulated monopoly to achieve optimal social welfare should be the short-run marginal costs.

We have seen in the simple case example that the *short-run marginal cost of generation* at any given time is the increment in the production cost to respond to a unit (one more kWh) increment of demand. Since this happens in the short run (e.g. in 1 hour, or less), the installed generation capacity must remain constant and the increment in demand must be met with the available units at the considered time. As said before, under perfect competition conditions this short-term marginal cost should be equal to the short-term market price of generation. Therefore the consumers' payments must equal the generators' required marginal revenues for their three products: energy, operating reserves (only if the security constraint is active) and available installed capacity (only if the reliability constraint is active). Plus any other adjustments to these prices due to any additional energy policies that the regulatory authority may want to achieve by means of economic signals.

We have also seen that it is possible to define a *long-run marginal cost of generation* as the increment in the production cost to respond to a unit (one more kWh) increment of demand in the long-run context. What is meant by "a long run context" here? Firstly, there is enough time for the generation company to include in the response any changes in the installed capacity that may be best to meet the new demand at minimum cost. Therefore the response now will be a combination of adaptation of the installed capacity and adjustment of the output level. Second, in the long term a meaningful increment of demand would be a sustained increment over an extended period of time, rather than just an increment in just 1 hour or less. Third, as the future is uncertain, the long-run marginal cost must be an expected value over all the considered possible scenarios.

One finding of the examination of the model is that *the values of the short- and long-run marginal costs are equal* if the following conditions hold: (a) the generation system is perfectly adapted to the demand, i.e. it is the optimal generation mix that minimises the total production cost; (b) the short-term marginal cost is averaged over the considered uncertain scenarios and the interval of time for which the long-run marginal cost has been defined; (c) there is no reliability constraint that forces investment costs in the long run when demand grows (since there are no short-run costs corresponding to this constraint).

Finding: Recovery of incurred generation costs under competitive market conditions.

In the considered power system model the total generation supply cost comprises investment, energy production and operating reserves costs. These costs are fully recovered from the marginal payments to generators for the three concepts (including the above-mentioned adjustments due to a possible number of policy or physical constraints, if required). Note that there is not a one-to-one correspondence between prices (energy, reserves, capacity) and costs (fixed and variable costs) components; for instance, the marginal price of generated energy pays for the total generation costs—fixed and variable costs—when the security and reliability constraints are not active and no generation price adjustments are needed.

The existence of active constraints on the future installed capacity (e.g., exhaustion of new investment resources, target levels of penetration of some technologies) creates a positive or negative mismatch between revenues and costs of the generators. A physical or policy constraint that limits new investment will result in less investment than what is theoretically optimal and cost over-recovery for the existing generators from application of market prices. Depending on the particular case, the regulator may consider convenient to allow that the market prices may incorporate adjustments to compensate for the mismatches.

Another cause of mismatch is the existence of nonlinearities in the generation cost functions (economies of scale, start up costs, heat rates, etc.), which have been assumed linear throughout this study. In particular, the results from this analysis are not valid in small and isolated or weakly interconnected power systems, where economies of scale in generation still exist.

These cost recovery properties hold for generation technologies that are "perfectly adapted in capacity" (i.e. the optimal capacity for that technology that would result from the centralised management model, including all the corresponding constraints and the installed capacity of the remaining technologies), even if other technologies are not well adapted. Optimal operation is also assumed, i.e. generation dispatch in accordance with the minimization of the short-term supply cost.

How does a perfectly competitive market evolve to reach a long-term equilibrium where this properties hold? A stable equilibrium in a power sector is a theoretical concept, since it will be never reached: demand, costs, technology and regulation continuously change. However, the rational behaviour of the agents of the market continuously makes the power system evolve towards this unattainable equilibrium condition that has been described above. This is how it theoretically happens:

- In the long run, all the companies competing on a market, where entry and exit are free and all players have access to the same technological opportunities, eventually exhibit similar cost structures.
- When companies earn extra profits,[25] new entrants are attracted to the industry by its high profitability.
- Companies posting losses exit the market under the pressure of low short-term prices. Note that, in the presence of sunk investment costs (such as in most power plants), the firm should reasonably keep the plant operative for as long as its expected short-term revenues exceed its expected operating costs.

Consequently, in the long term, the number and cost structure of the companies ideally evolve towards the equilibrium situation that has been described in this section.

[25] Note that the total cost of production must include a return on capital, so by "extra profit" is meant a return on invested capital higher than normally expected for a company of these characteristics.

The theoretical results presented here have been verified quantitatively with large computer models with more realistic representations of power systems. One example is [5].

Finding: Recovery of incurred generation costs under regulated monopoly conditions.
Cost recovery is in principle guaranteed in a regulated monopoly. The issue here is whether the short-run marginal costs of generation, which constitute perfect short-term economic signals for consumers, will be able to recover the incurred total generation costs.

Since the short-run marginal generation costs are equal to the short-run prices in a perfectly competitive market, the same properties of cost recovery that have been observed in competitive markets will apply here, including the need for adjustments to complete the cost recovery process. This issue of "revenue reconciliation" is examined in detail in the pioneer work of Fred Schweppe's group at MIT [8].

Economic signals to consumers are equally important under monopolistic or market conditions, since their purpose is to elicit an efficient demand response. The material that has been presented here is useful in the design of the generation component of efficient electricity tariffs, as discussed in detail in Chap. 8.

2.4 Monopolies

As mentioned before, regulated monopolies have been the preferred regulatory regime for electric utilities during the second half of the last century in most countries of the world. It is still the regulatory framework of choice in many power systems and it is almost universally used today for the network activities: distribution and transmission of electricity.

The regulated monopoly framework is necessary when reasonable conditions for competition do not exist. This may happen because of the nature of the activity, as it is the case with distribution and transmission electricity networks, since they present strong economies of scale that require, for the sake of efficiency, that a single company provide the service. Or it may happen in electricity generation because the structure of the potential market is not suitable for competition: either because the power system is too small to accommodate multiple firms with several generation technologies, or perhaps because the structure of the sector is too concentrated to sustain competition, e.g. too few and too large firms.

In addition, it is possible that the regulatory authority, at the highest level, decides that a regulated monopoly is the regime of choice for the generation activity, even if the system size and the sector structure are adequate to sustain a competitive market.

The regulation of monopolies will be addressed later in the book in Chap. 4. The general-purpose analysis of the economically rational behaviour of an

unregulated monopoly is presented in Annex A of this chapter. Here we simply summarise the findings to understand the need for regulation.

A monopolistic electricity company[26] has exclusive control of production or price (since demand will respond to any of these two outputs from the generation company) in a given market. The monopolistic condition may occur for the entire demand or for a layer of it (base, intermediate or peak). A monopolist operating on a market in the absence of regulation will try to maximise its earnings by raising the price. In the short term, in which neither investment decisions nor the possible entry of new competitors are considered, the monopolist will reduce the output, since this will raise the market price of electricity according to the demand curve, seeking to maximise its profit (i.e. the difference between the revenues from selling electricity and the production costs). How much reduction? The monopolist increases its earnings as long as the output of one extra unit raises revenues (marginal revenue) more than the associated costs (marginal cost). Consequently, maximum profit is earned where marginal cost equals marginal revenue. This increase in the producer's surplus comes at the expense of a greater reduction in consumer's surplus. In other words, a monopolistic situation produces lower net social welfare than a perfectly competitive market.

In the long term, the monopolist decides on new investments. Following the same reasoning, the monopolist maximises its profit with a level of investment lower than the competitive optimum. This underinvestment results in the same negative outcome of a loss of net social benefit.

Therefore, from the standpoint of economic efficiency, unregulated monopolies should not exist. If a power system chooses a monopolistic structure for its power sector (or for any of the major activities: generation or networks) the operation and investment of the monopolistic company must be regulated to prevent a loss of global social welfare. We have seen in Sect. 2.3.3 how to maximise global social welfare with regulated monopolies.

2.5 Market Structure, Concentration and Market Power

This section deals with market structure, identifying economies of scale and entry barriers as the factors that primarily determine the number of competitors in a given industry. This review is followed by a discussion of contestable markets and vertical integration and lastly an analysis of market power and related factors.

[26] It is easier to follow this discussion from the viewpoint of a vertically integrated electric utility or a purely generation company. However, the discussion also applies to the activities of transmission and distribution.

2.5.1 Determinants of Market Structure

The number of competitors on a market depends on the entry conditions and possible technical or regulatory barriers, i.e. on how easy or difficult it is to enter the market. Economies of scale are a determinant for the number of competitors on a given market. Strong economies of scale (such as in electricity transmission) result in a natural monopoly. If, once the economies of scale are exhausted, total demand only accommodates the existence of very few companies (two, or maybe three, as in some small electricity systems), the result will be an oligopolistic market with a small number of producers. If the market can be freely entered and economies of scale are not an issue, the number of competitors tends to be high in the long term, constituting a competitive market characterised by slender cost-price margins. Moreover, easy market access is perceived by incumbent competitors as a real threat, further encouraging competitive prices.

The existence of entry barriers favours the existence of highly concentrated markets. Several types of entry barriers can be distinguished: legal constraints, long entry delays, cost disadvantages, uneven access to technology and information, regulatory uncertainty or exit hurdles. A (non-exhaustive) list for electric energy systems is set out below.

- Administrative concessions or licences to operate on the market or conduct certain activities do apply to electricity generation, transmission and distribution companies. Complex bureaucratic procedures to obtain a licence or NIMBY[27] related political and social obstacles are frequently encountered in the electricity industry.
- Environmental regulations and legislation on access to natural resources such as water may be a high entry barrier for the construction of hydroelectric plants.
- In all tightly regulated industries, regulation itself constitutes a significant entry barrier. When regulations are unclear or where new entrant rights with respect to incumbent rights are uncertainly defined, new competitors are charged a higher risk premium on the borrowings they need to raise the huge amounts of capital required.
- The long lead times inherent in building and commissioning large-scale electric power facilities such as hydro or nuclear plants likewise constitute a barrier to entering the electricity industry.
- The experience acquired by companies already operating on a market affords them a certain advantage in terms of cost and access to technology and information with respect to potential new entrants. In the capital-intensive electricity industry, however, where personnel costs account for a fairly small proportion

[27] NIMBY stands for the phrase not in my back yard.

of the total, new entrants can hire experienced personnel with no significant increase in their average costs.[28]

- Investments in electrical infrastructures should generally be considered sunk costs, since they cannot be easily applied to other purpose or moved to other location. Since sunk costs cannot be recovered when exiting a market, new entrants have to assume the risk of failure and the inability to recover all or a significant part of their investment.

Since the electricity industry exhibits the above characteristics, it has traditionally been regarded as a monopoly that must be regulated. This historical context is the starting point for deregulated electricity markets and explains why such markets generally comprise only a few supply-side agents if structural issues are not addressed.

In addition to the aforementioned, other barriers to enter highly concentrated markets may be created by strategic behaviour on the part of incumbents, geared to lowering new entrants' expected earnings. The item below on strategic competition discusses two examples of such behaviour in electricity markets: limit pricing and investment in cost-reducing capital.

One natural aim of regulation is to attempt to eliminate or lower entry barriers, which hamper competition and favour the existence of higher than break-even point prices that are nonetheless unable to attract new competitors. The existence of incumbent competitors that are more efficient than the new entrants is not always detrimental, even though it may hamper entry. Consequently, the analysis of situations where efficiency and competition clash should always be based on a comparative study of net social benefit.

A simple case example

Before starting a review of the theory of monopolies, oligopolies and market power, it will be illustrated here, in a very simple case example, one of the diverse forms in which a market agent can exercise market power: short-term price manipulation. The other basic approach—withholding new investment—was already illustrated in another simple case example in Sect. 2.3.3.1.

Figure 2.14 represents supply and demand in a power system at a given time. The supply curve shows the staircase of prices and quantities for the several generators in the system. Assume that one of the generation firms owns plants A, D and F, with F being the current marginal plant. By removal of plant F (for instance bidding much higher than its actual variable cost, or pretending the plant is not available) the firm forces the market price to raise from P_F to P_G, therefore increasing the marginal rents of its plants A and D. Given that plant F was not earning any inframarginal rents, the firm loses nothing by the removal of plant F,

[28] Moreover, new entrants could also capitalise on technological progress to acquire a competitive edge by building units that are more efficient than the incumbents' facilities. In light of the long amortisation periods, existing facilities must be kept in operation for many years with lower returns except where further investments are made to implement new technological advances.

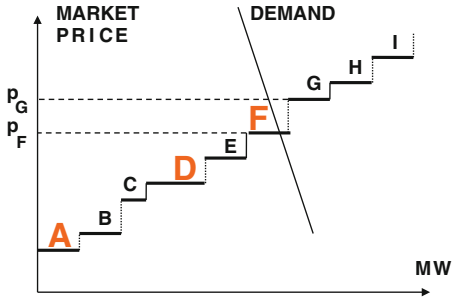


Fig. 2.14 Case example of exercise of market power

but benefits from the higher market price with A and D. This is an example of exercise of market power, since the firm has modified prices from the competitive level for its own benefit.

Assume now that another firm owns only plant E. This firm could remove plant E and also raise the market price. But it will not benefit from this action, since the company does not own any inframarginal plants. Unless this firm somehow colludes with the previous one and they somehow share the benefits, the behaviour of the owner of plant E is foolish. It does not have market power, since it cannot modify the market price in its own benefit.

2.5.2 Contestable Markets

The existence of a large number of competitors and the real threat of potential future competitors are two natural developments that drive market agents to be more efficient and competitive. Indeed, a number of real-life concentrated markets can be identified in which prices are competitive due to incumbents' fear of new competitors. These are known as contestable markets.

The three conditions for perfect contestable markets defined by economic theory are closely related to the entry conditions described in the preceding item: new entrants must be at no disadvantage compared to incumbents, their sunk costs must be nil and their entry must not be delayed. In the power sector the satisfaction of all three conditions (the second one, in particular) cannot happen in practice, although, for a market to be regarded as contestable, the most important condition is the ease with which new entrants can enter and exit, for this is perceived by incumbent companies as a real threat.

As most electricity industry activities are capital intensive, market exit entails the sale of very costly assets. This, in conjunction with the possible lack of liquidity attributable to the short number of potential buyers, means that electricity generation is characterised by sizeable sunk costs. Electricity retailing is often cited as an example of a possible contestable market because entry is both speedy and easy and exit is not contingent upon the ability to sell physical assets.

2.5.3 Vertical Integration

The preceding items in this section deal with horizontal concentration, i.e. the number of competitors engaging in the same stage of production. The present item addresses vertical integration as another determinant of market structure.

Traditionally, electricity supply was a vertically integrated industry in which the same company conducted all four major activities (generation, transmission, distribution and retailing). The reasons for vertical integration in electricity are discussed in the following paragraphs.

- As a rule, during the birth and early development of an industry, companies find more efficient to centralise the various stages of production under a single umbrella. Such integration tends to disappear as markets mature, with the advent of specialised companies able to conduct one or several of the vertically integrated businesses more efficiently (assuming the existence of competitive markets).
- Vertical integration is a business strategy that attempts to turn a small number of market transactions into a competitive advantage. This advantage exists where the company's purchase and sale operations entail high transaction costs or high levels of uncertainty. Long-term agreements are an alternative to vertical integration, although in the electric power industry such agreements are complex, due to the difficulty inherent in envisaging all the details associated with possible contingencies in an industry where continuity of supply is a key factor. A relevant example is vertical integration of electricity distribution and retail, which may allow the incumbent retail activity to compete with an advantage over retailers that do not have the same access to information on consumers or that cannot use quality of supply as a commercial instrument.
- The existence of natural monopolies or high levels of concentration in one of the stages of production favours vertical integration. This is due to the predominance of a monopolist or a small number of competitors engaging in purchase and sale operations compared to the number of companies operating in other stages. Vertical integration of upstream gas and electricity production when the gas market is not competitive enough is a representative case example.
- The technological interdependence among the various production processes may also encourage players to capitalise on economies of scale.

A company's vertical integration enhances its efficiency with no perceptible effect on competition, providing the horizontal segments of the market are competitive. Where only a small number of firms compete in an industry, however, vertical merging would prevent rivals from trading with vertically integrated firms. This situation, e.g. when retailers only buy electricity from their own generators, is known as market foreclosure, and is regarded as detrimental because it reduces competition.

2.5.4 Defining, Measuring and Mitigating Market Power

Market power is closely related to high levels of horizontal concentration, although in the electricity industry analysis of this power must take the non-storability of product into consideration: a small company may bear heavily on the price, for instance, if its output is indispensable to cover demand. This item defines market power, introduces the indices most commonly used to measure such power in the electricity industry and suggests how it might be reduced. A fuller discussion can be found in [12].

Market power exists when one or several companies are able to exercise some degree of price control for their own benefit. Several items in this short definition must be highlighted:

- The ability of a company to modify the price of a market is not considered market power if there is no benefit for this company. For instance, the owner of a single power plant may withdraw it from an electricity market with the result of raising the price significantly (this may occur under conditions of a tight margin of supply over demand). Since the plant does not benefit from the price increase, unless there is some collusive behaviour with another company that participates in this market, this is not an exercise of market power. This is just a foolish behaviour by the owners of this firm.
- The standard for the "normal" or "expected price" for the purpose of measuring the impact of the exercise of market power is the competitive equilibrium price.
- Other definitions of market power specify that this ability must be maintained over a significant period of time. In the adopted definition this is already included under the term "for their own benefit", which is understood not to be negligible.
- One thing is the existence of market power (i.e., the ability defined above, a normal occurrence in most electricity markets) and another one the abuse or exercise of this market power by the agents of a market (an illegal activity under most jurisdictions). In the belief that it is unrealistic to expect agents to exercise self-restraint or perfect market monitoring by the regulatory authorities, existence of market power should in itself be regarded as a significant threat to the correct functioning of a market.
- The existence of market power mostly depends on the structure of a market and not on the rules in a competitive market. However, flawed market designs may facilitate the existence and utilisation of market power.

As indicated above, some official definitions of market power differ in some key elements from the definition that has been adopted in this book. For instance the US Federal Trade Commission and the US Department of Justice define market power as the ability of a single or several competing firms to set prices above their competitive level or consistently withhold supply to raise prices for their own benefit for a given period of time. Note that price dumping, i.e. to lower prices to draw weaker competitors out of a market or to deter entry, is also a manifestation of market power.

In its Guidelines on the assessment of horizontal mergers under the Council Regulation on the Control of Concentrations between Undertakings, the [3] defines "increased market power" as "the ability of one or more firms to profitably increase prices, reduce output, choice or quality of goods and services, diminish innovation, or otherwise influence parameters of competition". This broader definition substantially agrees with the one adopted here.

Although "market power" is theoretically the basic term for the aforementioned definition, the European legislation opts for the term "dominant position", defined by the European Court of Justice in the United Brands case (European Court, 1978), as "a position of economic strength enjoyed by an undertaking which enables it to prevent effective competition being maintained on the relevant market by giving it the power to behave to an appreciable extent independently of its competitors, customers and ultimately of its consumers".

Market power can be exercised in multiple ways. Article 82 of the Treaty of the European Community describes the different possibilities:

"Any abuse by one or more undertakings of a dominant position within the common market or in a substantial part of it shall be prohibited as incompatible with the common market in so far as it may affect trade between Member States.

Such abuse may, in particular, consist in:

(a) directly or indirectly imposing unfair purchase of selling prices or unfair trading conditions:
(b) limiting production, markets or technical development to the prejudice of consumers;
(c) applying dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage;
(d) Making the conclusion of contracts subject to acceptance by other parties of supplementary obligations which, by their nature or according to commercial use, have no connection with the subject of such contracts"

Market power fundamentally depends on the structure of the market rather than the rules, under sound market designs. This is explained in detail in Chap. 7, in the context of electric power markets.

A number of indices are in place to measure market power or the level of market imperfection, such as companies' ability to alter prices and their effect on reducing net social benefit. These indices have been widely used in anti-trust policy, market monitoring and merger valuation. Most, however, were developed for slower paced industries than electricity, where demand changes constantly and there are numerous ways to influence prices. For that reason, multiple indices have been developed, not to replace but to supplement the measurement of market power in the electricity industry.

Indices can be classified in three categories, depending on whether they measure concentration, impact on prices or the proportion of demand covered by the company analysed.

The m-firm concentration ratio, abbreviated C_m, is the simplest and most widely used parameter to measure market concentration. Defined as the aggregate market share of the m largest companies, it is normally calculated for the four most prominent companies:

$$C_m = \sum_{f=1}^{f=m} \alpha_f \tag{2.23}$$

where α_f represents company f's market share.

Other simple metrics commonly used, see for instance [3], are the number of generating companies representing at least 95 % of the market share, or the number of main electricity generating companies (companies can be considered "main" if for example their market share is larger than 5 %).

Another widely used concentration-based measure is the so-called the Hirschman-Herfindahl Index, abbreviated R_H or HHI, that measures industry concentration as the sum of the squares of each participant's market share. This index includes more information on the distribution of company size than the simple concentration ratio.

$$R_H = \sum_{f} \alpha_f^2 \tag{2.24}$$

The R_H for a monopoly, which has a market share of 100 %, is 10000. In a hypothetical market consisting of ten companies of the same size, the HHI would be equal to $10^2 \times 10 = 1000$. A figure of 2500, defining a market with four companies of the same size, is generally considered to be the upper limit for reasonably efficient market operation with no need for regulatory intervention. Criteria on this matter are not uniform and they also depend on the nature of the market. In power systems values of HHI of 1000—1800 points are typically regarded as corresponding to moderately concentrated markets, and markets with an HHI in excess of 1800 points are considered concentrated.

Concentration indices are widely used in anti-trust policy, for experience has shown that prices tend to be higher in markets with high concentration indices. Another advantage to these indices is their undemanding data requirements, making them readily usable to measure concentration in a regulated industry before it is liberalised or to determine the impact of potential mergers on competition. Figure 2.15 contains a schematic diagram for evaluating the impact of a merger based on the absolute value of HHI after the operation and the pre- and post-merger differential.

Because of their simplicity, the C_m and R_H indices have been traditionally used as market imperfection indicators and in anti-trust policy and market monitoring. Another reason for their popularity is the clear conclusions that can be drawn, in the case of C_m and R_H, about market structure in terms of supply-side concentration, and in the case of L_I, about market operation in terms of pricing.

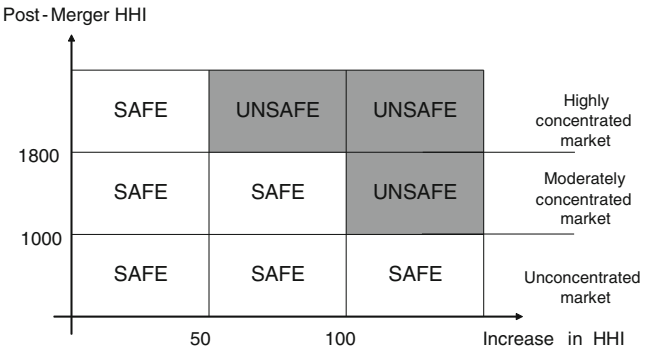


Fig. 2.15 Merger guidelines and HHI (*Source* [13])

While they provide useful information, these metrics based on market shares are in themselves not a reliable measure of market power, see e.g. [10], since they cannot reflect certain key aspects, such as resources or technology control. These difficulties are more accentuated in power markets, where crude indexes are unsuitable for studying market power. When the value of these indexes is significantly high (for example an HHI of 2200), it can be asserted that there is a clear evidence of the presence of market power, but on the contrary, in the case of electricity markets, a moderated value of HHI may fail to capture a non-competitive market structure. For example, an electricity market with 10 companies with 7 % market share and one company with 30 % but owning most of the peaking plants and some of the base load ones would result in an HHI of 1390 (apparently moderated) that fails to detect that the large company has a significant capacity to exercise market power. Although the general-purpose indices can furnish a first idea of the concentration of a market structure, they do not provide information on other factors with a sizeable impact on market power, such as the amount of base loaded production (which never sets the price) or the marginal capacity, the technology mix, the availability of manageable hydro resources or the margin of supply over demand at any given time.

Electric power is virtually non-storable, demand must be met instantaneously and the demand curve for electricity is very inelastic. Under these circumstances, a sole supplier could raise the market price for electricity substantially if its capacity were indispensable to meet demand. Pivotal methods measure the portion of demand covered by a given supplier and can therefore be used to predict high price situations on electricity markets.

The pivotal supplier indicator (PSI_f) measures whether or not a generation company f is pivotal in the market. A company is regarded as pivotal if all other producers are unable to cover market demand. PSI is therefore a binary indicator (pivotal or otherwise) that determines whether a given supplier is needed to supply demand.

The residual supply index (RSI) is a non-binary alternative to PSI. RSI_f and it is defined as the ratio between the total capacity of all a company's competitors to total demand.

$$\begin{aligned}\mathrm{RSI}_f &= \frac{\text{Company } f\text{'s residual supply}}{\text{Total demand}} \\ &= \frac{\text{Total supply capacity} - \text{Company } f\text{'s supply capacity}}{\text{Total demand}}\end{aligned} \tag{2.25}$$

RSI_f adopts values of under 100 % when a generator is pivotal. The main advantage of RSI compared to PSI is that it measures the company f's relative weight continuously, whereby minor market changes can be analysed.

Another frequently used index is the number of hours during which a generator is marginal, that is to say, when it fixes the price. Since market power may be exercised not only by raising the price but also by withdrawing capacity, and this indicator furnishes no information about infra-marginal capacity, it is also unable to properly identify market power.

If the definition of market power were to be applied directly, a reasonable measure would be the difference between prices in perfect competition and actual prices (the so-called output gaps), assuming that agents use all their potential to manipulate market outcomes and maximise their profit. The Lerner index is based on this difference, and it appears to be the perfect measure of market power, according to the adopted definition. However, its practical implementation often entails a considerable degree of simplification.

The Lerner index, L_I, is a behavioural index that measures market imperfection as overpricing with respect to a perfect market. It is therefore a totally adequate measure of the impact that market power abuse may have on market prices.

$$L_I = \frac{p_{\text{realmarket}} - p_{\text{perfectmarket}}}{p_{\text{realmarket}}} \tag{2.26}$$

Unlike the R_H, the Lerner index calls for information on market prices and underlying costs and hence is usually applied to markets once they are in operation, although it can also be estimated by ex-ante simulation. It is useful for monitoring and controlling the operation of a given market but is much more difficult to compute than concentration indices. Its usability obviously depends on the physical and economic characteristics of the power system and the accuracy of the model applied to compute what the price would be (system marginal cost[29]) given the behaviour expected of market agents under perfect competition. The primary advantage of the Lerner index is that, since it is based on market

[29] Marginal cost-based analyses are said to be difficult to justify in electric power generation for a number of reasons, including, among others:variable costs are confidential, long- and short-run as well as start-up costs are discontinuous, and the value of water is not a marginal but an opportunity cost.

operation, it may take company behaviour into consideration, including the reaction to the threat of potential new entrants.

The reduction of market power in electricity markets where prices are substantially higher than they would be on a perfectly competitive market is a recurring issue that is extremely difficult to address successfully in practice.

Market power arises as a result of horizontal concentration in a given industry. In principle, the methods for effectively mitigating or countering market power are structural, i.e., measures that focus on reducing company size and increasing the number of competitors, which can be achieved in a variety of ways. Measures that do not limit the ability to manipulate prices, or the incentive to do so, will not be effective. Here a list of the most representative measures is provided. A deeper discussion will be provided in Chap. 7, once wholesale markets have been described.

- The most effective and direct manner to curb market power consists of requiring all generators with an overly large market share (over 20 % or 25 %) to sell part of their assets, with a view to reducing market power metrics to acceptable values. Divestment measures are very difficult to implement in practice when companies are privately owned, however. A variation on this theme is to prohibit dominant operators from increasing their market share.
- The existence or imposition of long-term agreements that fix the revenues earned by dominant firms for part of their capacity likewise reduces the impact of the absence of competition. An alternative format is the virtual power plant auctions, which entail the temporary assignment of plant owners' generating capacity to third parties.
- Another measure consists of facilitating the entry of new producers by removing any regulatory difficulties or uncertainties that might serve as deterrents. This measure is very important if the long-term aim of lower concentration is to be reached and is, in any event, a measure required to ensure that a market remains competitive in the long term.
- Similarly, competition can be enhanced by furthering interconnections between neighbouring electric power systems to heighten competition between adjacent markets.
- Finally, a more elastic demand, able to react to high prices, would also reduce market power. This can be achieved, e.g. by demand response programs, information, or educational programs, and facilitated by new telecommunication technologies.

Unlike the preceding measures for mitigating market power, which aim to reduce the size of the dominant firms, other methods act directly on their behaviour. The regulator may, for instance, through the Market Operator, amend or even eliminate certain offers regarded to entail anti-competitive strategic behaviour. Regulators have also attempted to mitigate market power by introducing other rules. For example, price caps interfere with satisfactory market operation, since they bar high prices when productive capacity falls short of demand, discouraging new investment. Another example is to introduce pay-as-bid

rules in wholesale markets. However, these rules often make the market more opaque, rendering small company operation more difficult and favouring larger firms, i.e., achieving exactly the opposite of what was intended.

2.6 Oligopoly, Collusion and Strategic Competition

The preceding sections of this chapter deal with the reasons that may lead to high levels of concentration in electricity markets. The present section introduces the basic principles that explain company behaviour on markets with only a few competitors. Oligopoly is analysed, dividing this market structure into non-cooperative situations, collusion and cartel. This is followed by a discussion of strategic competition. Note that some of the anti-competitive behaviours described here may be illegal under some legislation.

2.6.1 Oligopoly

An oligopoly is a market structure in which only a small number of companies compete. Contrary to perfectly competitive markets and monopolies, in oligopolistic markets each company must necessarily bear in mind the interdependence between its decisions and the decisions of all other actors, present or future. This strategic interdependence varies with the time frame of the decisions to be made.

- The primary aim of oligopolistic companies' long-term strategic decisions is to ensure their position of privilege on the market. Consequently, they attempt to establish barriers to discourage new entrants. This, which is known as strategic entry deterrence, is addressed below in the item on strategic competition.
- These companies' medium-term objective, assuming maximum output to be constant, is to maximise their profit, bearing in mind that their competitors aim to do the same. The oligopolistic models based on game theory discussed in this section attempt to explain the rational behaviour of companies operating on real oligopolistic markets.
- Short-term decision-making involves two key factors. Firstly, since in the short term competitors lack response time, each incumbent must predict the competitors' behaviour on the grounds of their normal response, as defined by their supply curves. Secondly, short-term decisions should enable the company to attain its medium-term objectives, in the realisation that these may not concur with nearsighted shorter term aims. Some of the oligopolistic models reviewed below also explain companies' rational, short-term behaviour.
- The importance of oligopolies lies not in companies' strategic or interdependent behaviour, however, but in the fact that they are the most widespread situation in electricity markets, due to the entry barriers and cost structure of electric power generation described in the preceding sections.

2.6.2 Non-cooperative Oligopoly Models

Non-cooperative oligopoly models attempt to explain how companies compete on concentrated markets. Oligopolistic companies may capitalise on the absence of competition to hold prices higher than marginal costs and thereby earn super normal profits. Consumers are price takers whose behaviour is consequently modelled using the demand function, which also relates price to total demand.

Oligopolistic models are based on game theory, which explores players', in this case companies', rational behaviour, assuming that decisions should take competitors' reactions into consideration. Under Nash equilibrium conditions, each firm's decision maximises its profit given the decisions chosen by the other firms and this condition holds simultaneously for all firms. In other words, Nash equilibrium is reached when no firm believes it can improve its position by making a unilateral move. Price and quantity are the decision variables used in these games. Investment in capacity or the threat of new entrants is discussed below in the item on strategic competition.

To find the short-term market equilibrium, the problem is expressed in terms of each firm's optimisation, in which the objective function consists of maximising its profit:

$$\prod_f = p \cdot q_f - \mathrm{TC}_f \quad \forall f \tag{2.27}$$

where p is the market price and TC_f is firm f's total cost to produce q_f.

Competition among firms is based on their decision variable, in this case the total output, q_f, each brings to market. With output thus established, price becomes a dependent variable that can be obtained from the demand curve:

$$p = f(\sum_f q_f) \tag{2.28}$$

Market equilibrium defines the set of q_f values that meets the conditions of each firm's first order optimality condition:

$$\frac{\partial \prod_f}{\partial q_f} = p + q_f \cdot p'_f - \mathrm{MC}_f(q_f) = 0 \tag{2.29}$$

The above equation is obtained assuming that the derivative of price with respect to the firm's own output, p'_f, depends not only on the slope of the demand curve, as in a monopoly, but rather on the slope of the residual demand curve.

$$\frac{\partial p'_f}{\partial q_f} = p'_f, \quad \forall f \tag{2.30}$$

Residual demand is the market demand open to a company after eliminating the demand supplied by its competitors. Therefore, a company's residual demand is a function that relates price to output, obtained as the total demand function minus the supply curve for all its competitors.

As in a monopoly, each optimality Eq. (2.29) shows that at equilibrium, where each company earns its maximum profit, its marginal revenue (MR_f) is equal to its marginal cost MC_f.

$$\mathrm{MR}_f(q_f) = p + q_f \cdot p_f' = \mathrm{MC}_f(q_f), \quad \forall f \tag{2.31}$$

First-order conditions (2.29) show that the equilibrium price is higher than the marginal cost, since the residual demand slope, p_f', is negative. This overcharge with respect to the perfect competition price lowers net social benefit.

$$p = \mathrm{MC}_f(q_f) - q_f \cdot p_f' > \mathrm{MC}_f(q_f), \quad \forall f \tag{2.32}$$

As noted in the item on market power, one of the corrective measures consists of requiring companies to enter into long-term agreements. Take, for instance, a certain portion of output, Q_f, sold under a long-term agreement, which yields revenues that do not depend on the market price: the resulting equilibrium price can be readily shown to be smaller than the previous market price, further to the following expression:

$$p = \mathrm{MC}_f(q_f) - (q_f - Q_f) \cdot p_f', \quad \forall f \tag{2.33}$$

Coming back to Eq. (2.29), note that the difference between that equation and the equations that describe optimal behaviour under perfect competition or by a monopoly is the value of p_f'. In fact, all that distinguishes one market model from another is companies' conjectures about the impact of their decisions on price, which translate into different values for p_f'.

The various market models whose equilibrium conditions adopt the form of Eq. (2.29), but whose p_f' conjectures differ, are explained below.

- In the Cournot model, each company determines its production on the assumption that changes in its output will have no effect on its competitors' output. This is the same as assuming that a change in one company's output will only translate into a change in price through the demand curve. Consequently, p_f' is equal to the slope of the demand curve, p'.
- Under perfect competition, companies compete assuming that their decisions have no effect on price because their output accounts for only a small portion of the total production. This means that the value of p_f' is zero.
- The Stackelberg model proposes an asymmetric game in which two types of companies make decisions in two stages. Firstly, leader firms decide the output

that will earn the maximum profit. The follower firms then decide their output on the grounds of the leaders' decisions. In this case, the follower firms' conjecture is similar to the Cournot model conjecture (they regard their competitors' output as constant) and the value of p'_{follower} is equal to the slope of the demand curve, p'. The leader firms conjecture differently, however, for they bear in mind the follower companies' future reaction. The value of p'_{leader} is not equal to p', but it is obtained as the residual demand surrendered by the followers.

- The dominant or leader firm price model also envisages two types of companies that adopt decisions in two stages. In this case the leader or dominant firm decides the price and the follower companies decide their output on the basis of that price. Consequently, the follower firms' conjecture is the same as competitive companies', for they regard their effect on price to be nil (p'_f equals zero). The dominant firm anticipates followers' competitive behaviour by obtaining its own residual demand as total demand minus the followers' supply curves (which concurs with their marginal cost curves).

Other oligopolistic models can also be cited, such as the Bertrand model in which the decision variable is the bid price for the entire output. Each company is regarded to be able to meet 100 % of the demand, resulting in market price equilibrium that is very similar to the price on a competitive market, where companies bid at marginal cost. However, it can be proved that the result is Cournot equilibrium if companies first compete on the grounds of output as in the Cournot model and then on the grounds of price as in the Bertrand approach.

All these models lead to market equilibrium prices that lie in between the prices resulting from competitive and monopoly structures. While there is no theoretical model able to fully depict competition among a handful of companies, the models listed above describe behaviour patterns observed in companies on real electricity markets. These patterns vary with relative company size and the time frame considered.

The following is a list of behaviours observed on real markets that conform to the theoretical patterns described in the foregoing (as noted, these behaviours may be illegal).

- The most rational behaviour for small companies operating on oligopolistic markets is to compete by acting like follower firms in the dominant firm model.
- Large companies, on the contrary, should rationally assume the leader's role in that same model.
- The dominant firm model description of leader firms' behaviour, i.e. profit maximisation on their residual demand, also appears to suitably depict such companies' short-term conduct. Residual demand is obtained as total demand less the competitors' estimated supply curve, based on their usual behaviour.
- The sequential nature of the Stackelberg model makes it particularly apt for studying the new investments in of generation capacity. The company that reacts

first, the leader, makes its decision knowing that other companies will subsequently expand to optimise their position, but only after the leader has established its capacity. Companies that delay their expansion decisions are followers, who make their optimal expansion decisions in the awareness of what the leader's capacity will be.

- The Bertrand model is applicable to competitive tendering in which the company with the best bid is awarded the contract.

2.6.3 Cooperative Models: Collusion and Cartel

The oligopolistic models discussed above are based on a market equilibrium in which, under certain conjectures, all competitors maximise their profit simultaneously. The profit obtained in such models is lower, however, than would be earned if these companies coordinated their strategy to maximise joint profits and distribute them in accordance with some rule. In an infinitely repeated Cournot game, for instance, smart players learn without explicitly informing one another that they could earn more if they jointly lower output to increase all actors' profits. Collusion is defined as an explicit or tacit agreement to obtain an objective by gaining an unfair advantage. If the agreement is explicit the result is a cartel. Regulation should prevent both tacit and explicit collusion, because both lower the net social benefit.

Collusion is more readily sustainable over time in highly concentrated markets, where coordination among companies and agreement monitoring is easier because only a few competitors are involved. If collusion is explicit, rules for distributing market share must be established, whereas in tacit collusion coordination may be based on the use of conventional mark-up pricing rules. Another coordination alternative, similar to the dominant firm model described above, consists of the implementation by all incumbents of the price strategy defined by the leader firm, which is usually the company with the largest market share. Agreement monitoring calls for sufficient price and market share information, which is normally available in electricity markets. The main threat to the stability of collusion is failure by a given company to comply with the agreement in an attempt to raise its profit by increasing its market share at the expense of its higher priced competitors. Credible penalisation mechanisms must therefore be in place to discourage companies from failing to honour the agreement. A price war after the detection of a new free rider is the most widespread penalisation mechanism on highly concentrated markets. This measure has occasionally been observed on electricity markets.

2.6.4 Strategic Competition

An attempt to improve a company's future market position is known as strategic competition. For a strategic decision to be effective, it must be credible and able to reduce present or future competitors' expected profit.[30] The two natural types of strategic competition in electricity markets are predatory pricing and strategic entry deterrence.

Predatory pricing is the most extreme form of strategic competition. The dominant firm lowers its prices to force the least efficient competitors to exit the market. This strategy must obviously be sustainable over time to be effective.

Strategic entry deterrence is the branch of strategic competition that engages in creating entry barriers. The two most common examples are limit pricing and investment in cost-reducing capital, in which incumbent firms forfeit short-term profit in pursuit of high levels of longer term concentration.

- *Limit pricing* consists of increasing output over its optimal level to maintain prices lower than would be obtained with a profit maximisation strategy. One example of limit pricing is where the dominant firm seeks a price level in which it obtains super normal profits without attracting new competitors. That price would be just slightly below the new entrants' long-term average cost. In this case, additional entry barriers must be in place to enable the dominant firm to obtain long-term cost advantages over new entrants.
- *Investment in cost-reducing capital* involves investing in capacity over and above what the maximum profit level would be in the absence of new competitors. Since capacity investments are an irreversible decision, potential new entrants evaluate the real risk of intense post-entry competition that would lower their expected profits. One example of investment in cost-reducing capital is where a company behaves like the leader in the Stackelberg model and decides to invest in capacity knowing that its competitors, the followers, will have to accommodate its decision. The incumbent firm knows that the larger its capacity at the time of new competitor entry, the larger its market share. Conversely, seeing that their market share will be smaller, new entrants may be discouraged from entering the market.

2.7 Market Failures and Externalities

As noted earlier, when a market is not perfectly competitive, its output is inefficient. One of the underlying reasons, imperfect competition, has already been addressed. Other market failures include: incomplete information, lack of complete markets, transaction costs, and externalities. This section deals with last of

[30] The decisions accorded the highest credibility are irreversible decisions.

these factors. A more complete discussion of the first may be found in [11], for instance.

Externalities arise when an economic agent (a consumer or a producer), is affected by another agent's production or consumption decisions, which are not taken into consideration by the latter in its production or utility function. The failure to take this impact into account leads to inefficient resource allocation, as explained in greater detail below. The most common, although not the only, externalities in the power industry have to do with environmental impact.

Externalities may be negative or positive. The classic (and truly old fashioned) example of a negative externality is the damage caused by a coal power plant emitting soot that soils the laundry of the households nearby. Because of the soot, the laundry must be washed over and over again, but the power plant owner does not take this detriment to third parties into consideration when deciding how much electricity to produce or which fuel to use. An example of a positive externality is the flood regulation afforded by a hydroelectric dam.

Why do externalities produce inefficient resource allocation? This can best be illustrated by an example. Take a power plant located in the middle of a city that pollutes the air, with adverse effects on city dwellers' health. Assume for the time being that these impacts are limited to the city.

As shown in previous sections, the power plant produces electricity, and with it air pollution, until the marginal benefit of doing so equals zero. But this benefit does not take the harm to public health into consideration. Assume also that the marginal harm increases with the level of pollution as illustrated in Fig. 2.16, where electricity production is represented in the horizontal axis and the economic value of marginal damages and benefits of this electricity production in the vertical axis.

In the absence of externalities, the efficient production level for the power plant would be E. The figure shows, however, that this production also causes significant harm. If the power plant reduces its production somewhat, its profit also declines, but at the same time it lessens the harm inflicted on the population. At level X, the loss of marginal profit resulting from lowering production is much smaller than the marginal damage avoided by doing so. In fact, the city might pay the power plant to reduce pollution by an amount higher than its *lucrum cessans*, but lower than the harm prevented, and both would be better off. In other words, a reduction in power plant production would be efficient. As can readily be seen in the figure, efficiency rises until the loss of marginal benefit equals the marginal harm prevented (level EF, for efficient), which is the point at which marginal joint benefit is zero.[31]

Enlarging on this last point, imagine that the city owns the power plant. Under these circumstances, when determining the optimal power output, the city (the owner) will take possible harm to public health into consideration, and factor it

[31] Note that the efficient level of pollution is not necessarily zero. That is, the externality may disappear, but that does not imply that pollution also disappears.

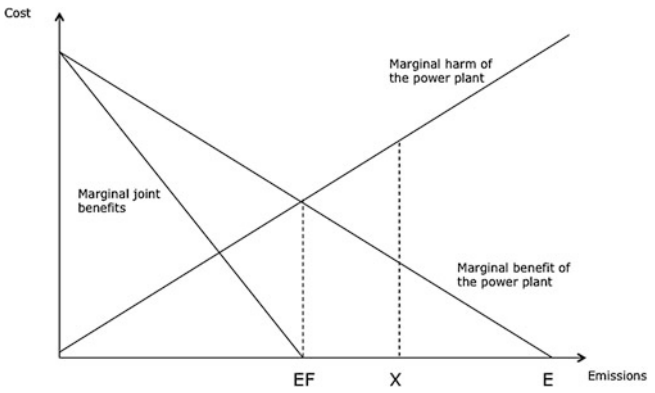


Fig. 2.16 Marginal benefits from electricity production and marginal harm to public health

into the production function. How much will the plant produce? The amount is shown as EF, because this is the level that maximises the city's total profit. The externality ceases to exist because all costs are reflected in the production function.

Why does the power plant fail to take this type of harm into account when it is not owned by the city? That is, why do externalities arise? The usual reason is that inasmuch as property rights are not suitably allocated, the plant owners have no incentive to take the harm to public health into consideration. If either of the two agents in the example were allocated the right to pollute or to enjoy clean air, the externality would disappear and the market outcome would be efficient. If the city is awarded property rights to enjoy clean air, and assuming the power plant wants to go on producing (and polluting), it will have to compensate the city for the harm caused, and include that consideration in its production function. Here also, the final outcome would be efficient.

Interestingly, this efficient outcome does not depend on how the property rights are defined. If the power plant has the right to pollute, the city will be willing to pay it to reduce its production, thereby reducing the public health menace. If the city has the right to enjoy clean air, the power plant will be willing to pay the city to allow it to produce, as long as the marginal benefit earned for from production is larger than the amount paid (which should at least cover the marginal harm). In both cases, electric power output and the adverse effects on health will reach efficient levels, although the amount of pollution and, depending on the allocation of property rights, the distribution of wealth, may vary. This result, known as the Coase theorem [2] is valid as long as no transaction costs are involved.[32]

[32] See Annex A of this book, "Grandma's inheritance theorem", for a folk version of Coase's theorem.

In addition, if the city's preference for clean air[33] is quasi-linear (dependent only on the amount of pollution, irrespective of income), the final amount of pollution is also the same, regardless of the allocation of property rights. This is a fairly bold assumption: typically, preferences for clean air rise with income.

Externalities, then, would appear to merit little concern: the market itself provides incentives to abolish them, providing property rights are allocated and no transaction costs are involved. In such cases, the regulator should not intervene to compensate anyone, since that would remove the incentive to attain an efficient solution through negotiation. If the property rights are allocated to the power plant but the government decides to compensate the city for the marginal harm, the power plant would continue to produce all the electricity it wants, and the city would have no incentive to negotiate. In fact, the city might even prefer more pollution, for it might mean higher compensation! The final outcome is not efficient, however. On the contrary, everyone (including the government) would benefit if the power plant lowers its production to reduce the adverse effects for health and the compensation to be paid.

But, what if property rights cannot be allocated? In such cases externalities cannot be readily eliminated. This is generally the case where public and common goods are involved. These are special types of goods, characterised by two elements: excludability and rivalry.

Excludability, the possibility of excluding an agent from participating in the good, depends on physical, legal, or economic factors. No one can be excluded from breathing their city's polluted air, for instance, unless very expensive solutions are adopted (an artificial bubble, for example). Pollution is therefore a non-excludable "bad". TV channels were also non-excludable until technology made it affordable to use decoders.

The inability to exclude someone from enjoying a good is the primary reason that externalities exist and poses problems for the efficient allocation of goods. Essentially, no charge can be established for the use of the good, because it can be used anyway: no property rights can be allocated. Consequently, the price system, which is at the heart of an efficient market, is useless in such cases.

Rivalry refers to the decline in the amount of a good available after being used. Typical consumer goods such as butter or cars are rival products. But public TV is not: when one person watches a certain programme, others' possibility to do likewise is unaffected. The same is true of pollution, or to put it positively, with the reduction of pollution: when pollution declines and air is cleaner, the clean air breathed by any one city dweller has no effect on the air breathed by others. Therefore, lowering pollution is a non-rival good (or pollution is a non-rival "bad"). The problem posed by non-rivalry is that, if no opportunity cost is

[33] Here the adverse effect on health has been used as a proxy for preferences. When harm or benefits are used to represent preferences, the quasi-linear assumption usually holds, since they are expressed in monetary terms which generally take income into consideration. This is also applicable when the externality is borne by a firm and the utility function is therefore a production function.

attached to the use of a good, the marginal cost of providing it is zero. Therefore, even if users could be charged, the efficient price would be zero, because the good is to be supplied to as many consumers as possible. If the price is zero, the market undersupplies the good, for producers (who nonetheless incur production costs) will not be able to recover their costs.

Combining excludability and rivalry yields different types of goods. Public goods are non-excludable and non-rival. Some examples are pollution reduction, public TV, or national defence. Common goods are non-excludable but rival. Examples include common pastures or land, fish in international waters or health services. The focus here is on public goods, which are the most relevant for the power sector.

Public goods pose two problems in connection with market trading. Firstly, they cannot be allocated property rights, because they are non-excludable. This spawns an externality that has no market solution. Secondly, since their marginal cost is zero, they will typically be undersupplied by the market.

To return to the aforementioned example, now assume that the power plant has adverse effects on the health not only of city dwellers, but also of the population in the entire region. Under such circumstances, no one in the region can be excluded from the polluted air, making the allocation of property rights much more difficult or even impossible. Negotiation is also more difficult, since everyone in the region must agree to it, and transaction costs, if any, are higher. Finally, even if they decide to negotiate, people may have different preferences respecting clean air, and different incomes, but different amounts of pollution cannot be allocated depending on these factors, due to non-excludability. Furthermore, people are given incentives to free-ride when deciding how much to pay for reducing pollution: inasmuch as no one can be excluded, and the marginal cost is zero, each individual benefits by letting others pay for the clean air he breathes for free.

When externalities are generated by the existence of public goods, resource allocation is inefficient and market agents have no incentive to reach an efficient solution. Regulation may therefore be needed to establish the efficient outcome and the incentives required. The regulation of environmental impact-related externalities is addressed in Chap. 11.

Annex A
Standard Microeconomic Theory

This annex discusses the economic behaviour of markets in very general terms, which need to be adapted to the specific circumstances prevailing in electricity production, as shown in Sects. 2.3 and 2.4 of this chapter.

How does a perfectly competitive market evolve from short-term to long-term equilibrium? Following the definitions of costs components of electricity production in Sect. 2.1.2, the marginal cost (MC) (additional cost of producing one

more unit) and average cost (AC) (total cost divided by total output for each level of production) curves are used here to understand these dynamics, see Fig. A.1.

The "U" shape of the average cost curve is directly related to the existence of economies of scale. For many productive activities, average cost is high for low output, lower in some range of output, and high for high levels of output. This yields the familiar "U-shaped" average cost curve. This shape is a function of quantity: for small quantities, fixed costs are high; for large quantities, variable costs increase as quantity approaches production capacity.

- When the average cost declines, marginal cost is below average cost. This is because if a unit of production costs less than the average of all previous production, the average cost falls.
- If average cost is constant (its derivative is zero), the cost of producing one more unit is equal to the average cost of producing the previous units of production. Marginal cost is consequently equal to average cost and the two cost curves intersect at the so-called break-even point as shown in Fig. A.2.
- Finally, if a unit of production costs more than the average, the average cost rises. In that case marginal cost is greater than average cost.

In the electricity sector the typical case is to be situated either in the first or in the second part of the curve: that is, to have either decreasing or flat average costs. In practical terms, electricity generation shows an average cost curve in which there is a large flat area (when economies of scale have been exhausted, but efficiency losses have not been reached yet). This is due to the fact that, when demand is high enough, it is better to serve it with multiple power plants than to keep building bigger ones.

As explained earlier, the company's optimal output, q^*, is the point where its marginal cost (MC) curve intersects with market price, p. The company depicted in Fig. A.2 reaches optimal production at an average cost (AC) lower than the price. This positive margin between the selling price and average production costs earns the company a super-normal profit that can be calculated as the difference

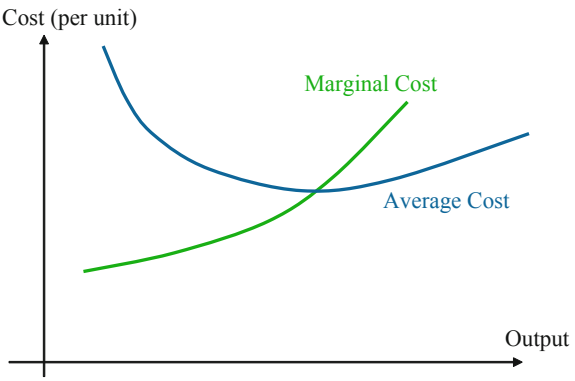


Fig. A.1 Average and marginal costs

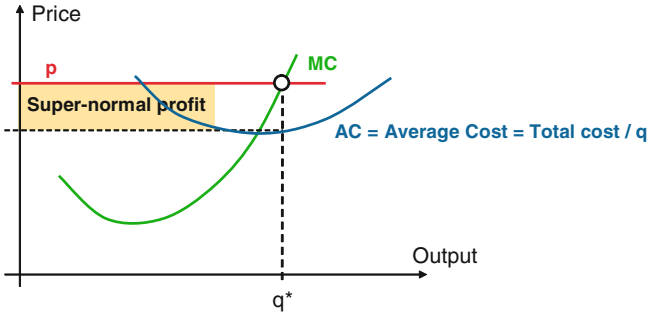


Fig. A.2 Market equilibrium under perfect competition with super-normal profits

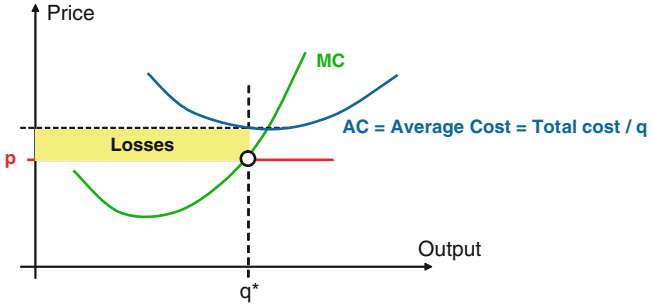


Fig. A.3 Equilibrium on a perfect market with losses

between the price and the average cost multiplied by optimal production. Graphically, this profit is the area between the price and the average cost.

Another company operating on the same market but with a less efficient cost structure might reach optimum output at an average cost higher than the price, as shown in Fig. A.3. This would lead to losses for the company that may be calculated as the difference between the average cost and the price multiplied by optimal production. Graphically these losses are represented as the area between average cost and price.

In the long run, companies competing on a market where entry and exit are free and all players have access to the same technological opportunities eventually exhibit similar cost structures.

- When companies earn super-normal profits (Fig. A.2), new entrants are attracted to the industry by its high profitability based on short-term high prices (ST price). These new entrants bring about a decline in the long-term price (LT price), since according to the demand curve, prices fall as supply rises. This is illustrated in the upper half of Fig. A.4.

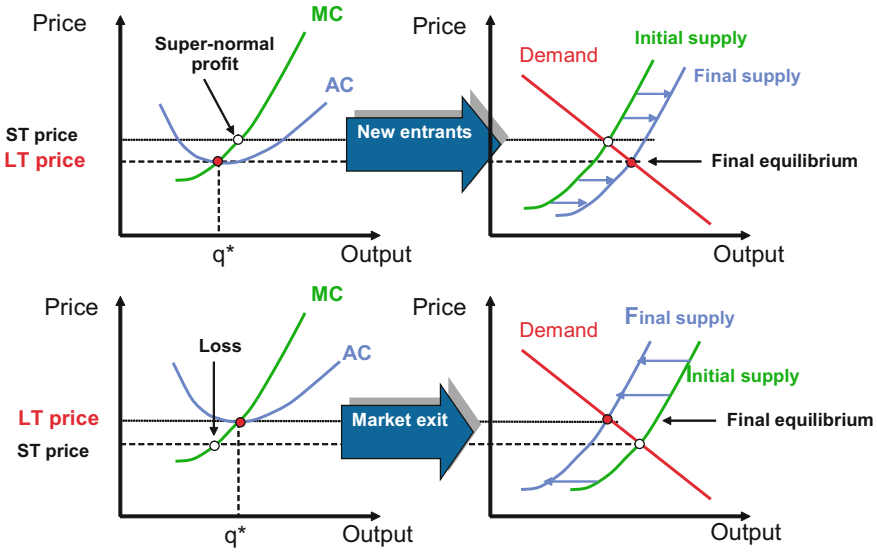


Fig. A.4 Perfect market equilibrium dynamics at the break-even point

- Companies posting losses (Fig. A.3) exit the market under the pressure of low short-term prices (ST price).[34] Their disappearance (or an increase in demand) brings about a rise in the long-term price (LT price) because, again according to the demand curve, a decline in supply raises prices. This is illustrated in the lower half of Fig. A.4.

Consequently, in the long term, as shown in the figure below, all companies adapt their cost structure to reach what is known as the break-even point, in which price equals average cost. Note that the average cost must include a return on capital, so this point does not imply nil earnings, but rather a long-term sustainable profit in keeping with the investment risk involved.

Another factor related to economies of scale is economies of scope, a situation in which average costs decline when products straddle markets. The energy sector, for instance, may benefit from synergies between combined cycle plant electricity generation and gas distribution facilities.

Monopolies

The regulation of monopolies is addressed in detail in Chap. 4. Here we present a general assessment of the economic efficiency of unregulated monopolies to justify the need for their regulation.

[34] In the presence of sunk investment costs (such as in most power plants), the firm should reasonably keep the plant operative for as long as its expected revenues exceed its expected operating costs, even if it is not operated most of the time.

We look first at an intuitive explanation of the behaviour of a monopolist in the short term, considering only decisions about the output. Then we will extend the explanation to the long term, including decisions on new investments.

A monopoly is said to exist where a single company has exclusive control of a product in a given market. As noted earlier, monopolistic companies may take advantage of this situation to raise their price if their activity is unregulated or uncontrolled, lowering net social benefit. This section describes the behaviour of a monopolist operating on a market in the absence of regulation. The time frame considered here is the short term, in which neither investment decisions nor the possible entry of new competitors is considered. These factors are addressed in Sect. 2.6 on strategic competition.

If a monopoly firm is free to set the price it will maximise its earning by applying the following equation:

$$\prod_f = p \cdot q_f - \mathrm{TC}_f \tag{A.1}$$

where price, p, is determined from the demand function[35]:

$$p = f(q) \tag{A.2}$$

The output matching maximum profit is obtained by setting to zero the derivative of that expression with respect to output:

$$\frac{\partial \prod}{\partial q} = p + q \cdot p' - MC_f(q) = 0 \tag{A.3}$$

where p' is the derivative of the price with respect to monopolistic production, which is generally negative: in other words, the higher the output the lower the price.

The maximum profit is the point at which the company's marginal revenue (the extra revenue obtained from producing one extra unit) is equal to marginal cost. Note that because p' is negative, marginal cost, MC, is a function of monopolistic output and is smaller than price for each production level.

$$\mathrm{MR}_f(q) = p + q \cdot p' = \mathrm{MC}_f(q) \tag{A.4}$$

As Figs. A.5, A.6 show, a monopolist reduces output (from perfect competition level) to increase the price paid for the rest of its production, maximising profit, as in the upper half of the figure.

A monopolist's behaviour can be reasoned intuitively as follows. Taking nil production as the starting point, the monopolist increases its earnings as long as the output of one extra unit raises revenues (marginal revenue) more than the

[35] Since microeconomics defines the demand function in terms of price, strictly speaking Eq. (A.2) is the inverse demand function.

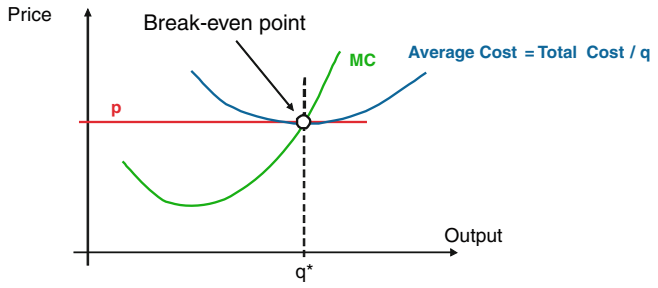


Fig. A.5 Break-event point under perfect competition conditions

associated costs (marginal cost). Consequently, maximum profit is earned where marginal cost equals marginal revenue.

Figure A.6 also shows that an unregulated monopoly would set a higher price and produce less than on a perfect market because it grows its profit by raising the price through reductions in output. This increase in the producer's surplus comes at the expense of a greater reduction in consumer's surplus. In other words, a monopolistic situation produces lower net social welfare than a perfectly competitive market. The deadweight loss caused by the monopoly in net social benefit is labelled DWL in Fig. A.6.

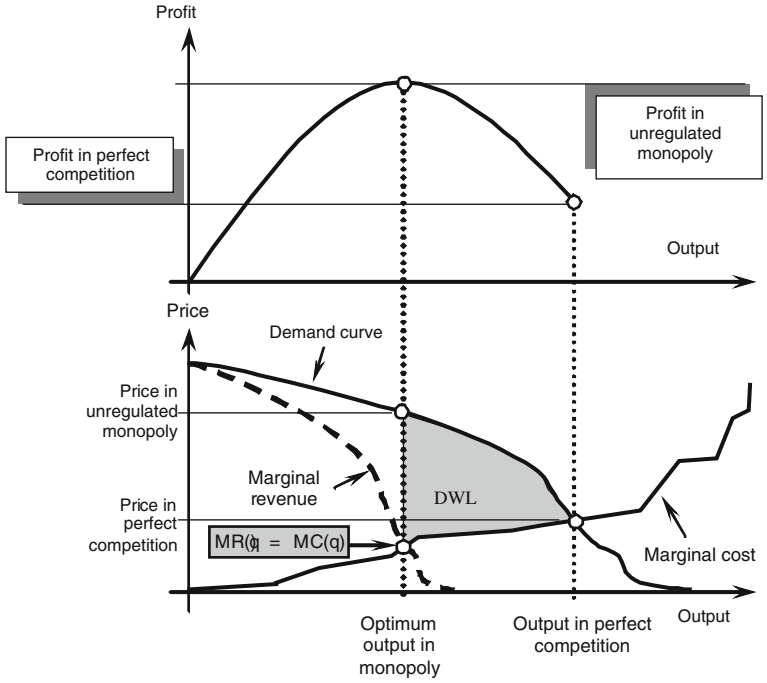


Fig. A.6 Price setting by an unregulated monopolistic firm

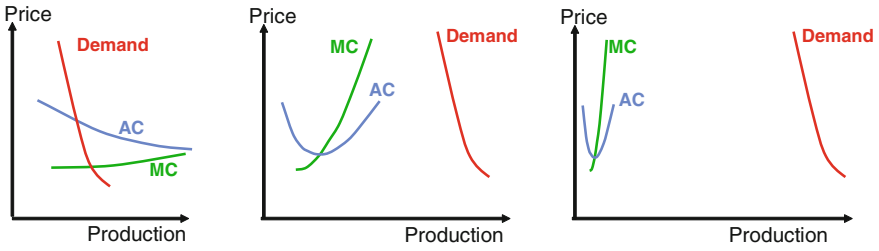


Fig. A.7 Cost patterns, economies of scale and market structure

In the long term, the monopolist must decide on new investments. Following the same arguments, it is easy to see that the maximum profit is achieved with a level of investment lower than the competitive optimum. This underinvestment results in the same negative outcome of a loss of net social benefit.

Therefore, from the standpoint of economic efficiency, unregulated monopolies should not exist. The corollary to this is that if economies of scale or the structure of an industry impede the creation of a potential competitive market, the operation of the monopolised industry must be regulated to prevent an undue decline in social welfare. Regulators must set prices or total revenues that enable the monopolistic firm to cover average total costs, including a rate of return on the capital invested in its assets to avoid low levels of investment. In fact, a well-regulated monopoly could theoretically attain the same outcome as a perfectly competitive market, as shown in Sect. 2.3.3 of this chapter.

Determinants of market structure

Economies of scale are a determinant for the number of competitors on a given market. Figure A.6 shows three possible situations with respect to average and marginal cost curves and market size. An industry in which there are strong economies of scale, and average costs are consequently consistently higher than marginal costs, will be a natural monopoly (left-hand panel of Fig. A.7). If, once the economies of scale are exhausted, total demand only accommodates the existence of a few companies, the result will be an oligopolistic market with a small number of producers (central panel of Fig. A.7). Finally, if demand accommodates many companies after exhaustion of the economies of scale, the resulting market tends toward perfect competition (right-hand panel of Fig. A.7. Please note that this classification is independent of the shape of marginal cost and average cost curves.

The existence of entry barriers limits the dynamics described in Fig. A.7, favouring the existence of highly concentrated markets.

References

1. Caramanis MC., Bohn RE, Schweppe FC (1982) Optimal spot pricing: practice and theory. IEEE Trans Power Apparatus Syst PAS-101(9): 3234–3245
2. Coase RH (1960) The problem of social cost. J Law Econ 3:1–44
3. Eurostat (2012) Electricity market indicators. epp.eurostat.ec.europa.eu
4. Pérez-Arriaga IJ (1994) Principios económicos marginalistas en los sistemas de energía eléctrica (in Spanish). Technical Report IIT-93-044
5. Pérez-Arriaga IJ, Meseguer C (1997) Wholesale marginal prices in competitive generation markets. IEEE Trans Power Syst 12(2):710–717
6. Phillips D, Jenkin FP, Pritchard JAT, Rybicki K (1969) A mathematical model for determining generating plant mix. Third PSCC Conference, Rome, 1969
7. Rodilla P, Batlle C (2012) Security of electricity supply at the generation level: problem analysis. Energy Policy 40:177–185
8. Schweppe FC, Caramanis MC, Tabors RD, Bohn RE (1988) Spot pricing of electricity. Kluwer Academic Publishers, Boston ISBN 0-89838-260-2
9. USDOE (2006) Benefits of demand response in electricity markets and recommendations for achieving them. United States Department of Energy, U S A
10. UK Office of Fair Trading (OFT) (2004) Assessment of market power, draft competition law guideline for consultation (http://www.competitionlaw.cn/upload/05062800559712.pdf)
11. Varian HR (1992) Microeconomic analysis. W.W. Norton, New York
12. Vázquez B, Pérez-Arriaga R (2006) Market power mitigation proposals for the Spanish wholesale electricity market. Working Paper IIT-06-026A
13. Viscusi WK, Vernon JM, Harrington JE (2005) Economics of regulation and antitrust. MIT Press, Cambridge
14. WREZ initiative. (http://www.westgov.org/rtep/219)

Chapter 3
Electricity Regulation: Principles and Institutions

Carlos Batlle and Carlos Ocaña

Error of omission begets new rules.
Toba Beta (Master of Stupidity)

Building a high-rise is subject to many principles, but primarily the laws of physics, which require the foundations to be apt for the materials used and the height of the building, which may otherwise collapse. While any number of buildings can be designed to that requirement, the use to which the structure is to be put reduces the number of alternatives considerably. Similarly, the electricity industry is subject to the laws and principles that govern both the physical characteristics of electricity and also to the fulfilment of the expectations of utilities and consumers. These principles, together with other practical considerations, ultimately limit the number of ways that electricity can be regulated.

This chapter introduces the basic principles underlying the regulation of the electricity industry and discusses the most prominent opinions regarding regulatory design. More than 20 years ago when most modern electricity markets were first proposed, many alternatives appeared to be open and few underlying principles seemed necessary. Today the contrary is true. With the experience accumulated in the interim, the debate on regulatory models for the electricity industry has gradually focused on a small number of options that have been successfully implemented in one or several countries, while the issues of concern and the rules to address them have grown considerably.

C. Batlle (✉)
Institute for Research in Technology, Comillas Pontifical University,
Sta. Cruz de Marcenado 26, 28015 Madrid, Spain
e-mail: Carlos.Batlle@iit.upcomillas.es

C. Batlle
MIT Energy Initiative, MIT, Cambridge, MA, USA

C. Batlle
Florence School of Regulation, European University Institute, Florence, Italy

C. Ocaña
FUNCAS, Caballero de Gracia, 28, 28013 Madrid, Spain
e-mail: dirgen@funcas.es

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_3,

The content of this chapter may be more readily assimilated if the question is viewed from the perspective of a legislator, intent upon reforming the electricity industry. When a country decides to create electricity market (or reform the existing market), a debate among the stakeholders generally ensues over the legislation to be adopted to regulate the industry. This chapter reviews the main questions that would have to be posed when drafting such legislation and the possible answers. In the last two decades, several hundreds of experts in mostly European and American and some Asian countries, as well as in Australia and New Zealand and some African nations have been faced with these same problems.

The design of industry reform calls for institutions to enforce the legislation drafted, and these bodies have to be provided with the necessary resources and authority to do their job. While several approaches and institutional designs are possible, certain basic principles need to be honoured, as indicated before. Consequently, the present chapter analyses regulatory institutions along with the basic principles of regulation.

It is explained in this chapter why an "electricity act" is insufficient by itself for regulation to be operational. Rules and regulations on multiple specific subjects—what is termed "secondary regulation"—are also needed: network tariffs, quality of supply, operating and market trading procedures or the formation of spot electricity prices, as well as many others that are usually addressed in very general terms in "electricity acts". Such rules and regulations are discussed in the second and third parts of this book.

3.1 Introduction to Regulation

3.1.1 What Is It, and What Is It for?

The best way to introduce regulatory mechanisms and their design is to briefly review the various definitions of "regulation" to be found in the literature. According to Wikipedia, which provides what is perhaps the most concise definition, regulation is a way to control individual and collective human behaviour by means of rules and restrictions. In other words, it is a series of principles or rules (with or without the authority of law) used to control, direct or manage an activity, organisation or system. From the legal perspective, regulation can be defined as the rules based on and written to implement a specific piece of legislation. It is a form of secondary legislation issued by a government minister or a regulatory body under the authority of primary legislation, whose effective application it is meant to secure.

The primary aim, therefore, is to alter the outcome that would be reached if the human species were allowed to interact freely, i.e. to prevent (or attain) a series of inefficient (or efficient) results in different places and time frames which otherwise may (or may not) happen.

Governments regulate industries to improve their unregulated performance. In the context of government and public services, regulation means establishing rules to control, but not to prohibit, an activity. Regulation can therefore be viewed as a compromise between prohibition (or more precisely command and control) and no control at all. Consequently, it is an expression of the exercise of authority through which the regulator (such as a government or the management of any given organisation) attempts to impose guidelines under the conviction that they lead to a better result. Here is where two of the main conditioning factors that summarise the regulation issue arise. First, what criteria should be applied to determine when one result is better than another? And assuming that a consensus is reached in that regard, what is the most efficient way to reach the objective sought?

Regulation should aim to steer an industry's performance towards improving "general welfare", i.e. the collective benefit gained by consumers and operators. An industry's performance can be measured in terms of consumer surplus, service availability, profitability and affordability, range of services offered, quality and degree of innovation.

Regulation seeks to protect consumers from the market power that may enable monopolies and oligopolies to set unjustifiably high prices or lower the quality of their goods or services. Regulations limit the prices that companies can establish and set quality and service continuity standards as well as rules about service coverage. Regulators also participate in investment planning in a number of ways: the State may plan investment directly, for instance, or subject companies' plans to administrative authorisation.

Regulation also seeks to protect investors from the State, which might act opportunistically by setting supply tariffs and obligations that would preclude recovery of the investment. In regulated industries, where investors must make huge investments in specific assets that cannot be used anywhere besides where they are installed, investors could do very little once the investment is made to protect themselves against such opportunistic behaviour. The State, in turn, may be keen on acting opportunistically to benefit consumers or reduce inflation, for instance. Regulation prevents such behaviour by requiring prices to reflect costs and, much less often than might be desired, transferring the power to set prices to independent regulatory bodies.

A distinction is generally drawn between "economic" and "social" regulation. The former is industry specific and focuses on prices, quality and safety, market entry and exit and investment. The latter, in turn, addresses social concerns such as health, security or the environment. Economic regulation deals primarily with correcting monopolistic markets or imperfect competition, whereas social regulation attempts to correct externalities, such as safety information, quality or environment-related problems. The present chapter deals essentially with economic regulation, but issues of social regulation are addressed elsewhere in this book.

Regulation is not the only way to protect consumers and investors. Their interests are also defended by courts and legislation, including laws on competition. The difference is that, while regulatory action is ex-ante, judicial action is ex-post. Moreover, courts generally act fairly slowly and are subject to very strict

procedural rules that may make it difficult to prove and penalise certain types of behaviour. Because of this difference, judicial action is not a perfect replacement for regulation, and monopolies are regulated in most of the world.

3.1.2 Scope of Regulation

Regulation consists of three basic elements:

- The first is the design of rules for steering agents' behaviour towards the objectives defined by the regulator. To use an example that has nothing to do with electricity, in soccer leagues, wins are assigned three points and ties only one, to encourage teams to strive to win, thereby enhancing competitiveness. Similarly, the number of player substitutions is limited to three per match to narrow the gap between teams with larger and smaller budgets. In some electric power markets, agents are required to buy and sell their energy on a spot market, first to enhance transparency and ensure that all agents know the price at which electricity is traded, and second to enable smaller actors to compete more effectively.
- The second is the structure of the power industry. When market mechanisms are adopted, an appropriate business structure is generally needed. In the electric power industry, for the market to operate properly, a sufficient number of similarly sized competitors must participate. Unfortunately, this precondition is frequently unmet in practice. To return to the soccer simile, a poorly balanced distribution of television revenues among different teams compartmentalises competition and reduces the number of teams that can aspire to be champions. Or, for instance, it does not make sense to establish a championship mixing teams from the NFL and college football. In the long run, this lowers audience interest in the championship. In the electricity industry, the implementation of market mechanisms has on occasion been preceded by the sale of separate parts of a former public utility with a view to ensuring certain minimum levels of competition, i.e. the presence in the market of at least a few strong companies of significant size. In mergers, regulators must establish conditions to guarantee that the augmented size and capability of the resulting company will have no adverse effect on market competitiveness.
- The third is the supervision of agents' behaviour: Even where market structure is suitable on paper and the rules of the games well established, the regulator must supervise agents' behaviour. This involves monitoring, taking legal actions and penalising the infringement of rules, as well as permanently reviewing their effectiveness to achieve the objectives sought. To conclude with the soccer example, national federations attempt to persecute and punish doping, and referees are needed in the matches to make sure that the rules of the game are not violated. In the electric power industry, the regulator must ensure that agents do not abuse any dominant position they may have to manipulate prices, or that

they do not conceal relevant information from other market participants (e.g. nuclear unit outages must be immediately reported to the System Operator and the market to maximise both system operation and market transparency).

Regulatory concerns typically include a wide diversity of topics, such as: consumer prices and tariffs, quality of service, economic viability of the companies involved, environmental impact of industry activities, policies for supplying the less advantaged or others without access, market structure and market power, proportionality between investment volumes and operational efficiency and demand, and asymmetries between information available to the regulator and to market agents.

Regulatory entities have a wide range of instruments of different nature (economic incentives, structural constraints, etc.) with which to attempt to manage and influence these issues positively:

- Cost-of-service subject to regulatory oversight: as discussed briefly below and in greater detail in Chap. 4, one alternative is for the regulator to closely supervise all costs and make all items subject to its approval.
- Benchmarking of regulated monopolies: for regulated operators such as electricity distributors, the regulator may link remuneration to their comparative performance, whereby companies providing better service obtain higher revenues than the ones providing lower quality service.
- Price or revenue caps: taking that idea one step further, a ceiling may be set on the price that the supplier is allowed to pass on to consumers. The company is free to establish its expenditures, but the remuneration or profit earned is subject to a limit set by the regulator.
- Unbundling of activities: if a company conducts activities that the regulator believes can be performed more efficiently if unbundled, or if the joint exercise of activities may result in some sort of dominant position in the market, the regulator may require separation of the different businesses.
- Introduction of competitive pressure: the regulator may establish rules to enhance competitiveness among agents. It may, for instance, impose limits on larger players to enable new competitors to grow.
- Application of other incentives, such as the creation of quality standards (ISO-9001) or many other regulatory measures that may also be deployed, such as command and control (standards, targets, penalties, etc.), operating licence requirements, establishment of prerequisites for mergers and acquisitions, information gathering and analysis and market behaviour monitoring.

3.1.3 Regulatory Bodies

Three types of institutions often share authority in the industry: the ministry concerned, a regulatory commission more or less independent from the ministry and the competition authority. In federal systems, these institutions may exist at both the

central and regional government levels, as it is for instance the case in the US, with the Federal Electricity Regulatory Commission and the Public Utility Commissions at state level, or in the European Union with the Agency for the Cooperation of Energy Regulators (ACER) and the regulatory authorities of the Member States.

These three chief actors hold real power. Other organisations, such as ministerial agencies and independent advisory agencies, may also play a potentially significant role, albeit with no legally sanctioned regulatory powers.

3.1.3.1 Ministries and Ministerial Agencies

The ministry concerned is the part of the executive branch of government specifically responsible for the considered industry. The ministries that usually deal with energy affairs are the ministries of industry or economy, although some OECD countries have a separate ministry for energy affairs. Some countries have created agencies (under different denominations) associated with the respective ministry. These institutions operate with an independent budget, are independently managed and may be subject to separate legislation (because they are not subject to the rules governing civil service, for instance), although they are in one way or another ultimately subordinate to the ministry.[1] Even so, ministerial agencies operate with relative independence in many countries.

3.1.3.2 Independent Regulatory Commissions

Independent regulatory commissions (often called independent regulatory agencies, authorities, bodies, commissions or committees) are public bodies entrusted with regulating specific aspects of a given industry, a characteristic that distinguishes them from mere advisory bodies. Regulatory commissions share regulatory powers with other public institutions, the ministry concerned in particular, and are fairly free of short-term political influence. Their duties often include regulating grid access and determining grid and end-user tariffs. Regulatory commissions may have judicial or quasi-judicial powers, such as establishing fines and penalties for non-compliance or acting as arbiters in disputes between industry players. Some commissions are also explicitly mandated to protect end users and regulate entry and exit through licences, where the rights and obligations of the licensees are specified. Some commissions also intervene in the processes of authorisation of mergers and acquisitions of companies in the particular sector.

Independent commissions may be characterised by the degree of political independence and the extent of their authority. In addition to being vested with

[1] Take for instance the Colombian case: the Comisión de Regulación de Energía y Gas (CREG) is organized as a Special Administrative Unit of the Ministry of Mines and Energy, and it is chaired by the Minister of Mines and Energy.

independence and specific powers, independent regulatory commissions share regulatory responsibilities with the competent ministry. Theoretically at least, decision making (the establishment of a general framework and rules) is generally left to the ministry, while the regulatory commission's responsibility is to enforce the ministry's rules and sometimes to develop them in detail. The boundary line between policy and regulation is blurred, however, and some overlapping is inevitable. The annex to this chapter contains a fuller discussion of regulatory authority design.

Independent advisory agencies
Some countries have created independent advisory agencies. These institutions do not answer to the ministry, which cannot even revoke the appointment of agency members. While the terms of reference of their advisory mandate are broad, they are vested with no decision-making power in regulatory matters. They are often responsible for monitoring areas such as grid access and dispute settlement.

3.1.3.3 Competition Authorities

The competition authority is the entity or group of entities entrusted with enforcing competition law. It acts *ex-post* to enforce prohibitions on collusion or abuse of market power and curtail unfair competition, and *ex-ante* to prevent mergers and acquisitions that would have an adverse effect on competition. In some countries, these entities may also implement structural measures, such as ordering asset divestiture when the industry structure blocks competition. Regulatory authorities, on the contrary, act essentially *ex-ante*, establishing rules that serve as a framework for the market. Regulatory institution and competition body competence overlaps substantially in many areas, including grid access, investigation of market power abuse, price setting and structural measures governing generation, end user supply and mergers and divestiture.

3.1.4 Regulatory Models

3.1.4.1 Regulatory Reform: An Overview

It is very possibly true the statement that no two countries anywhere in the world have taken the same approach to the regulation of the power sector. And the adopted approaches change often. At a broad level of detail, the two major dimensions that characterise a regulatory model, in essence, are:

- Which activities are conducted separately from all the others?
- Which activities are conducted on a competitive basis?

In other words, models can be classified on the grounds of which activities have been "unbundled" and which deregulated. In principle, many different models can

be envisaged, given the many possible combinations of unbundled and deregulated activities. However, as noted above, some activities are not readily liberalised and this limits the number of models that it makes sense to consider in practice.

On one extreme, there is the vertically integrated monopoly, in which none of the activities are separated or deregulated. This is what is known as the traditional model. The other extreme is the full retail competition model. In this model, activities are vertically disaggregated and generation, supply and trading or transacting are performed competitively.

The early power sector developments at the end of the nineteenth century and beginning of the twentieth century were mostly driven by private initiative and competition. However, in most countries this was replaced by strong governmental intervention in the form of public ownership or treatment of the electricity companies as regulated monopolies. During most of the past century and until the 1990s, the electricity industry regulation worldwide was based on a quasi-standard regulatory approach involving heavy State planning and intervention, with the State being the sole regulator. Under this traditional model, prices are fixed to cover the costs incurred by electric utilities, while investments either require regulator authorisation or are State planned. This approach precludes any electricity market per se and transactions are conducted to the rules laid down by the regulator. Consequently, the vertical and horizontal structure prevailing in the industry is of scant importance, which often has led to the existence of a single utility, a vertically integrated monopoly with a territorial franchise.

Today, this traditional regulatory paradigm has changed in many countries. Chile (1981) started the process, followed by England and Wales (1990), Norway (1991), Argentina (1992) and many others in all continents. The main component of this change is the creation of electricity markets that provide the platform for trading and establishing prices. The first step in creating such markets is the elimination of the limitations to competition that characterises the traditional scenario. While calling for broad knowledge and technical analysis, all it involves is drafting and approving suitable legislation for the industry.

Abolishing the limitations to competition suffices to create markets for many manufactured or agriculture products such as textiles or wheat, where suppliers abound. In the electricity industry; however, regulatory reform must address an additional element, namely vertical and horizontal restructuring.

A competitive marketplace requires a certain minimum number of competitors. Since traditionally a few companies at best—and often only one—operate in a given electric power system, certain preliminary measures are frequently required. These may consist of dividing the monopoly (horizontally) into different companies, opening up the market to companies from other electric power systems or, if possible, other (interconnected) countries or facilitating market entry to newcomers.

Moreover, physical delivery of electric power is what is known as a network industry. In general, grid or network operations such as transmitting and distributing electric power cannot be performed competitively. Therefore, prior to the creation of a competitive market, these network activities, which must continue to be run as a monopoly, have to be separated (vertically) or "unbundled" from the

activities that can be conducted in competition. Measures must be taken to ensure that all market agents have grid access. This is what is known as third-party access to the grid or TPA.

3.1.4.2 Traditional Cost-of-Service Regulation

Under traditional regulation the State—via the corresponding ministry, ministerial agency or public utility commission—supervises, approves or even makes the decisions of operation and investment for the electricity sector, demanding a minimum level of quality of service in the assigned territorial franchise area and paying the cost of service in return, via regulated tariffs charged to consumers. Traditional regulation is known as cost-of-service or rate-of-return regulation.

A well developed and documented example of traditional regulation is the case of vertically integrated companies in the United States in the 1980s described in greater detail in Chap. 4 and Annex A to this book.

In practice, the tariff review procedure, known as the rate case, is the core of traditional regulation. Companies present their investment plans and estimated future operation costs, which regulatory commissions may accept, reject or propose to amend. Accepted investments are built into the tariffs paid by consumers.

Electric utilities under cost-of-service regulation are typically bound to public service mandates in their franchise area. Many voluntarily engage in transactions of lesser importance to coordinate service with neighbouring utilities. They constitute regulated monopolies. In this scheme, the electric utility makes all economic and technical decisions, i.e. planning and operation are centralised and subject to regulatory review. Public ownership and regulation frequently overlap. The company's yearly remuneration from regulated tariffs paid by electricity consumers is set periodically, annually for instance, depending on the outcome of the rate case.

Under cost-of-service regulation, the company is allowed to charge end customers its total incurred costs (investment costs plus operation costs) costs, where the investments costs include a fair return on investment. This type of regulation is analysed in depth in Chap. 4.

The chief advantage of traditional regulation is that it is meant to ensure fair prices at any given time. Since revenues are equal to incurred costs, if the regulator has done a good job, consumers do not overpay and investors are not undercompensated. This affords regulatory stability and guarantees cost recovery (via suitable remuneration), providing a favourable investment climate, reducing capital costs and guaranteeing high levels of security of supply for electricity consumers. Finally, with cost-of-service regulation, social obligations can be more readily instituted. These include social tariffs, specific R&D programmes, support of domestic fuels, energy diversification and environmental protection.

From an economic perspective, however, traditional regulation is problematic because the incurred cost being considered may be inflated, i.e. they may not be the most efficient cost. Three factors may contribute to cost inflation under traditional regulation:

- Information asymmetries: utilities have much more precise cost and demand data than the regulator, who needs them in the tariff review process. Information may therefore be manipulated by regulated companies to bring in higher revenues that cannot subsequently be recorded as earnings, but which can be earmarked for certain cost items (such as higher salaries or a larger headcount).
- Lack of incentives for efficient management: keeping costs as low as possible (for a given amount and quality of service) calls for some effort from company managers. Under the traditional system of regulation, managers have no incentive to make this effort since, if costs grow, revenues are in principle automatically adjusted to absorb the difference (see the discussion on the Averch–Johnson effect below).
- Regulator capture: utilities usually have a wealth of resources that can be deployed to influence regulator decisions in their favour. This undue influence on regulatory decisions, called "regulator capture", may be exerted in a variety of ways, including all forms of lobbying, communication campaigns, regulator hire by the regulated utilities and vice versa (so-called revolving doors).

3.1.4.3 Incentive-Based Regulation

The problems cited above have given rise in some countries to a revision of the regulatory mechanisms in place with a view to establishing explicit monetary incentives for the regulated companies (monopolies) to minimise costs while at the same time avoiding monopolistic pricing. The idea, essentially, is to allow utilities to make a profit when they are able to lower costs. This means that prices do not necessarily mirror costs at any given time but, in exchange, companies have an incentive to cut costs. If this incentive is effective, it can be expected that costs should be lower (and efficiency greater) and, in the long term, prices may also fall.

Under the traditional approach, the basic outline of incentive-based regulation can be summarised in a single idea: to refrain from calculating tariffs that reflect each year's costs. Therefore, if a utility lowers its costs more than planned, the difference between revenues and costs, or part of it, is not returned to consumers, but raises the company's earnings. The expectation of such earnings is the incentive for companies to lower costs.

The most common form of incentive-based regulation is known as "CPI-X" (or "RPI-X") regulation.[2] Under this scheme, the revenues or maximum prices that the company can charge are set for a longer period, 4 or 5 years, for instance. The revenue target is fixed on the grounds of the previous year's target, updated to adjust for inflation (RPI), less a certain efficiency, productivity or simply adjustment factor, X. Therefore, the company has a significant incentive to reduce its costs below the revenue target established ex-ante.

[2] CPI and RPI stand for Consumer Price Index and Retail Price Index, respectively. They are inflation measures.

Under incentive regulation, the company can make decisions that would not be possible under the traditional scheme. For instance, if the company believes a given investment programme to be profitable despite regulator objection, it would be free to make the investment and benefit from any operational savings ensuing.

RPI-X schemes provide for revising remuneration to compensate for circumstances that the company is unable to control, such as demand growth or variations in input costs. Target revenues, in turn, may include incentives associated with non-cost objectives such as quality of service improvements or environmental criteria.

The alleged advantage of incentive-based regulation is that prices and costs can be lowered in the long term, even if short-term prices rise. This is illustrated in the graph below (Fig. 3.1). Chapter 4 contains a more detailed description and analysis of different methods of incentive-based regulation.

The question is whether incentives work and the prices resulting from incentive-based regulation decline (as shown on the chart) or whether, on the contrary, the two price lines never intersect. A number of arguments have been put forward suggesting that the effectiveness of this approach should be viewed with some scepticism. First, company managers can logically foresee that the cost reductions achieved will eventually give rise to lower cost and price targets. Therefore, the argument goes, the incentive consisting of earning a short-term profit is weakened by the knowledge that in the medium term allowed earnings will decline. The net result of these two opposite forces, called the ratchet effect, is uncertain.

Second, criticism has also been levelled at incentive-based regulation from the standpoint of its credibility. When the result of regulation is high earnings for utilities, consumers and other actors will exert pressure to change the incentive formula. If the opposite occurs and utilities lose money, pressure is likewise expected to change the formula to prevent the quality of service from deteriorating or to escape the threat of company failure. Anticipated changes in the formula cancel the beneficial effects of incentives.

Finally, this regulatory formula has been criticised because in early stages of its practical application to power systems, it was overly generous with some companies, suggesting that the values for X and the cost targets had been inadequate, although empirical evidence in this regard is inconclusive. RPI-X has gradually become the approach of choice for the regulation of network utilities. However, it has been recently realised that a regulation that is only focused on reducing costs

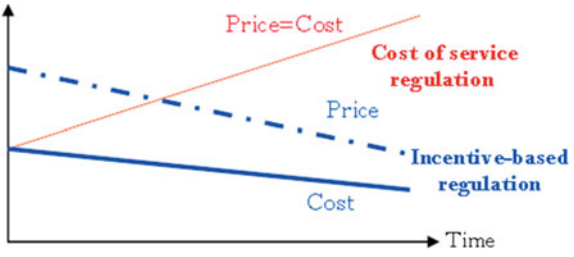


Fig. 3.1 Static and dynamic incentives in regulated industries

might stifle innovation, which is nowadays considered of essence to address the challenges of the deployment of distributed generation, the integration of information technologies in electricity networks, the utilisation of electric vehicles or more active demand participation.

3.1.4.4 Restructuring and Liberalisation

The new paradigm of electricity industry regulation is based on the conviction that competitive electric energy markets are possible. As shown in Chap. 2, competition is possible in generation today, because power systems in developed countries are large enough and new technologies have significantly reduced former economies of scale. Moreover, markets have grown with the increase in interconnection capacity between regions and countries. Since in the electricity industry, generating investment accounts for the largest share of the total investment costs, this is the area where the potential benefit from reform is most promising. By contrast, significant economies of scale can still be identified in transmission and distribution, which continue to be regarded as natural monopolies regulated under cost-of-service or incentive-based arrangements.

This basic strategy, market creation, guides the organisation of economic activity in many other industries. Its implementation differs in the electricity industry; however, because the supply of electric power depends on transmission and distribution grid monopolies whose control affords absolute power over the electricity market.

The problem is that the existence of a monopoly may preclude competition. Consequently, grid-associated activities must be independent of competitive businesses such as generation and retailing (also known as supply). Since, moreover, the baseline situation for liberalisation prior to introducing competitive mechanisms is normally a vertically integrated utility (a company that conducts all businesses from generation to end customer billing), industry organisation and ownership must usually be restructured. Business unbundling is discussed in detail in a later section of this chapter.

Regulation under the new circumstances also addresses many other matters, such as the creation of a retail market in which all consumers may exercise the right to choose a supplier. Mechanisms and institutions for coordinating organised markets, and especially the physical operation of the system, are also established. Other questions include transmission and distribution grid access, expansion, remuneration and service quality, as well as usage tariffs and the design of the transition from a traditional to a competitive market in which the legitimate interests of both consumers and producers are protected.

In short, electricity industry regulatory reform consists of unbundling network activities from the businesses that can be conducted under competitive terms and letting competition happen in the latter. Therefore, two consecutive steps are needed: restructuring and liberalisation.

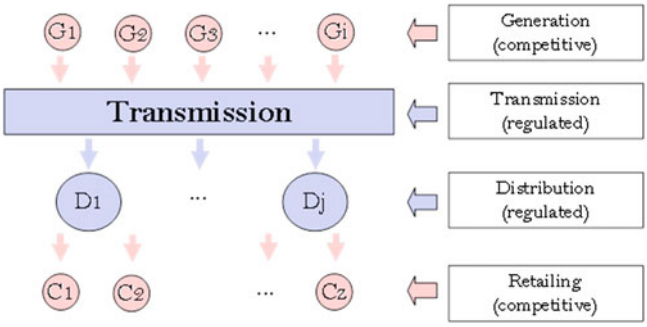


Fig. 3.2 Unbundling of electricity industry activities

Restructuring: unbundling of activities

As defined by the Body of Knowledge on Infrastructure Regulation,[3] restructuring is the separation of functions in a vertically integrated firm, leading to the unbundling of services. This first step involves separating businesses that can be liberalised (opened up to competition) from the activities that cannot.

Effective separation or unbundling is imperative for satisfactory market operation. Figure 3.2 contains a simplified diagram of the organisation of a restructured electric power industry. Competitive activities, generation and retail (or commercialisation), are separated from regulated activities, transmission and distribution.

Several important regulatory issues should be borne in mind to ensure effective unbundling.

The basic rule for separating or unbundling activities in the new regulatory environment is that no single agent should be allowed to conduct a regulated activity (such as transmission or distribution) and an activity open to competition (such as generation or retailing) simultaneously.

The reason for unbundling is the appearance of conflicts of interest when a player engages in both a monopolistic and a competitive activity. In other words, monopoly power may be used to distort competition in the monopolist's favour in the liberalised activity. Competition can be distorted or competitors discriminated against in a variety of ways:

- Cross-subsidies: the monopolist charges high prices for network services and applies the earnings to underpin competitive activities at less than cost. Competitors with no access to such revenues cannot compete at those prices or can only do so at a loss.
- Restrictions on third-party access to the grid: the monopolist can distort free third-party access to the network by creating technical obstacles to access, denying access on the pretext of insufficient available capacity, granting

[3] www.regulationbodyofknowledge.org

dispatching or access priority to its own business units or concealing information on grid availability. The existence of such practices is difficult to detect and conclusively prove.

A monopolist participating in liberalised activities has not only the capability, but the incentive to discriminate against competitors, since such discrimination is a source of higher earnings.

The separation of activities attempts to avoid this conflict of interest by eliminating either the capability or the incentive to discriminate. Unbundling may be implemented at different levels, and should be adjusted to each individual case. Regulators may determine the degree of unbundling required, which is usually one of the four listed below, shown by increasing degree of separation:

- Accounting separation: liberalised and regulated activities are conducted by the same company, which is obliged to keep separate accounts for each activity. The company must charge the same fees for grid usage to its business units and customers as it charges to third parties.
- Management separation: in addition to keeping separate accounts, each activity is managed separately and the company conducting both businesses commits to providing the same information on the regulated activity to all competitors.
- Legal separation: the regulated and competitive activities are run by legally separate entities (different companies) so as to formally prevent common financial interests from biasing decisions in favour of the competitive activities, although both firms are allowed to belong to the same corporation.
- Ownership separation: regulated activities are conducted by separate utilities that do not engage in deregulated activities and the two types of companies have different owners.

Ownership separation is often referred to as a structural remedy because it eliminates the incentive to discriminate, whereas the other forms of separation are called behavioural remedies, because they reduce the monopolist's ability to discriminate but not the incentive to do so. Economically speaking, structural remedies are preferable because they strike at the root of the problem, whereas behavioural remedies call for continuous regulator supervision, and intervention in the event of abuse. Behavioural remedies are sometimes preferred for political or legal reasons, however, because they are easier to implement. Accounting separation, for instance, is readily accepted by companies, whilst ownership separation (which entails selling assets) generally runs into strong opposition and the necessary legal grounds to enforce such action may be lacking.

Liberalisation: Freedom to choose and market implementation

In most countries, the change from an electricity industry with just one or a handful of vertically integrated monopolies to a marketplace where regulated and competitive activities are unbundled has been phased in step by step, see Sect. 7.3 for a detailed explanation. The final outcome is an industry that revolves around a wholesale market, often followed by the institution of a retail market.

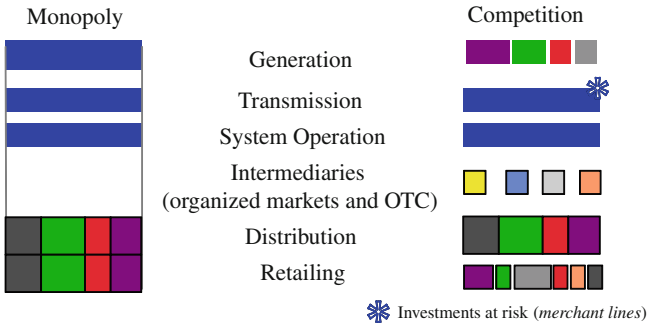


Fig. 3.3 From monopoly to market

Figure 3.3 illustrates the restructuring of a vertically integrated company, regulated as a monopoly and engaging in generation, transmission and system operation and distributing and supplying power across several regions. Its unbundling yields: (i) several generation companies; (ii) a single transmission company[4] that may or may not also include the system operation function; (iii) new entrants acting as traders or intermediaries for market operation, under the format of organised or over the counter (OTC) markets[5]; (iv) several regulated distribution companies in the different former service areas and (v) a number of competitive retailers selling energy to end customers.

Under the new circumstances, transmission, system operation and distribution companies are regulated as monopolies under cost-of-service or incentive-based arrangements. Generation, trading, market agency and retailing, on the contrary, are competitive activities in which market rules govern inter-player commercial relationships. It is up to regulators to ensure fair competition. The graph in Fig. 3.4 provides a simplified view of economic regulation and market transactions under the new organisational structure.

Competition is organised around a wholesale market, which is normally a spot market for short-term electricity transactions. Medium- and long-term contracts of different types are usually established in conjunction with or as an alternative to the spot market. The players qualified to trade on such markets are generators, authorised consumers, several categories of suppliers and, for medium- and long-term contracts that do not specify the final destination of the power, any interested agent.

In some cases, such as in many Latin American countries, only the wholesale market is operational, with no retail market. In this case, generators compete to sell energy to distributors and a few large-scale consumers on an organised pool. Distributors sell energy to commercial and residential customers at regulated prices. In other words, retail competition has not yet been implemented in these markets.

---

[4] Multiple transmission owners is another possibility. In any case, all transmission assets must be operated by a single entity, which could also own some transmission.

[5] See Chap. 9.

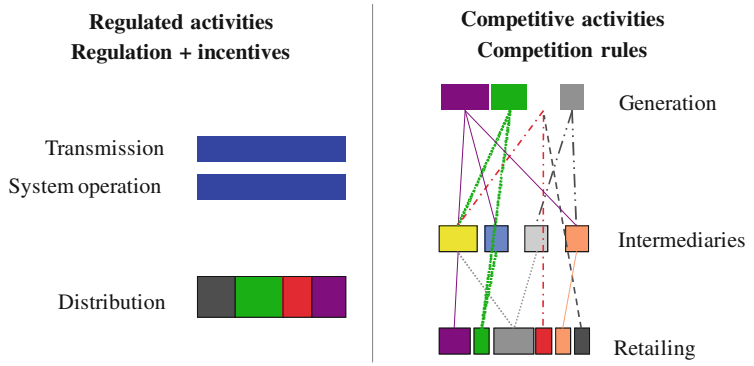


Fig. 3.4 Regulated and competitive activities in the new market structure

In Europe, however, retail competition has been progressively introduced. The latest version of the European electricity directive requires all Member States to institute full competition at retail level. In the US, the choices of different states cover the full range of retail liberalisation possibilities.[6]

An additional issue in unbundling is the separation between the two major activities related to the transmission networks: ownership—typically associated to investment and building of infrastructure—and system operation, which frequently also includes network capacity expansion planning. In most European systems, a single company that is termed the "transmission system operator" or TSO performs both activities. TSOs must be independent—although some carefully specified exceptions may exist—from the generation companies. In other countries, the operation of the system is assigned to an entity that is independent from the transmission and generation owners: the independent system operator, ISO. The ISO manages a grid that it does not own. This second approach is more frequent in countries such as the USA where the transmission grid is in private hands and largely fragmented, and ownership separation would have involved the—scarcely feasible—expropriation of the transmission grid. Similar issues are starting to emerge at distribution level.

Risk allocation issues

The liberalisation of traditionally regulated activities is meant to further the efficiency in the operation and planning of power systems, and also in the allocation of risks between producers and consumers. Since planning errors committed in regulated environments are paid for by customer tariffs, utilities have only a weak incentive to make efficient decisions. Examples of erroneous planning investment decisions under traditional regulation are not difficult to find: the nuclear development plan in Spain in the 1980s, later followed by an expensive moratorium imposed by the government, for instance; or the large hydro projects undertaken in

[6] See US Energy Information Administration webpage in the references.

Latin American countries that overran their budget, drastically increasing State debt. On a competitive market, by contrast, each agent makes its own investment decisions. Investment profitability is consequently not guaranteed and the consequences of errors are borne exclusively by the company committing them. If it is unable to recover its investment at the market price of electricity, the sums invested are lost. Carried to the extreme, this situation would require the company to withdraw from the market.

The market would efficiently allocate the risk of poor decision making between consumers and companies. The latter have a stronger incentive and are better positioned to be accurate. Consumers have a choice and only the best solutions would be chosen. Collective decision making is expected to enhance efficiency. It is as if the market is willing to pay for the best technology only. If the decision proves to be more expensive, the extra cost is borne by the investor and the consumers choosing that technology. Competitive systems are also expected to adapt quickly to technological change. If a new technology appears to be more efficient, it is quickly adopted in competitive markets, whereas traditional companies must take the fixed costs already incurred into consideration.

Many other topics must be addressed in the new regulatory environment, including: the mechanisms and institutions needed to co-ordinate organised markets or power exchanges, and especially technical system operation; access to, expansion of and remuneration for the transmission and distribution grids, as well as quality of supply and the establishment of tariffs for the use of these facilities; and the design of the transition from a traditional to a competitive market, protecting consumers' and utilities' legitimate interests. These questions are discussed in greater detail in the following chapters.

Annex A
Regulatory Authority Independence and Structure[7]

Energy market reform covers a wide range of measures that entail much more than mere legislative change. Institutional reform to adapt regulatory bodies to their new functions is a basic component of any effective reform policy. The institutional framework in place has a substantial impact on the quality and effectiveness of regulation, acting specifically on company, investor and consumer incentives and expectations.

Electricity industry reform requires regulators to engage in new tasks such as ensuring network access or protecting consumer freedom of choice. The expertise and resources needed to perform these tasks often differ from the know-how drawn on in the past. The new markets need regulators that operate impartially, without discriminating amongst the various players. Regulators must be independent of all

[7] Prof. Gaspar Ariño was a key support for the development of the annex.

stakeholders and their activities must be as transparent as possible. Rising to such challenges generally entails institutional reform.

A.1 Independence

Regulator independence is two-dimensional, since regulators must be independent from both political and stakeholder interests.

Independence from Stakeholders
Independence from stakeholders means that regulated agents have limited influence over regulatory decisions. Such independence is required to guarantee that regulation does not favour one community of interests over any other. All approaches to regulation are based on the principle that regulators should not be "captured" by the interests of industry actors.

To avoid capture, regulators are often subject to restrictions on their relationships with the regulated parties during and after their term of office. Such restrictions, which apply to regulators both in ministries and regulatory bodies, include the prohibition to hold a financial interest or a position in the industry during or after their term of office as regulators. Hiring procedures may also encourage independence if, for instance, company owners and employees are not eligible for membership on regulatory bodies. Rules on procedures to guarantee decision-making transparency and limit regulatory discretion likewise enhance independence.

Political Independence
Political independence means that regulators are protected against short-term political influence. The need for political independence is not an unquestionable principle, although in recent years it has gained considerable support. Political independence is furthered through irrevocable terms of office and other measures such as the existence of a separate budget for the regulatory agency, and independent human resource and wage management. The actual degree of political independence is contingent not only upon the adoption of such measures, but also on the powers vested in the regulatory commission. When a board has broad powers, political independence has significant implications for the regulatory framework and industry structure. On the contrary, when regulatory commissions' terms of reference are limited to technical aspects or advisory functions, the consequences of political independence are less important. Appointments to independent regulatory commissions are irrevocable.

Political independence serves three purposes. First, it reduces short-term political pressure on regulation. It is generally agreed that regulatory policies should not depend on short-term political circumstances. Energy prices should not be used as a tool to control inflation, for instance. Second, political independence can reinforce regulator independence from specific lobbies. Third, when

companies are state owned, political independence is necessary to avoid conflicts of interest between the State's roles as owner and regulator.

Despite these advantages, full political independence is difficult to attain both in theory and in practice. In democratic systems, legislative power is held by elected bodies such as parliaments. This means that even if the regulator is politically independent, regulation is subject to a certain amount of political control and influence. In the US, where regulatory commissions are politically independent, in principle, and have broad powers, market reform has furthered legislative activity in areas traditionally handled by the commissions. Full independence may or may not be desirable, at least under certain circumstances, as it may make regulatory bodies more susceptible to capture. A small and specialised regulatory body, for instance, may find it very difficult to counter industry influence on public opinion, hold its ground during a lengthy dispute with the industry or even attract the qualified and highly specialised personnel needed to fulfil its duties.

In practice, individual regulators must be appointed by a political institution, usually Parliament. Political preference is therefore reflected to a certain extent in the choice of the people involved. The shorter the term of the appointment, the greater is this influence. Moreover, even if regulators attempt to act objectively, they may be sensitive to the political situation. Regulator personality and preferences are important in determining to what extent regulatory processes are affected by political circumstances.

Fostering political independence has implications for public sector management, since it often entails the creation of a new independent body. This is costly, and more importantly may cause long-term problems due to the tendency of such bodies to grow and self-perpetuate.

By way of summary, regulatory independence from short-term political pressure fosters effective action. However, a certain amount of political control (and, with it, influence) over regulatory structures is both necessary and inevitable. Political independence is ultimately a question of degree.

A.2 Structure

Regulatory commissions may be designed in different ways, depending on their role (or "mission"), governance, duties, specific regulatory processes, resources and internal management and implementation strategy.

Objectives

The aims of regulatory commissions are generally limited to economic matters. In network industries, regulators aim to protect both users and investors: users from abuse on the part of companies with market power and investors from the possibility of arbitrary governmental action, such as the establishment of exceedingly low tariffs. Regulatory commissions usually seek to balance these two objectives. Prices are set to protect users from monopolistic practices, for instance, but

attempts are made to ensure that investors recover and earn a reasonable return on their investment. Regulatory commissions may also have broader objectives, such as favouring economic efficiency or introducing competition. The institutions that specifically serve these purposes may play an important role, for legislative constraints and regulation constitute the greatest obstacles to competition in many key industries for the economy. Regulatory commissions do not usually have social objectives, which are assigned to other institutions, such as environmental bodies and ministries. However, the growing relevance of environmental concerns in energy regulation is broadening the scope of responsibilities of regulatory commissions, as discussed in Chap. 14.

Powers

In addition to their regulatory duties, independent commissions may also assume responsibility for other areas of public activity, such as the implementation of merger policies or the enforcement of other provisions of competition law. These tasks may range from cooperation with competition authorities in the fulfilment of their obligations (market control, reports on possible infringements and advice on decisions to be adopted) to leadership in the application of competition legislation. In the latter case, the regulator and competition authority are one and the same institution.

The main reason for keeping these two institutions separate is that regulation and competition policy are different activities that call for a certain degree of expertise and specialisation. Separation would favour the regulator's specialisation in a single industry. At the same time, however, merging regulatory bodies and competition authorities may afford the industry watchdog a fuller set of tools with which to reach its objectives. When the two are separated, as in most countries, regulatory commissions generally advise the competition authority or supplement its activity.

Regulatory commissions may also play an active role in certain policy issues, such as market entry or the drafting of investment or privatisation plans. Regulators are usually assigned an advisory task in these areas. When they are given executive power, it is normally in conjunction with the respective ministry.

Coverage

The breadth of regulatory commissions' responsibilities may vary, from a single industry such as electricity to several, such as all energy industries or even all network industries. The advantages of cross-industrial regulation are listed below.

- Shared activities may lead to savings in areas such as information gathering and administration. One important factor when assessing the significance of these savings is the size of the industry regulated. Having specific commissions for each industry may be prohibitively expensive in a small country, while the impact of the savings stemming from joint commissions would be only minor in large economies.

- The risks of regulator capture and undue political influence are lower, since the commission depends less on a specific industry or group.
- Distortions among related regulated sectors due to regulatory inconsistencies can be avoided, particularly when the commodities involved are mutually replaceable, such as electricity and gas.
- The commission can directly settle problems relating to blurry inter-industry borders, which exist, for instance, when regulated companies engage in the supply of both electricity and gas.
- The expertise acquired in one industry can be applied to related businesses in questions such as third-party access to different network industries.

By contrast, specific regulators for each industry are more specialised. They are also less likely to act discretionally or waste resources, since they have less power and less information. Such regulators have other implementation-related advantages, because they can leave some room for experimentation and the impact of possible failure is more limited. Lastly, specific regulators can be more effective when regulation differs significantly between industries. This would be the case of an electricity industry open to competition and a tightly regulated gas industry, for instance. And it is certainly the case between energy and telecommunications.

Decision-making structure

Decision making may be assigned to a commission or a single person. A commission is better protected against conflicts of interest and regulator capture, given the plurality of standpoints involved. This potential also constitutes a risk, however, since pressure may be exerted to appoint a commission with a given political leaning or a tendency to represent certain parties' interests or as a microreproduction of the existing political forces. A commission may also be more stable than an entity consisting of a single person if its members' terms are staggered. At the same time, advantages to having a single regulator may also be identified: in addition to being less costly, decision making is clearly speedier and more responsible and predictable. In practice, all commissions have an odd number of members, from three to seven in most countries, to avoid tie votes.[8]

Appointment of regulators

Regulatory commission members' personal traits and professional qualifications are instrumental in commission operation and independence. In a number of cases, in fact, a change in commission membership has been seen to significantly affect its position on important issues such as prices, restructuring or competition. Consequently, one and the same set of legal provisions may lead to very different results depending on the personality and preferences of regulatory commission members.

[8] The Committee that regulates the Single Electricity Market of Ireland is an interesting example. The Committee consists of seven members: Three from the Republic of Ireland, three from Northern Ireland and one independent member from another country and without any prior ties with electricity activities in Ireland.

Integrity, competence, independent reasoning and the strength to resist pressure are indispensable in regulators. Technical experience in the industry regulated is also of essence in industries with so many distinctive features as the power sector. This renders the selection process difficult, since candidates linked to the industry are often excluded from the selection process to prevent conflicts of interest and ensure independence. Appointments may be made by Parliament or the Government or, in some countries, proposed by the government and confirmed by Parliament.

Protecting independence

A number of mechanisms may be used to protect regulator independence. Commission members are usually appointed for fixed, renewable terms (this is questionable, from the viewpoint of independence from the appointing authority), for a total of up to 7 years; their positions are seldom tenured. Their appointments are irrevocable; i.e. governments may not shorten their terms at will. They may only be dismissed under fairly extreme circumstances, such as mental incapacity or proven corruption. Provisions are also in place to prevent conflicts of interest from arising during and after their term. These include the prohibition to hold a financial interest in or receive any manner of consideration from regulated companies or accept employment with such companies after their term is over (to avoid the possibility of revolving doors). Other measures designed to reduce regulatory commission susceptibility to capture include provisions to ensure stable and predictable financing and the establishment of salaries comparable to private sector remuneration, i.e. not subject to the rules applicable to other public officials.

Regulator's duties

The duties that may be assigned to regulatory commissions include:

- regulation of monopolies (restructuring, price setting and conditions for grid access, operating rules and system safety),
- establishment of end user tariffs,
- definition of quality standards,
- oversight of company behaviour and market operation,
- enforcement of the law and secondary regulations,
- entry regulation (granting of licences and authorisations),
- drafting of reports for the government,
- conflict settlement.

The definition of tasks depends both on the regulatory framework (i.e. the purpose of regulation) and the institutional structure (how regulatory tasks are distributed across the commission, ministry and other institutions). The possibilities range widely. Regulatory duties may be assigned to a single organisation, or less frequently to two or more bodies (such as in the United Kingdom, where licences may require approval by both the ministry and the regulatory agency). Legislative bodies may change the regulatory scenario by approving new laws. Legislative action should change the rules only exceptionally, for legislation

pursues stability. That notwithstanding, the need for reform has intensified legislative activity in the electricity and gas industries in recent years.

An industry regulator is generally responsible for economic regulation, addressing the consequences of imperfect competition and monopoly issues, or the latter only. Regulatory commissions rarely assume responsibility for social regulation, which would include questions relating to the redistribution of wealth, such as alleviating the "lack of fuel" by affording low income families access to basic energy services. Regulatory duties may also be divided between two organisations, one to write the rules and another to enforce them, in an attempt to prevent conflicts of interest. Such separation would aim to highlight the fact that formulating legal provisions is a more "political" endeavour than enforcing them. In practice, however, the boundary between the two is difficult to define.

Yet, another consideration arises in federally structured countries: the division of regional and nationwide responsibilities. The advantage of nationwide regulation is its uniformity. It is more efficient than the regional approach when markets overrun regional boundaries. Nationwide commissions also have access to more information and avoid duplication in some regulatory processes. Regional regulation, in turn, leaves room for experimenting with different approaches and favours innovation. Where industry conditions differ between regions, decentralised regulation can adapt policy to local circumstances. In practice, nationwide and regional regulators' responsibilities vary enormously and reflect the degree of political decentralisation in place in a country. As a rule, regional regulators engage in activities with a fairly minor impact on other regions, such as distribution and retailing, whereas nationwide regulators' tasks cover activities with a broader geographic scope, including international and inter-regional trade.

Regulatory processes and appeals

Regulatory activity may focus on the ex-ante formulation of regulations, with a view to guaranteeing impartiality and justice for all stakeholders. In such cases, it tends to be explicit and formal. Alternatively, it may focus on reaching a consensus among the stakeholders, a process that tends to be less formal. Third, the regulator's activity may be geared to supervising the industry and adopting corrective action wherever necessary, in which case it focuses on supervision. All regulatory schemes combine these three approaches to some extent. Some procedures are more advisable than others for certain regulatory duties. The consensus approach may be more suitable for technical and operating standards, for instance, whereas the establishment of the revenues allowed to a regulated monopoly may require a more formal, almost "judicial" procedure. Some systems, such as in the US, have adopted the ex-ante approach, while in others, such as in Finland, have been based on ex-post control and resources.

All independent regulatory commissions are subject to standards that favour transparency and predictability and provide stakeholders the opportunity to submit their opinions as alternatives to the commission's position. This process includes formal consultation processes, "hearings" and advisory committee meetings prior to decision making, as well as the publication and justification of the decisions

adopted. In some systems, these procedures are nearly judicial, whereas less formal processes are in place in others. Moreover, all systems establish procedures for appealing decisions either to an independent institution or a governmental body. In some cases, the grounds for appeals are restricted to objective or procedural errors to guarantee that the appellate institution cannot reverse a regulator's resolution. Other systems, however, allow the appellate body to review the content of a decision, effectively limiting the regulator's power. Regulatory design almost always entails a trade-off between foreseeability on the one hand, which limits regulatory risk, and simplicity and effectiveness on the other. As a rule, more formal processes reduce regulatory risk but call for more resources, such as advisors and legal auditors.

Coordination

The larger the number of regulatory institutions, the greater is the need to coordinate their action. All public bodies must adopt a concurrent approach to avoid the inefficiency that stems from having to deal with multiple organisations that together form a "regulatory labyrinth". Coordination contributes to guaranteeing regulatory consistency and lowers the costs incurred by stakeholders having to deal with different authorities and comply with their regulations. Coordination may be based on informal cooperative agreements or formal rules that require mutual inquiry or the exchange of reports or information on certain subjects.

Funding

The availability of suitable financial and human resources is essential to effective regulation. Commissions usually operate with an independent budget approved by the Government or Parliament, which ensures some administrative independence. The origin of the funds may be general taxation allocated under the national budget, or more frequently specific charges levied on industry participants, such as end consumer quotas or grid usage tolls. In the latter case, the funds may also be allocated through national or regional budgets to increase transparency and facilitate control. The size of the budget is important for effective operation: insufficient resources compromise the commission's ability to fulfil its tasks, whereas overly generous resources may result in a lack of focus, as well as a waste of scarce public funds. Financial stability over several budget periods is also important to ensure effective management and independence, as mentioned earlier.

Human resources

The questions posed by human resource management in regulatory commissions are similar to the issues faced in other public organisations. In any event, staff competence and specialisation must be guaranteed. Regulatory commissions need specialised, qualified experts for whose services they must compete with the private sector. Salaries are therefore often established in accordance with market standards. Moreover, regulatory commissions tend to call in outside consultant firms for certain services and resort to specialised training programmes, particularly in the early phases.

Reports and auditing
A number of mechanisms are in place to audit agency governance and ensure that its activities are conducted responsibly. Most agencies report to Parliament or the respective ministry. They are also subject to audits and other controls, which are normally performed in accordance with the procedures applied in other public organisations.

Implementation strategy
Certain transitional problems arise when a regulatory agency is created. The time of their launch, whether before or after reform, affects how reform is perceived by the players concerned and defines the commission's role in the formulation of regulatory provisions. Initial financing, particularly if the regulatory commission was created in the wake of reform, may be allocated by the government. Staff transferred from the industry or the ministry may provide the expertise required, which would not otherwise be available.

A.3 Other Matters

Another factor to be considered in regulatory commission design is the interaction with other institutions and stakeholders. Official procedures may be established to regulate the relationship with ministries, competition authorities and other institutions. Regulation may be governed by detailed schemes to determine stakeholders' rights and obligations ("regulation by contract") or be based on more general rules. These legal issues are defined in depth in the country's administrative and legal systems.

Basic references and sources

1. The International Energy Regulatory Network provides information on electricity and gas regulation and it is maintained by the energy regulatory commissions. www.iern.net
2. Public Utility Research Center, University of Florida, The Body of Knowledge of Infrastructure Regulation. www.regulationbodyofknowledge.org
3. Comparing regulatory agencies. Report on the results of a worldwide survey, European University Institute, Working Paper RSCAS 2009/63. http://hdl.handle.net/1814/12877
4. Council of European Energy Regulators (CEER). CEER Publications and Press. www.energy-regulators.eu/portal/page/portal/EER_HOME/EER_PUBLICATIONS
5. US Energy Information Administration. Status of Electricity Restructuring by State. www.eia.gov/cneaf/electricity/page/restructuring/restructure_elect.html
6. Asociación Iberoamericana de Entidades Reguladoras de la Energía (ARIAE). www.ariae.org
7. International Directory of Utility Regulatory Institutions. www.worldbank.org/html/fpd/psd/ifur/directory/index.html
8. Ocaña, C (2001) Regulatory institutions in liberalised electricity markets. OECD, Paris

Two Interesting Journal Special Issues

9. European Review of Energy Markets—volume 3, issue 3, October 2009. http://www.eeinstitute.org/european-review-of-energy-market
10. Energy—Electricity Market Reform and Regulation, Volume 31, Issues 6–7, Pages 745–1114 (May–June 2006)

References for In-depth Regulatory Study of Specific Topics and Power Systems

11. Sioshansi, FP and Pfaffenberger, W (eds) (2006) Electricity market reform. An international perspective, Amsterdam, Elsevier
12. In: Joskow, PL Introduction to Electricity Sector Liberalization: Lessons Learned from Cross-Country Studies, pp 1–32, Elsevier
13. Batlle C, Barroso, LA and Pérez-Arriaga IJ (2010) The changing role of the State in the expansion of electricity supply in Latin America. Energy Policy 38(11) 7152–7160. Nov 2010, doi: 10.1016/j.enpol.2010.07.037

Already Classic, Pioneer Texts on Power System Restructuring and Liberalization

14. Schweppe, FC, Caramanis, MC, Tabors, RD, Bohn, RE (1988) In: Spot pricing of electricity. Kluwer Academic Publishers, Boston
15. Joskow, P and Schmalensee, R (1983) In: Markets for Power. MIT Press, Cambridge

Chapter 4
Monopoly Regulation

Tomás Gómez

For a few decades after September 1882, when Thomas Alba Edison commissioned the first electric power plant on Pearl Street in New York City, the electric power industry was composed of small firms in competition with one another. Then, for many years until barely two decades ago, it has been considered as a natural, regulated monopoly.

This chapter is an introduction to the fundamentals of natural monopoly regulation, particularly as it applies to utilities providing what are regarded to be public services: electricity, water, telecommunications and gas, although given the subject matter of this book, the focus is logically on electricity.

Monopoly regulation is justified by the need to prevent a monopolistic provider from overcharging consumers or delivering a service of unacceptable quality, given the loss of economic efficiency this would entail for society as a whole.

When the monopolist is a state-owned company, the state can act as both owner and regulator without clearly separating the two functions. On the other hand, when the monopolist is a private company or there is a separation of functions, the state takes responsibility of establishing the regulation, whose implementation and supervision may be delegated to regulatory commissions. For instance, in the US, regulatory commissions were created to control utilities in all 50 states, and were in place throughout the twentieth century. The US also has nation-wide regulatory commissions such as the Federal Energy Regulatory Commission (FERC) and the Federal Communications Commission (FCC), whose mandates cover interstate transactions or services that fall outside the regulatory powers of individual states.

The preceding decade has often been said to have witnessed the collapse of natural monopolies, but this is not entirely true. For a while, competitive markets have in fact been created for telecommunications, electricity and gas as well as water supply services, these markets have focused on wholesale production and retail sales of the product in question: information, kilowatt-hour of electric or thermal power, cubic metres of water and so on. However, building and

T. Gómez (✉)
Instituto de Investigación Tecnológica, Universidad Pontificia Comillas,
Alberto Aguilera, 23 28015 Madrid, Spain
e-mail: tomas.gomez@upcomillas.es

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_4,

maintenance of the necessary network infrastructures continue to operate as natural monopolies (this is only partly true in telecommunications).

One of the chief characteristics of electrical network infrastructures, which depends on large investments, is that they are "bound" to the physical space where they are located. Indeed, one often cited example that clearly illustrates the inefficiencies inherent in introducing competition in this type of activities is the duplicate expense involved in two competing electricity distribution companies building the same type of infrastructure in the same area to provide the same service. Their networks would be redundant and users would end up paying roughly double the price for the same service.

The first section of this chapter establishes the fundamentals and principles on which monopoly regulation is based and introduces the variables that can be regulated: company revenues, quality of service standards, past and future investment and the conditions to be met by the exclusive licensee.

The second section gives a detailed description of the procedures normally followed by regulators to establish revenue ceilings and tariffs under the most traditional regulatory method, known as cost-of-service or rate-of-return regulation. The advantages and drawbacks to this approach are also discussed.

The third section addresses the differences between incentive-based and cost-of-service regulation and justifies the growing acceptance of the former as an alternative to enhance efficiency where electric power transmission and distribution are separate and regulated activities. The two most popular methods, price and revenue caps, are described at length.

Finally, the last section contains detailed information intended to provide an in-depth understanding of the various stages and basic elements involved in a price review process, and other issues that should be regulated, such as quality of service.

4.1 Fundamentals of Monopoly Regulation

A monopoly exists when, for whatever reason, a given company becomes the sole supplier of a product or service. Further to the discussion in Chap. 2, markets are the most efficient way of producing and selling goods or services. Under certain circumstances, the conditions required for an acceptable level of competition are not present, however. In other words, markets can fail as a result of market concentration, economies of scale, public goods, externalities, incomplete information or transaction costs. Market failures must be corrected by regulatory intervention to ensure the optimal outcome for society. When economies of scale exist, economic reasons can be put forward to explain why a monopoly is the only feasible option to provide a product or service. Natural monopolies may be characterised by one or more of these features: (i) economies of scale, (ii) capital intensity, (iii) non-storability with fluctuating demand, (iv) location-specific delivery generating location rents, (v) production of essential services for the community and (vi) direct connections to customers. While all these circumstances

are present in the electricity industry, some segments of this industry, generation and retail, can migrate from a monopoly to a market structure. The transmission and distribution grids, however, must continue being regulated as a natural monopoly.

An unregulated monopoly would be in a position to charge consumers a price much higher than its production costs, with the concomitant loss of economic efficiency. Consequently, some form of regulation is required to ensure efficiency under these circumstances [19].

The electricity industry has been traditionally dominated by national or regional monopolies subject to price control regulation, with the regulator setting new tariffs from time to time. After the wave of industry deregulation and liberalisation in the 1990s, the generation and sale of electric power to end consumers are viewed as activities that can be conducted under competition, whereas network activities, i.e., electricity transmission and distribution, are considered to be natural monopolies and still in need of regulation. The justification, in terms of economic efficiency, is immediate: quite obviously, allowing two or more companies to build power lines across the same region to supply the same community of consumers with electricity would be both prohibitive and wasteful.

According to economic regulation theory, the monopolistic supplier of products or services regarded to be in the public interest should be prevented from exploiting its market power, either through due regulation or by some other form of control, such as state or pubic local institution ownership. Yet, the same ideal of economic efficiency that underlies the operation of perfect markets—namely, that the profit motive induces innovation, investment in new projects and reduced costs, and competition puts downward pressure on prices—should also inform the design of efficient monopoly regulation.

Regulators may choose from a number of regulatory variables as tools to reach efficiency objectives [27]. Arguably, the most important of these approaches is the regulation of the revenues from electricity sales that the company is allowed to earn. Such revenues must be sufficient to enable the utility to cover its operating costs and make any necessary investments, while earning an adequate return on the capital invested. In other words, revenues should ensure the company's medium- and long-term economic and financial viability, without driving it to bankruptcy. Conversely, such revenues should not be detrimental to consumer interests. Moreover, in the case of an essential service such as electricity, unjustifiably high prices also have an adverse impact on the competitiveness of a country's industry.

Generally speaking, the (conflicting) objectives that appropriate regulation should pursue in determining the allowable volume of regulated revenues are:

- Economic and financial sustainability of the utility.
- Productive efficiency, attempting to provide the service or product at the lowest possible cost.

From the standpoint of tariff design for the end users, the following objectives must be borne in mind:

- Equity, whereby the receipts from any given community of consumers cover the cost incurred by their consumption, ruling out cross-subsidies among consumer groups.
- Pricing efficiency, whereby the amounts charged should be kept as close as possible to the marginal costs of providing the service or product.
- Sufficiency, whereby the receipts from tariffs concur with the revenues allowed by the regulator.

Where network companies are privately owned, sight should not be lost of the large, immovable investment required to build the associated infrastructure. If the regulator yields to political or populist pressure to lower prices, so far that the return on the company investment is insufficient, the result will very likely be a firm decision on the part of the utility not to invest even the amounts strictly necessary. Regulation should strike a proper balance between operating costs and capital investment. If heavy investments are made at the expense of sufficient staffing; for instance, the outcome is not economically efficient. If, on the contrary, the company skimps on investment to pay shareholders higher dividends, the medium-term quality of supply will be compromised.

Another variable that regulators may draw on to enhance efficiency is the service quality standards that the company is required to meet. In the electricity industry such quality is related to: (1) reliability of supply, i.e., the number and severity of power supply outages; (2) voltage quality, defined as the existence or otherwise of disturbances that may affect the proper operation of apparatus and equipment connected to the mains and (3) consumer satisfaction with the service standards, for instance time for providing new connections, maintained by the company. All these quality indicators are directly related to operation and maintenance costs, but also to company investment and the quality of the installed infrastructure.

Regulation that excessively encourages cost cutting or lower investment may lead, as noted above, to a gradual deterioration of the quality of the electric power delivered to consumers. For all these reasons, the regulator should explicitly establish performance standards and link them to the revenues the company is allowed to receive.

A variable that sometimes is explicitly controlled by regulators is the investment in new infrastructure proposed by the company in its transmission or distribution grid. This also has to do with the problem discussed above. The regulator's primary long-term objective is to ensure that sufficient installed capacity is built to meet the expected demand at suitable levels of quality. This, as noted, is directly related to the rate of return on investment and any deviation, upward or downward, has undesirable consequences. The regulator may attempt to solve this difficult problem by establishing criteria to assess the suitability and necessity of the investments proposed by the company. One example can be the use of network planning models for cost/benefit evaluation of the proposed investments. The increase in regulatory costs entailed in the concomitant control and information gathering must also be taken into consideration, however.

Yet, another variable controlled by regulators is the entry into or exit from the business by companies other than the incumbent monopolist. This variable is very important where market regulation is concerned. In the case of electricity distribution utilities, regulation implies an agreement that grants the supplier the right to distribute electric power in a territory exclusively, in exchange for submission to regulatory control. One of the conditions typically imposed on the monopolist is the obligation to supply power to all users, regardless of the associated cost, since electricity is an essential service. Companies may attempt to refrain from servicing high-cost consumers or areas and focus on areas where costs are lower. Regulators must require operators to also provide service of sufficient quality in higher cost remote or rural regions, which must be remunerated accordingly. Alternatively, regulators may establish incentives for other potential suppliers to enter the market: granting territorial franchises to small local cooperatives, for instance.

Practically speaking, the aim is to define the regulatory design that achieves an optimal trade-off between efficiency and quality of service. The following sections review the two most common types of regulation: the traditional method used in the electricity industry for many years, known as *cost-of-service* or *rate-of-return* regulation, and an alternative regulatory instrument, which is an extension of the previous one, and has become increasingly popular in many countries, known as *incentive-based* regulation.

Other less common regulatory methods are also used, and are mentioned here briefly to complete this section. In *yardstick competition*, the tariffs that a company is allowed to charge are calculated on the basis of the costs declared by other regulated operators in the same business [23]. Implementation of this scheme requires substantial amounts of comparable information on the characteristics and costs of a number of companies. Each operator's remuneration is normally established on the grounds of statistical analyses that determine average efficiency patterns for the group as a whole. This gives companies an incentive to lower their costs, since the method for setting revenues is unrelated to their own individual balance sheet. As a result the yardstick value will decline, lowering prices to end consumers. The implementation of yardstick competition also encounters practical problems, however. The regulator must have data for a sufficient number of similar companies, but the companies involved may not all be similar enough to assume that the differences among them are due exclusively to different degrees of efficiency; collusive behaviour may be encouraged, and dissociating revenues from costs may drive some companies into financial straits. The yardstick approach may also be viewed as an input to incentive-based regulation, when benchmarking techniques are used to compare relative efficiency among regulated companies in the same industry, as described in Sect. 4.4.4.

A third type of regulation is *light-handed* regulation, exercised by the company itself under regulatory supervision. Under this scheme, the company sets the tariffs to be charged to consumers subject to regulatory approval, which may include mandatory changes in tariff structures and rates.

4.2 Cost-of-Service Regulation

In cost of service, also known as rate-of-return regulation, the tariffs charged by the utility are authorised and set by the regulator. Tariffs are periodically "re-negotiated" by the regulator and the company in what are known as rate case proceedings. The regulatory process essentially entails two successive stages. The discussion below refers to the way this type of traditional regulation has been implemented in the US and applied to vertically integrated utilities, which may vary slightly from the arrangements in place in other countries.

The revenues the company is allowed to receive (rate level) are determined in the first stage, a process that involves: (1) identifying company total costs and investments and (2) establishing the allowed rate of return, to provide the utility with suitable remuneration for the capital invested. Rate cases are based on the data furnished by the company for the preceding accounting period and a forecast of the needs of expenditure for the next control period, and the tariffs established remain in effect for the following period, until the next rate case revision. In the US, the rate hearings happen at irregular intervals.

The second stage consists of determining the tariff structure, in other words, of defining the tariff components to be charged to each type of consumer for each cost item. These tariffs are designed to enable the company to collect the revenues allowed by the regulator as calculated in the first stage. For further details on tariff design and calculation, see Chap. 8 of this book.

The rate case or revision process usually comprises the following steps [21]:

- The regulator decides to initiate the tariff revision process because the established period has lapsed, or more frequently, because the utility makes the request of a rate revision.
- After the company submits detailed accounting information and "negotiates" with the regulator, the latter determines the rate of return to be applied to the capital invested and the suitable level of expenses to be covered.
- Finally, tariffs for end users are adjusted to the allowed revenues calculated in the preceding stage. This requires taking into account the energy demanded in the period in question; as this may change with prices depending on demand elasticity, information is likewise required on the latter parameter.

4.2.1 Determining the Cost of Service

Box 4.1 presents a typical general breakdown of the allowed revenues for a vertically integrated electric utility. The revenues are intended to recover the total costs of providing the service, which must include the generation, transmission, distribution and retail activities. The considered costs essentially include operation and maintenance costs, return on capital, depreciation and taxes.

Box 4.1. Regulated Revenues Under Cost of Service

The following accounting formula represents the balance that the regulator must strike when reviewing a rate case.

$$AR = TC = O\&M + DP + s \times RB + TAX - ADR \quad (4.1)$$

where

AR	is the allowed revenues,
TC	is the total cost of service,
O&M	is the allowed operating and maintenance costs,
DP	is the depreciation expenses on the company's gross assets,
s	is the allowed rate of return,
RB	is the rate base, a measure of the value of the company's investment, calculated as its net assets, defined to be its gross assets less depreciation,
TAX	is the taxes for which the company is liable, and
ADR	is the additional revenue.

The precise definition of terms in the previous formula for vertically integrated electric utilities used in cost-of-service regulation in the US follows (see the American standards on systems of accounts for electric utilities [6] for further details).

- Operating and maintenance costs: these include the cost of fuel, material and replacement parts, energy purchases, supervision, personnel and overhead.
- Depreciation: the straight line method is generally used. Fixed assets in construction progress are not depreciated.
- Tax: all the taxes for which the utility is liable, i.e., on profit, revenue and property, as well as social security and construction tax (except as relating to fixed assets in progress, since such tax is built into the value of the asset).
- Rate base: this includes net fixed assets (plants, transmission and distribution facilities, other tangible and intangible fixed assets and nuclear fuel, less the cumulative depreciation for all these items), plus current assets (fuel and other material and replacement part inventories, advance payments and deferred revenue, research and development expenses and current asset requirements).
- Rate of return: this is the average weighted interest rate on the company's long-term financial resources (bonds, debt certificates, shares and preferred shares).
- Additional revenue: this consists of the expenses/revenues deriving from the sale of the company's property, revenues from wholesale energy sales and other revenues not directly related to producing electric power.

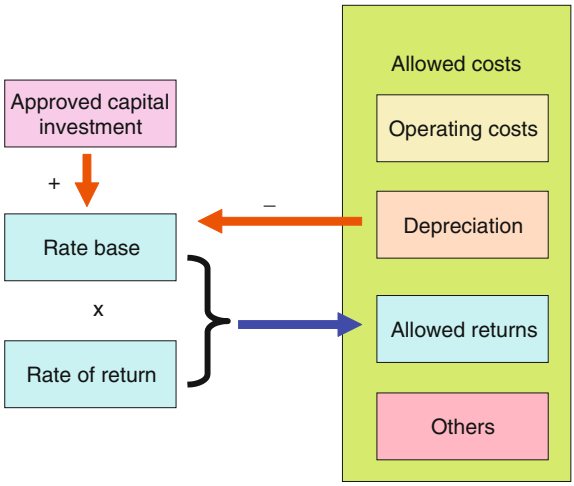


Fig. 4.1 Allowed revenues under cost-of-service regulation

Figure 4.1 illustrates how the regulator calculates the allowed cost of service. Allowed costs are the sum of the operational costs (operating and maintenance costs) plus rate base depreciation expenses, plus the allowed returns (rate of return times the rate base), plus miscellaneous (taxes minus additional revenues). As the figure shows, the approved capital investment is added to the rate base and the depreciation is subtracted, and the allowed returns are calculated as the rate base times the allowed rate of return.

Two issues take most of the effort in the negotiations between utility, regulator and stakeholders to determine the allowed revenues for the next price control period: the allowed rate of return, *s,* and the investments to be included in the rate base. As may be deduced from Eq. (4.1), ultimately the key figure is the product of these two quantities. Note, also, that the above formula takes account of the *allowed* O&M expenses, which may not necessarily be the same as the actual expenses incurred. This provides an incentive for the company to enhance its efficiency and constitute a financial penalty for inefficient company management. The simple numerical case example in Table 4.1 presents the list of the accounting items used to calculate the rate of return, in keeping with Eq. (4.1). An adjustment is made in the calculation of the allowed revenues to improve the allowed rate of return.

Volumetric tariffs are obtained by dividing the allowed revenues by the amount of estimated energy consumption ($/MWh). Volumetric tariffs are common in the US, at least for small consumers.[1] When the resulting tariffs (per MWh) remain

[1] Volumetric tariffs are nowadays criticised in the context of pricing distribution and transmission services, because the companies increase their profits by increasing sales. This incentive can act as a barrier for the implementation of energy efficiency and demand response programs. In many US states, revenue decoupling from sales has been proposed as a remedy to counteract this problem. See the website of the National Action Plan for Energy Efficiency http://www.epa.gov/cleanenergy/energy-programs/suca/resources.html.

Table 4.1 Example of utility accounts and calculation of rate case (in k€) [21]

	Anticipated rate case	Adjustment	Rate case after adjustment
Revenues	30,000	1,600	31,600
Expenses			
Fuel	24,000		24,000
Operating	3,000		3,000
Depreciation	1,000		1,000
Total expenses	28,000		28,000
Net operating revenues	2,000		3,600
Rate base (RB)			
Assets minus depreciation	42,000		42,000
Working capital	350		350
Total rate base	42,350		42,350
Rate of return (%)	4.72		8.50

unchanged until the next revision, the system encourages the company to cut costs. If it incurs lower operating costs than anticipated in the rate case, it obtains a higher rate of return. Conversely, if costs are higher than anticipated the rate of return will decline. The incentive to lower costs rises with the length of the regulatory period between tariff revisions, a circumstance known as regulatory lag that underlies the incentive-based regulation approach discussed in Sect. 4.3. If mid-term tariff revisions are allowed and the regulator constantly adjusts the rate of return, however, there will be no incentive for the company to reduce costs.

4.2.2 The Rate Base

One important aspect of tariff revision during rate case negotiations is the criterion used by the regulator to calculate the rate base (RB) and determine which investments should be included. The rate base may be determined in a variety of ways, as described below.

- On the grounds of the original investment, i.e., the sum paid by the company for its facilities, less the cumulative depreciation. This is also known as the book value of the asset.
- On the grounds of the cost of reproducing the investment in question today, reproduction costs, i.e., an estimate of the present cost of rebuilding or re-purchasing the facilities acquired by the company.
- On the grounds of replacement cost or the new replacement value (NRV), i.e., the cost involved in replacing existing assets with new upgraded facilities available on today's market technology and costs, but which serve the same purpose.
- On the grounds of the market value of assets, i.e., the value that the company with its assets would command if sold on the market.

In the US electricity industry, regulators have traditionally used book value to assess the rate base and have focused their efforts on adjusting the rate of return; hence the name *rate-of-return regulation*.

For electricity distribution networks, if the book value is available and the technology is stable, there is no good reason to use another method, so no unnecessary economic risk is created for the company. In case new technologies appear, incentives should be devised for promoting innovation in the new investments. The other methods for rate base estimation can be used when a reliable rate base does not exist to start with, as it has been frequently the case in power sector privatisation processes around the world.

4.2.3 The Rate of Return

The most common method for defining the allowed rate of return is to calculate the weighted cost of the different forms of financing used by the company, such as bonds or shares traded on equity markets. This manner of calculating the rate of return is known as the weighted average cost of capital (WACC). An example is shown in Table 4.2. The most controversial item is usually the rate of return on the company's own capital, i.e., the remuneration of its equity, which the regulator typically establishes in accordance with the rates set for regulated companies with a similar level of risk: electric power distribution compared to gas distribution, water supply or communications, for instance, or the actual rates of return of other firms, again with an estimated similar risk level. This issue is discussed in greater depth in Sect. 4.4.3.

4.2.4 Strong and Weak Points

A classical weak point of cost-of-service regulation as the method for setting the rate of return is the so-called *Averch-Johnson* effect [3]. A company allowed a rate of return higher than the true cost of capital has an incentive to over-invest, giving rise to economic inefficiency. Conversely, if the rate of return is lower than the cost of capital, the utility will invest very little and its operating costs will rise, likewise generating economic inefficiencies. Further to that reasoning, regardless of whether in the real world the company behaves as predicted by the Averch-

Table 4.2 Example of calculation of allowed rate of return

	Capitalisation ratio	Rate (%)
Bonds	28	4.00
Quoted shares	12	12.00
Equity	60	6.50
Total	100	6.46

Johnson model, this type of regulation requires the regulator to accurately calculate the utility's real cost of capital.

One of the problems encountered in the tariff revision process is the existence of information asymmetries between the regulator and the company. This asymmetry is mitigated by the regulator requesting as much information as possible, and the company, which must do the work involved in furnishing it. The more involved the regulators become in investment planning and operating cost management, the greater is their understanding of the problems they must regulate. The trade-off, however, is the higher cost of regulation, as more specialised personnel and more sophisticated analytical tools are required, and more burden is also placed on the company.

Cost-of-service regulation has been criticised for not furnishing suitable incentives for companies to reduce costs when tariffs are revised frequently, for instance every year or two. Year after year, the company can recover all its duly substantiated costs and accepted by the regulator. The approach described in the following section, known as incentive-based regulation, attempts to correct some of these shortcomings.

Depending, then, on how it is implemented in practice, cost-of-service regulation has advantages and drawbacks.

The advantages associated with cost of service when judiciously applied are: (i) in principle, it enables the company to cover its costs, thereby providing financial stability; (ii) the cost of capital is a control parameter for the regulator, and must not be set above what is strictly necessary to sustain an efficient activity and (iii) it provides for a good balance between optimal investment levels and quality of service, if measures are taken to prevent over-investment in the event the rate of return allowed is too high, by lowering such rate (A-J effect).

On the contrary, when this type of regulation is implemented ineptly, the advantages may be offset by the drawbacks. Some of these are: (i) this type of regulation does not encourage efficiency and in some cases has evolved towards intrusive and legalistic regulation and (ii) it may encourage over-investment justified by technical arguments and provides an incentive for companies to incur higher costs, ultimately leading to higher prices for consumers.

Box 4.2. Cost of Service in the US Electric Power Industry During the 1980s[2]

In the US and many other countries with a privately owned electricity industry, some general criteria were followed to determine "fair" electricity rates in the context of vertical integrated utilities. Electric power rates were supposed to:

[2] A detailed version of the traditional US power sector regulation in the 1980s is presented in Annex A of this book.

- remunerate electric power suppliers and expenses incurred in providing the service,
- equitably distribute costs among all users, as far as practical given the limitations of metering and similar facilities,
- provide a reasonable return on capital and attract new resources of funding to finance any new facilities needed to cope with demand growth,
- reward service quality and system operating efficiency,
- promote revenue stability over time to facilitate planning for the future, and
- be simple enough to be readily applied by utilities and understood by consumers.

As noted above, ratemaking can usually be broken down into two stages: (i) obtaining the revenue requirement (total revenue authorised by the regulator), and (ii) formulating the rate structure for each type of user. US regulatory commissions traditionally have focused on determining the total revenues, a utility was to receive for providing the service, trying to ensure that it obtained a reasonable (neither excessive nor insufficient) return on its capital. Regulatory commissions also usually addressed the question of the allocation of total costs to different types of users by defining the tariffs for each one: residential, industrial or commercial; nonetheless, some of them did not actively exercise their regulatory powers in this connection.

The revenue requirement determined the extent to which revenues covered operating expenses, provided for a return on invested capital and were able to attract new funding. In practice, determining the cost of service was an extremely complex exercise. The most controversial aspects of this process were as follows:

- use of actual current or estimated future values (fuel or capital costs for instance) to prevent rates from lagging behind real costs (regulatory lag), a problem intensified by the duration of the regulatory process,
- establishment of a fair return on shareholder equity,
- valuation of assets on the basis of their historic cost (usual practice in the US), replacement cost or at a "fair value",
- inclusion or otherwise of the fixed assets in progress in the rate base, and
- accounting treatment for the deferred taxes generated when accelerated depreciation methods are applied.

In practice, regulatory commissions have dealt with all these issues in different ways. A discussion on some of the most relevant aspects follows. The regulatory implications of some of these issues are analysed in Sect. 4.4.

The rate base was defined to encompass current assets and net fixed assets, with the latter accounting for the major share of the base. Net fixed assets were evaluated to be the sum of the non-depreciated property used for company operations. The treatment for net fixed assets varied widely from

one regulatory commission to the next. The essential questions were: what should be included, when it should be included and at what value?

Standard practice in assessing fixed assets was to use the historic or original value, which was normally much lower than the current replacement value; certain regulatory commissions allowed utilities to use the replacement cost (computed from extrapolations based on the costs actually incurred), or a "fair" value, which each commission established at its discretion.

One particularly controversial area was the inclusion or otherwise of fixed assets in progress in the rate base. When inclusion was not permitted, no provision was made for remuneration for the fixed assets in progress, and the utility was authorised to include the respective interest costs of the capital invested in the works in progress item. If all or part of the fixed assets in progress was included in the rate base, the interest on the part in question was not charged. The trend in the years discussed here was for a growing number of regulatory commissions to allow the inclusion of the fixed assets in progress, to avoid sudden hikes in the rate base, undue interest fund growth during construction and excessive haste in commissioning facilities. A number of different procedures were in use to appropriately account for variations in the rate base throughout the year.

A peculiar American accounting practice known as standardised accounting lay at the root of another variation in the way the rate base was computed. This procedure consisted basically of keeping fictitious parallel accounts used exclusively to determine the company's tax liability each year: instead of the usual straight line depreciation used in the real accounts, these parallel accounts provided for accelerated depreciation to reduce taxes in the earlier years and increase the liability in subsequent years. This deferred tax was accumulated in a fund. Most American electric utilities used standardised accounting: the effects of this practice on the determination of cost of service and the tax treatment are discussed in greater detail by Suelflow [24].

One consideration of great practical importance for utilities in the reference period was the time that frequently lapsed between when a rate was calculated and when it was actually applied. This arose as the combined effect of two factors: the practice of calculating rates on the basis of the cost and energy sales figures prevailing when an application for a rate change was submitted, and the relatively long time lapsing between that date and when the regulatory commission's resolution was forthcoming. The procedures in place to alleviate this problem are listed below.

- Use of estimated values to cover the time expected to lapse before the new rates would come into effect. This practice was still uncommon at the time; however; regulatory commissions were much more prone to allow slightly higher rates to implicitly compensate for this shortfall.

- Clauses on automatic adjustment of rates to accommodate changes in fuel costs. Critics of this procedure sustained that it discouraged companies from seeking less expensive alternatives when fuel prices rose.
- Adjustment of the rate base as new facilities became operational.

The bone of contention in the ratemaking negotiations between electric utilities and regulatory commissions was the determination of the long-term rate of return on the company's financial resources (bonds, debt certificates, preferred shares and shareholder equity: normal shares plus reserves). Given that the average interest rate on debt (bonds and debt certificates) and the average rate of dividends on preferred shares were in fact known in advance, the problem was merely a question of determining the rate of return on common equity.

The general legislation in effect[3] at the time provided that the rate of return on common equity should (a) be comparable to that of other investments with similar risks, and (b) be sufficient to inspire confidence in the soundness of the company's financial position, so it could raise new capital when necessary. Much has been written about this important question in ratemaking (see [11] for a detailed discussion). In practice, regulatory commissions would set this rate after hearing the experts' opinions and considering issues such as the method adopted to assess the fixed assets, the estimated lag between the period when rates were to be applied and when they were calculated and the inclusion or otherwise of fixed assets in progress. The model used to quantify the cost of a utility's own capital on the grounds of the risk associated with the various types of businesses is known as the capital asset pricing model or CAPM, discussed in detail in Sect. 4.4.3.

4.3 Incentive-Based Regulation

The basic principle behind incentive-based regulation is the establishment of relatively long intervals between price controls or "rate cases". In each price control period, which generally covers 4 or 5 years, a specific revenue path is defined to create an incentive to lower costs and thereby increase profits. After the regulatory period lapses, all cost components are thoroughly reviewed (in a process similar to the rate case proceedings described in the preceding section). The outcome of this review is a new formula to set revenues or prices for the following regulatory period. In other words, this type of regulation weakens the relationship between prices and costs; in a way, it can be regarded to stand midway between cost-of-service regulation and deregulation with market-defined pricing. There are

[3] Supreme Court, Hope Natural Gas Case, 1944.

two basic alternative schemes for incentive-based regulation, the *revenue cap* and the *price cap*. These schemes may be combined with schemes for profit- or loss-sharing with users to reduce risk.

Box 4.3. A Scheme for Illustrating the Power of Incentives

Laffont and Tirole [12] introduced a simple model to establish the link between traditional and incentive-based regulation:

$$\mathrm{TR} = (1-b) * (\mathrm{costs})_{\mathrm{ex-ante}} + b * (\mathrm{costs})_{\mathrm{ex-post}} \qquad (4.2)$$

where

- TR is total revenues (ex post);
- b is a coefficient, $0 < b < 1$, established *ex ante*, to incorporate allowed costs;
- $(\mathrm{costs})_{\mathrm{ex\text{-}ante}}$ are the estimated allowed costs (ex ante);
- $(\mathrm{costs})_{\mathrm{ex\text{-}post}}$ are the costs incurred (ex post).

The incentive for the utility to reduce costs is inversely proportional to the value assigned to parameter b. Cost-of-service regulation with frequent rate cases can be likened to a *low incentive* type of regulation in which $b = 1$, since the company is allowed to recover all the costs incurred. A number of regulatory schemes with strong incentive mechanisms, in which b is close to 0, have been devised and include, for instance, (1) tariff freezing arrangements, (2) cost-of-service regulation with long intervals between rate cases or (3) incentive-based regulation with long regulatory periods.

Incentive-based regulation involves the *ex ante* raising of the proportion of the revenues that the company receives and the *ex post* lowering of the revenues calculated on actual costs incurred. In this way, the key point of incentive regulation is passing the efficiency gains achieved by the company to consumers in the next regulatory period. Regulation involving incentive mechanisms scantly adjusted to *ex post* costs may lead to excessively high or low total revenues or may even jeopardise the feasibility of the company's long-term business plan. Hence the need for a rate case, or price control review, every few years. Regulatory design should strike a balance between lower short-term costs, global economic efficiency and long-term viability.

4.3.1 Price Cap

In the *price cap* approach, a formula is used to set the maximum yearly price that the company can charge for each service provided, for a period of several years. These prices are adjusted annually to account for inflation minus a correction factor associated with expected increases in productivity:

$$\bar{P}_{m,t} = \bar{P}_{m,t-1} \times (1 + \mathrm{RPI}_t - X) \pm Z \tag{4.3}$$

where

- $\bar{P}_{m,t}$ is the maximum price that the company can charge for service m in year t.
- RPI_t is the annual price variation per unit (retail price index, RPI, or inflation rate) in year t.
- X is the productivity factor per unit.
- Z is adjustments owing to unforeseen events beyond the control of the utility, such as natural catastrophes, environmental regulation or tax hikes.

This form of regulation has been used in the United Kingdom, where it is known as "RPI – X", to regulate telecoms, gas and electricity network utilities, and in the United States to regulate telecommunications companies, under the term "CPI – X". The version known as "revenue cap" (see below) is presently used in many power systems worldwide.

Cost-of-service regulation with tariffs frozen for a period of time covering several years can be viewed as price cap regulation with no correction for enhanced productivity.

Price patterns under a price cap scheme for a regulatory period are shown in Fig. 4.2. When X is greater than 0 real prices decline. The larger the value of X, the more will consumers benefit. If the company is able to lower its costs by improving efficiency more than required by the regulator, it also gains. In some situations, prices may tend upward rather than downward: where the regulator recognises high investments in a given regulatory period, for instance. At the end of the price control period the regulator will establish the price for the following year, according to the updated estimates of costs, and will set a new value of X for the next period. The benefit for the consumers during the next period results from the combination of the new price for the next year and the new value of X.

The general RPI or CPI is not the optimum indicator for this purpose, because it reflects variations in consumer prices but not in specific industry costs. In practice, mixed indicators found as a weighted average of individual industry (or company) cost indices and the general inflation rate are also used.

In practice, the price cap formula may be applied to: (i) the average price for the company as a whole, after duly weighting each of the services provided; (ii) the average price to be charged to each consumer class or (iii) the fixed price of each of the terms comprising the end user tariff. The regulator may choose either to (1)

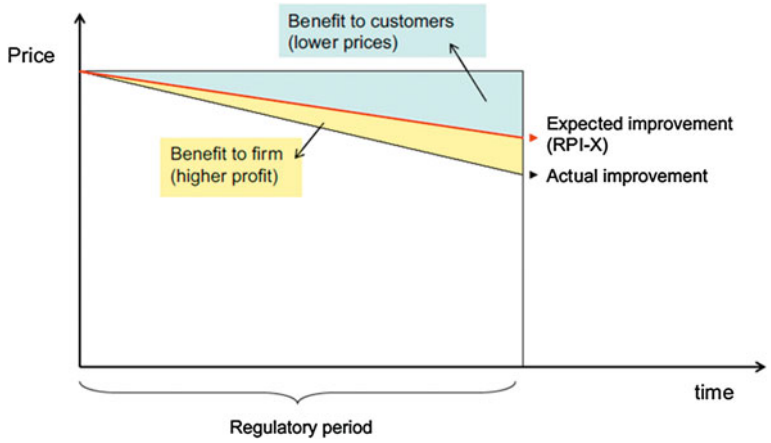


Fig. 4.2 Price trends in price cap regulation [1]

allow the company flexibility to design its end consumer tariffs or (2) design the end user tariff itself to prevent cost shifting among consumer classes.

The price cap used to regulate electric power distribution companies in the United Kingdom, for instance, known as the "average revenue" or "revenue yield" cap (see next item), sets the maximum revenue per sales unit. Sales are calculated as the mean amount of energy delivered to the consumers in each class or voltage level. Moreover, providing they honour this limit on unit revenues, companies are free to design their end user tariff structure, albeit under regulator supervision.

In Chile, Peru and Argentina, on the contrary, the regulator sets the maximum prices that can be charged for the various cost items, such as investment, operation and maintenance, customer management and the adjustment coefficients applicable to those prices throughout the regulatory period. Price caps are included in the resulting formulas for calculating the distribution tariffs to be paid by end consumers. In this case, price caps are translated directly to end-user tariff design.

4.3.2 Revenue Cap

Under the *revenue cap* approach, the maximum yearly revenues, the company is allowed to earn for a period of several years, are calculated with a formula that makes provision for yearly inflation less a correction factor associated with expected improvements in productivity. These revenues may be adjusted annually in accordance with one or several cost or revenue drivers that are beyond the control of the regulated company, such as the number of consumers, total energy supplied or, in network companies, the size of the network. Provision is also usually made for adjustments to compensate for exceptional events beyond the company's control.

The simplest expression for a revenue cap is given by:

$$R_t = R_{t-1} \times (1 + \mathrm{RPI}_t - X) \pm Z \tag{4.4}$$

where

- R_t is the allowed remuneration or revenues in year t
- RPI_t is the annual price variation per unit (retail price index, RPI or inflation rate) in year t
- X is the productivity factor per unit
- Z is any adjustment owing to unforeseen events beyond the control of the utility, such as natural catastrophes, environmental regulation or tax hikes.

Another possible way to express the revenue cap formula is shown in the following equation, in which the cost driver is the number of consumers [5]:

$$R_t = (R_{t-1} + \mathrm{CGA} \times \Delta\mathrm{Cust}_t) \times (1 + \mathrm{RPI}_t - X) \pm Z \tag{4.5}$$

where

- CGA is the consumer growth adjustment factor (currency unit/consumer)
- $\Delta\mathrm{Cust}_t$ is the variation in the number of consumers in year t.

Another revenue cap formula used by regulators is:

$$R_t = R_{t-1} \times (1 + \mathrm{RPI}_t - X) \times (1 + \alpha \times \Delta D_t) \pm Z \tag{4.6}$$

where

- α is the economies of scale factor, typically lower than 1, which provides an indication of how much the regulated costs and therefore revenues increase in proportion to a cost driver, represented by D.
- ΔD_t is the increment in year t per unit of the selected cost driver (the formula could include just one or several cost drivers), such as units of energy supplied, number of customers, network size or any combination of the three.

Revenue caps set at the beginning of the regulatory period and adjusted yearly in terms of RPI-X only are known as "fixed revenue" caps. "Variable revenue" caps are adjusted yearly by both the RPI-X factor and other cost drivers.

In yet another variation on the revenue cap theme, known as "revenue yield" or "average yield" caps, a yearly cap is set on the average revenue per unit of demand supplied. As in the case of power distribution companies, in the UK, demand can be expressed as a combination of the number of customers and the total energy delivered. For that reason, "revenue yield" caps are sometimes classified as price caps. The revenue yield cap can be formulated as (notation as in Eq. 4.6):

$$\frac{R_t}{D_t} = \frac{R_{t-1}}{D_{t-1}} \times (1 + RPI_t - X) \pm Z \tag{4.7}$$

In the revenue cap approach, the end-user tariffs defined by the regulator on the grounds of company proposals must be designed, so the total revenues do not exceed the allowed revenues. Since receipts may, of course, be higher than the cap in certain cases, an adjustment mechanism should be designed to correct for such deviations. This adjustment mechanism could also take into consideration when actual revenues fell short of the allowed revenues by providing the corresponding compensation to the company next year.

Although the incentives to lower costs are similar in revenue and price cap schemes, increases in sales generate very different effects in the two. Whereas both the price and the revenue yield cap encourage higher sales, revenue caps, depending on how remuneration for sales growth is incorporated in the formula, is always more neutral to this effect. It may therefore be deduced that compatibility between remuneration for the utility and its energy savings or demand-side management programmes can be more effectively attained with revenue caps. Moreover, due to the nature of network-related costs, the greater share of the costs of the regulated network company varies little with demand in the short and medium term. Consequently, revenue caps including appropriate cost drivers duly adjusted for economies of scale are the most popular scheme for price control in this type of regulated businesses. In the US, a good number of state regulatory commissions have implemented "revenue decoupling" from sales when regulating utilities in different forms of revenue caps, providing examples of how this partial decoupling can be designed and work in practice [14].

4.3.3 Mechanisms for Sharing Earnings and Losses

A variation on rate-of-return regulation often used in combination with price or revenue caps is what is known in the literature as the sliding scale. The chief property of this method is that it provides a mechanism for the utility and consumers to share the risk of very high earnings or losses.

Box 4.4. Sliding Scale

Expressed in cost-of-service or rate-of-return regulation terms, this scheme is based essentially on an adjustment of new rate case prices, so that the allowed rate of return in year t, s_t, adopts the following form [27]:

$$s_t = s_t + h(s^* - s_t) \tag{4.8}$$

where

- h is a constant parameter ranging in value from 0 to 1 and established by the regulator,
- s_t is the rate of return that the company would obtain in year t as a result of the tariffs set in the current rate case and prior to the application of this adjustment, and
- s^* is the target rate of return.

If $h = 1$, tariffs are set in each rate case to ensure the company a rate of return of s^*, i.e., *rate-of-return* regulation. If rate cases are scheduled at short intervals, the company fails to benefit from efficiency or to lose money for inefficiency. On the contrary, if $h = 0$, the result is constant or frozen tariff regulation, with all the earnings or losses attributed entirely to the company, i.e. *fixed-price* regulation. A value of $h = 0.5$, in turn, would indicate that the profit or loss would be shared by the company and its customers.

Sliding scale ratemaking may also be viewed as a mechanism for adjusting tariffs to maintain the rate of return within a pre-established margin [5]. Under this approach, the frequency of rate cases would be reduced by lengthening the regulatory interval. If earnings are so high that the rate of return exceeds the highest value on the scale, tariffs are revised downward. And conversely, if earnings dip below the lowest value, tariffs are revised upward. If the rate of return remains inside the established range, tariffs are not changed and the earnings or losses are attributed to the company. Practically speaking, then, this mechanism divides earnings and losses between the company and its customers.

Such sliding scale mechanisms are sometimes used by regulators in combination with price or revenue caps to protect both company and consumers from the risk of extreme earnings or losses that fall outside what is regarded to be a reasonable margin.

Sliding scale mechanisms may be progressive or regressive. In progressive mechanisms, the part of the profit retained by the company rises with cost savings. For example, a company might receive 20 % of the first 5 % saved, 40 % of the second 5 % and so forth. Since cost savings are more difficult to achieve as total cost declines, progressive mechanisms that provide for retaining a higher proportion of the savings constitute more effective incentives for the company than regressive mechanisms.

Examples of sliding scale mechanisms applied to performance-based ratemaking (see the end of next Sect. 4.3.4) for electricity companies in the US can be found [5].

4.3.4 Design of an Incentive-Based Regulation Scheme

Designing an incentive-based regulation scheme such as a price or a revenue cap calls for making certain key decisions, addressed in the discussion below (see [16, 25] for further details).

Definition of the regulatory period
Generally, four or five years is normally a good compromise, as it leaves sufficient time to create incentives for the company to lower its costs (productive efficiency) without running the risk of prices or revenues deviating too far from costs (seeking both financial viability for the utility and cost-of-service efficiency for consumers). If particularly significant changes take place that entail substantial deviations from the initial estimates for the regulatory period, a rate case may be initiated to review the price or revenue formula before the regulatory interval expires. Today, there is a discussion on the need of enlargement of regulatory periods for better ensuring the recovery of riskier investments in new clean energy technologies and innovation in energy infrastructures.

Determination of baseline and adjustment parameters
Throughout the review and before the regulatory period begins, regulators must proceed to analyse all the information furnished by the company on past costs and future investment plans. Thereafter, they must reach a decision on the costs that the company will be allowed to recover in the following period. This may involve the use of detailed analyses of each cost item, and wherever possible engineering or econometric models that can also process data on the other companies in the group for benchmarking purposes.

In practice, regulators classify costs in several categories.

- *Operating expenses* (OPEX) cover personnel, maintenance and operation.
- Capital expenditure (CAPEX) is related to investments. From the standpoint of the annual allowed cost this heading includes: (i) yearly depreciation, and (ii) the return on the rate base (RB) or the regulated asset base (RAB), which includes both the existing installations not totally depreciated and projected investment during the price control period.
- Uncontrollable costs refer to taxes, upstream fees and other exceptional cost items.

Regulators may use one of two approaches to set the baseline for revenue or price caps and adjustment parameters. These parameters determine how revenues or prices evolve during the regulatory period. The chief factors involved are inflation, productivity and variation in market size: the variation in the number of consumers or energy delivered, for instance. The two approaches are known as the building blocks and the total expenditure (TOTEX) approach. For a fuller discussion see [1, 18].

Under the building blocks approach, the regulator assesses the OPEX, CAPEX and uncontrollable cost caps separately for each year of the regulatory period.

OPEX and investment efficiency can be determined by benchmarking. At the end of the period the regulator monitors both OPEX and CAPEX efficiency gains.

The building blocks approach is used to calculate total maximum allowed revenues directly, which is the sum of each year's OPEX, CAPEX and uncontrollable costs. A revenue cap formula is calculated as follows: a starting value is determined and adjusted yearly to account for inflation and productivity (X factor) so as to equate allowed revenues to the total costs calculated for the year.

Under the TOTEX approach, separate caps are not explicitly defined for regulated OPEX and CAPEX. Rather, a single TOTEX cap is calculated by the regulator. In TOTEX, a company's productivity factor X for efficiency improvements is usually determined by the regulator by benchmarking. In theory, this approach affords the regulated company greater freedom to make the optimal choice in the trade-offs between OPEX and CAPEX to reduce costs.

Since, in practice, regulators use a mix of procedures or strategies to define the revenue cap formula, the resulting method can seldom be classified entirely under one approach or the other. Regulators may, for instance, use the building blocks approach to calculate the revenue cap formula and then implement the cap under a TOTEX scheme, whereby the company is assessed on the grounds of total savings.

Definition of secondary objectives associated with the incentive regulation scheme

In addition to the primary objective pursued with this type of regulation, i.e., lower costs, other specific aims relating to the characteristics expected of the service provided by the utility may also be sought. These objectives are related to company performance. This type of mechanism is also known as performance-based ratemaking or regulation (PBR), where the company's remuneration depends on its actual performance in meeting the specific targets defined by the regulator. Examples of such objectives include improvements in quality of service and consumer satisfaction, universal service for all consumers within the franchise area, reduction of the environmental impact caused by the company's activities and implementation of programmes in the public interest such as research and development or energy efficiency schemes. The regulatory scheme must include explicit financial mechanisms to reward or penalise the company for reaching or failing to reach the targets set for each objective.

4.3.5 Strong and Weak Points

As noted above, the main advantages to incentive-based regulation are related to the provision of clear and simple incentives for efficiency through cost reductions. In addition, the amount of information required from the companies and the cost of regulation itself are lower than under traditional cost-of-service regulation with frequent price reviews [9]. Incentive regulation has also proved to be very useful in countries with scantly developed auditing systems, where state-owned companies

were divided and privatised. In such cases, incentive regulation had to strike a balance between company and consumer interests under conditions in which the information available was incomplete.

The implementation of incentive-based regulation schemes also has its weak points, however. The first is the potential for *a decline in quality of service*. The incentives for companies to lower costs may have an adverse effect on service quality. Facility maintenance and investment costs are directly related to the quality of the service delivered. Consequently, the regulator must necessarily set both quality standards and financial penalties for the failure to meet them.

Excessive concern over the profits that companies may be making may lead regulators to gradually revert to cost-of-service regulation. Such a concern may induce regulators to increase the frequency of rate cases to review costs, ultimately slipping back into traditional regulation and losing the potential benefits of cost reductions.

In keeping with the pursuit of fair profit sharing between customers and the regulated company, key aspects that the regulator should resolve in each price control review are the starting point, the initial price or initial revenue and the choice of both the X factor and the inflation index, which may include a mix of price indicators. All these decisions impact the delicate balance between profit sharing and achieving efficiency and stability in the medium and long term.

Since incentive regulation may lead to laxer cost supervision by the regulator, companies may tend to *shift the costs incurred in the non-regulated line of business to their regulated activity*, if this is the case (e.g. in an attempt to show that their profits are falling and thereby obtain a larger tariff hike than would be strictly necessary). In addition, monitoring OPEX and CAPEX separately where the requirements for CAPEX are less strict than OPEX introduces the risk of *operating costs being capitalised* through investment [15]. One way of tackling this problem is to impose strict rules that require regulated companies to furnish accounting information on standardised forms with a high level of cost disaggregation.

In schemes that allow for company flexibility in establishing the end-user tariff structure, mechanisms should be in place to prevent the *shift of costs to consumers with fewer options* or influence over the company: from large industrial to residential consumers, for instance.

4.4 Implementation Details for Price or Revenue Caps

This section contains a practical description of some of the notions and issues mentioned above that should be considered by regulators during price control reviews when setting caps for the next regulatory period.

First, an alternative viewpoint of the "productivity factor X" will be presented. So far in this chapter, X has been considered as an external input to the process, which is obtained by knowledge about the potential efficiency improvements of the

considered company or comparison with similar firms. In practice, this use of X may be possible with some kind of companies or with specific activities within a company (e.g. operation and maintenance costs), but not with others (e.g. investment costs). Therefore, the use of a global productivity factor X encompassing the diversity of costs that a company has to incur to deliver a product might be considered impractical.

An alternative view focuses on the estimation of the efficiently incurred costs that the company will incur during the next control period. Each cost item (e.g. capital costs of the existing and new investments, operation and maintenance costs, costs of R&D, uncontrollable costs) is individually proposed by the company and acknowledged by the regulator, who takes into account the potential gains in efficiency that are possible in that particular cost component. This is how "productivity improvements" are accounted for in this approach. No "X factor" is used so far. Once the estimated costs for all items have been obtained for all years of the price control period, the net present value of the costs is computed for the initial year using a convenient discount rate. Then Eq. (4.4) is employed to determine a trajectory of allowed revenues that has the same net present value as the estimated efficient costs. In this approach, the role of X is reduced to be an "adjustment" or "smoothing" factor, whose mission is to make sure that the net present values of the streams of costs and revenues for the duration of the price control period are exactly the same. This is the viewpoint that will be adopted in what follows in this section.

4.4.1 Present Value of Costs and Revenues: The Smoothing X

Some basic cost definitions were given in Chap. 2. Now we introduce the concept of present value of costs and revenues, i.e., the change in value of money with time. In rate cases, present value calculations are used to determine the revenues required to cover the utility's expected costs (including a suitable rate of return on assets). This constitutes an economic, as opposed to a financial, analysis of cash flows in the years comprising the regulatory period.

Under cost-of-service regulation, each year's revenues should equal the sum of the annual operating costs, the rate of return on net assets during the year and the annual depreciation on gross assets (see Eq. 4.1).

Assuming that both costs and revenues for the year materialise at year end, all these amounts should be discounted at the beginning of the period to calculate the present net value.

Net assets, in turn, should be re-calculated each year by adding the new assets resulting from investment made during the year and subtracting the annual depreciation on gross assets in service.

The table below contains a sample calculation of revenues in cost-of-service regulation for a price control of 2 years duration. Estimated costs and allowed

Table 4.3 Present value of utility costs and revenues [7]

Value	Item	Start	End-year 1	End-year 2		Key to columns	
	Interest rate	10 %			r		
Actual	Operating costs		50	48		OC_1	OC_2
	Depreciation		10	10		D_1	D_2
	Investment		20			I_1	
	Assets	100	110		A_0	$A_1 = A_0 + I_1\text{-}D_1$	
	Return on assets		10	11		$Ret_1 = r*A_0$	$Ret_2 = r*A_1$
	Discount factor		0.90909	0.82645		$d = 1/(1 + r)$	d^2
Discounted	Operation costs		45.455	39.669		OC_1*d	OC_2*d^2
	Depreciation		9.091	8.264		D_1*d	D_2*d^2
	Return		9.091	9.091		Ret_1*d	Ret_2*d^2
	Revenue		63.636	57.025		$rev_i = oc_i + dep_i + ret_i$	
Actual	Revenue		70	69		rev_1/d	rev_2/d^2

Net present value of estimated incurred costs (2 year example)

revenues for years 1 and 2 are discounted at the starting point (i.e. at the beginning of year 1) (Table 4.3).

In price or revenue cap regulation over a regulatory period of several years, regulators may distribute the utility's revenues in different ways in each year, as long as its present value for the entire price control period is maintained. Then, the X factor can be calculated to smooth the trajectory of revenues across the regulatory period. As a consequence, revenues are set in a fashion that ensures gradual variation throughout the period. This is achieved by the *smoothing X*, computed by equating the present value of the revenues calculated according to the revenue cap formula (4.4) during the whole regulatory period to the present value of the allowed costs calculated year by year. The outcome is more gradual variations in prices and the corresponding tariffs.

Table 4.4 illustrates the calculation of the smoothing X.

First, the regulatory asset base (RAB) is updated yearly taking into account allowed investments and depreciation. All the figures in bold are input data. Next, the rate of return or WACC is set by the regulator. Note that the rate of return is expressed in real values, i.e. no explicit provision for inflation is taken into account in this exercise, and the WACC is a before tax figure, i.e. the revenues that the company would obtain before paying taxes are taken into consideration. This subject is explained in greater detail in the following items. In this simplified example, the demand has been assumed to be constant during the entire price control period.

The regulator then sets the allowed operational costs for each year of the regulatory period. In our example, these costs decline over time due to the

Table 4.4 Calculation of the X factor in a revenue yield cap scheme

Regulatory asset base	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Opening RAB		450.00	467.50	473.00	481.25	489.15
Gross capital expenditure Investment		**50.00**	**40.00**	**45.00**	**47.00**	**50.00**
Regulatory depreciation		32.50	34.50	36.75	39.10	41.60
Closing RAB	**450.00**	467.50	473.00	481.25	489.15	497.55
Average RAB		458.75	470.25	477.125	485.2	493.35
WACC (real pre-tax)(%)		**8**	**8**	**8**	**8**	**8**
Regulatory depreciation	Life	Year 1	Year 2	Year 3	Year 4	Year 5
Existing assets (15 years)	**15**	30.00	30.00	30.00	30.00	30.00
New investment in year 1 (20 years)	**20**	2.50	2.50	2.50	2.50	2.50
New investment in year 2 (20 years)	**20**		2.00	2.00	2.00	2.00
New investment in year 3 (20 years)	**20**			2.25	2.25	2.25
New investment in year 4 (20 years)	**20**				2.35	2.35
New investment in year 5 (20 years)	**20**					2.50
Total regulatory depreciation		32.50	34.50	36.75	39.10	41.60
Operational costs (OPEX)		Year 1	Year 2	Year 3	Year 4	Year 5
OPEX		**60.00**	57.90	55.01	51.98	49.90
X factor for OPEX (%)			**3.5**	**5.0**	**5.5**	**4.0**
Allowed costs		Year 1	Year 2	Year 3	Year 4	Year 5
Return on assets (WACC * RAB)		36.70	37.62	38.17	38.82	39.47
Depreciation		32.50	34.50	36.75	39.10	41.60
Operational costs (OPEX)		60.00	57.90	55.01	51.98	49.90
Total allowed costs (Revenue requirement)		129.20	130.02	129.93	129.90	130.97
Smoothed revenues	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Quantities		**100**	**100**	**100**	**100**	**100**
Average prices (revenue/quantities)	**1.320**	1.313	1.306	1.298	1.291	1.284
Smoothed revenues		131.27	130.55	129.84	129.12	128.41
Present value analysis						
PV (total allowed costs) (WACC is discount factor)	518.85					
PV (smoothed revenues) (WACC is discount factor)	518.85					
PV difference	0.00					
Calculated X-factor	0.55 %					

regulator's efficiency requirements, i.e. an ad hoc productivity factor for OPEX that is estimated by the regulator,[4] and are also expressed in real terms, i.e., net of inflation.

Each year's total allowed costs are the sum of the return on assets, plus yearly depreciation, plus the allowed OPEX.

In this exercise, the smoothed revenue requirements are calculated in such a way that the average price in each year, revenue R_t, divided by the estimated

[4] Not to be mistaken with the global smoothing X factor in Eq. (4.4), which is the outcome of Table 4.4.

quantities delivered, D_t, can be expressed as shown below, where the global smoothing factor X is given in percentage and the initial value in year 0, R_0/D_0, is known:

$$\frac{R_t}{D_t} = \frac{R_{t-1}}{D_{t-1}} \cdot \left(1 - \frac{X}{100}\right) \tag{4.9}$$

This type of cap was referred to as the "revenue yield" cap in Sect. 4.3.2. The inflation rate was not factored into the above equation, because this exercise was performed with real costs and prices. Nonetheless, including inflation in all the cost items is straightforward.

The X factor is calculated with an Excel spreadsheet using the "goal seek" tool, by equating the present value of costs to the present value of revenues and taking the WACC set by the regulator as the discount factor. The computed X can be considered to have a double meaning. On the one hand, it is the adjustment factor that makes possible that the net present values of costs and revenues are equal; on the other hand, X can be seen as a global productivity factor.

Note that, while OPEX declines by around 4.5 % yearly, the CAPEX rises in keeping with the investment pattern approved by the regulator, causing the RAB to climb progressively. As a result, the total productivity factor, at 0.55 %, is much lower than 4.5 % but still positive, signifying a reduction in average prices in real terms.

One final observation that should not be overlooked is that sales or the quantities delivered by the company remain flat throughout the regulatory period. In other words, projected investment and operating costs are calculated on the assumption of constant sales. In revenue yield schemes, revenues rise with sales. Variable revenue cap formulas in which economies of scale are taken into consideration call for explicitly calculating the rise in investment and operating costs with increasing sales volumes or selected cost drivers. In other cases, the regulator makes direct estimations of the diverse costs for each year of the regulatory period without recurring to prescribed formulas. The reader is invited to experiment with this more realistic problem with a spreadsheet similar to the one used in Table 4.4.

A critical issue when designing price or revenue caps is how to connect prices or revenues between two consecutive regulatory periods. The incentives for the company to lower costs are greater when prices move gradually from the final year of the previous regulatory period to the target point in the last year of the new regulatory period (option 1 in Fig. 4.3). This approach allows the regulated company to retain the benefits stemming from cost reductions achieved in the previous regulatory period for a longer time. Exceptionally, if the profits expected under this scheme are regarded to be too high, the average prices at the beginning of the new period may be changed in one step (option 2 in Fig. 4.3). Observe that under this latter option the regulated company would have a disincentive to be efficient at the end of each regulatory period, because the achieved efficiency gains would be quickly passed to consumers as a price reduction.

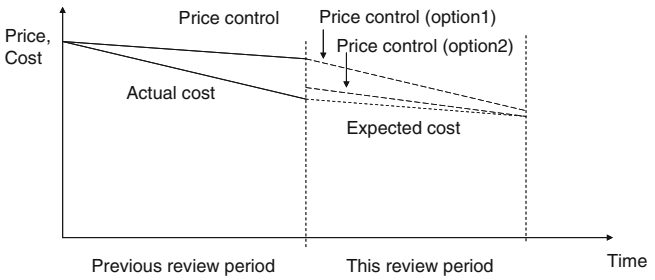


Fig. 4.3 Projected costs and prices: smoothed revenue pattern [7]

4.4.2 Regulatory Asset Base, Investment and Depreciation

The determination of the regulatory asset base (RAB), i.e. the company's net assets, is of crucial importance in the tariff revision process.

As noted earlier, the RAB is updated year by year by adding investment and subtracting depreciation. A key issue in rate cases or price controls is the determination of the RAB at the beginning of the regulatory period and the criteria for inclusion of new assets in the RAB.

The RAB for existing facilities

The first step in RAB assessment is the decision on how to evaluate the existing assets. This is necessary when a comprehensive method, such as the one described here, is implemented for the first time. Also if the adopted procedure to update the existing assets at the end of a regulatory period requires a complete evaluation (see below). This evaluation is typically done during deregulation and restructuring processes entailing, for example, the separation of regulated and competitive activities, the unbundling and privatisation of state-owned companies or the unbundling of vertically integrated enterprises into separate business units.

The four most popular methods to determine the initial RAB or, if needed, subsequent RABs, as mentioned earlier, are: book value, reproduction cost, replacement cost and market value.

Book value is reliable, providing procedures are in place to audit the general accounts and compare the entries to the aggregated and disaggregated investment in facilities. This criterion usually yields stable values that ensure a suitable rate of return while preventing prices from spiking and the company from earning overly large profits. It eliminates a major source of regulatory risk.

Reproduction cost is an estimate of the current cost of a replica of the assets considered. For instance, the reproduction costs for distribution assets are the current costs of building the present distribution grid, i.e. the electric lines and substations installed by the company in the past. The use of reproduction costs instead of the original cost rate base constitutes an attempt to assess capital costs at the current rather than the historic values. It creates a significant regulatory risk for network companies, but it might be appropriate in case of privatisation and

unbundling of former state-owned companies where no reliable accounting data are available.

Replacement cost refers to the cost of replacing current assets with new assets featuring the same functionality but updated technology. To use the distribution grid example, the replacement value of the lines and substations in a given region would be calculated by assuming that the lines and substations installed would be replaced by new lines and substations perfectly planned to meet the expected demand using the most efficient technologies currently available. This method is used in a number of Latin American countries, where the new replacement value (NVR) of the distribution network is calculated at the beginning of each regulatory period with the *model firm approach* [22].

The use of the replacement instead of the reproduction cost is in line with the economic theory that propounds using long-term marginal or incremental costs when pricing a service or product, from a forward-looking rather than a backward-looking perspective [11]. While in theory this is the most economically efficient way of appraising the value of facilities, it may lead to greater price fluctuations from one period to the next, for it introduces technological innovations and price oscillations in the source materials more abruptly. It should be noted that in case of network infrastructures, electricity or gas, due to the long life of those installations that will be on the ground for ever, this method can create an unacceptable regulatory risk for investors.

Finally, in privatisation processes, if the RAB is to be assessed from the *market value* of a privatised company, that market value must be equal to the value of the RAB to be calculated, multiplied by the rate of return established by the regulator for this type of utilities. This so-called "market value" is therefore the direct outcome of a regulatory decision. This may lead to a figure that differs substantially from the book value, giving rise to what is known as stranded assets or the opposite, a valuation above the book value. In any privatisation process, before selling companies, the regulator, subject to state-owner agreement, can establish sufficiently attractive rate schedules for the years ahead to command a selling price that satisfies state coffers. A balance must always be struck, however, between state and consumer interest, ensuring that the latter will not be over-burdened with high tariffs in exchange for large immediate gains for the public treasury.

Inclusion of new assets in the RAB

Additional assets are routinely required for three main reasons: (1) to deliver electricity to new customers or to meet growing demand, (2) to replace aged, deteriorated or obsolete facilities and (3) to improve service quality or comply with new legal or environmental requirements.

As explained earlier, price or revenue caps are applied *ex ante* to the years in the regulatory period to encourage efficiency in controllable OPEX and projected investment. If the company betters the OPEX target during the regulatory period, the extra cost reductions should clearly enhance its earnings. The allocation of the benefits of efficient investment, however, is less obvious and controversial.

Moreover, the subject is closely related to the mechanics of including new assets in the RAB.

When establishing a pattern for expected investment *ex ante*, the assets involved are included in the RAB, and the value of the allowed CAPEX is adjusted accordingly. This is the value that ultimately determines the allowed revenues for the next period (see exercise in Table 4.4).

The *ex ante* inclusion of investment in the RAB generates information asymmetries between the regulator and the company. Utilities normally have no lack of estimates, forecasts and projections on the amounts and characteristics of the infrastructure that will be needed to meet the needs created by expected market growth under different scenarios. Companies have an obvious incentive to overestimate requirements to raise their RAB and with it their allowed revenues. Company proposals in this regard must be critically re-assessed by the regulator, drawing from the necessary expertise, which should include investment cost benchmarking. The UK regulator has established a sliding scale mechanism to incentivise power distribution companies to make accurate projections for future investment. The aim is to avoid gaming and allow companies to retain part of the benefits gained from efficient investment [10, 17]. This mechanism is discussed in detail in Chap. 5, on the regulation of the distribution business.

Two situations may arise during the regulatory period: (1) the company invests less than initially estimated in the RAB on which its allowed revenues are based, consequently earning higher profits or (2) it invests more than planned, in which case it makes a loss.

In pure incentive-based regulation, in the absence of specific circumstances beyond the company's control that would justify deviations from the initial plans, such as unexpected demand growth, changes in legal requirements or similar, a pure *ex ante* approach is adopted. The regulated company keeps all or part of the earnings or losses (see for instance the profit and loss-sharing mechanisms implemented in the UK discussed earlier [10]). Further to Sect. 4.3, in incentive-based regulation, the strength of the incentive to reduce costs and enhance efficiency declines with the frequency of *ex post* cost reviews. Indeed, if such reviews are conducted yearly to include company investments in the RAB, the *de facto* result is cost-of-service regulation in which the incentive for efficient investment is very low.

The challenge is to be able to distinguish between intentionally postponed investment and genuine cost reductions. To prevent potential gaming in the form of postponing planned investment, regulators who are not in a position to monitor each new facility individually, such as in distribution grid companies, must monitor output variables, including quality of service indicators. The company's revenues should be lowered if the quality of supply does not improve as expected or even declines due to underinvestment.

When significant individualised investment, such as in transmission assets, for instance, is not undertaken within the expected time frame, some regulators have implemented an *ex post* review known as the *trigger approach*. The company is allowed extra revenues when the investment becomes operational or is penalised

when it is delayed with respect to when the asset in question was expected to come on stream [2].

As noted earlier, in practice most regulators adopt an *ex ante/ex post* approach. See Alexander and Harris [2] for other alternatives. Under this approach, the assets approved in the *ex post* review are the ones that replace the assets included in the RAB *ex ante*. The same two situations as described above are possible here: (1) the actual investment made by the company is smaller than initially included in the RAB, in which case the *ex post* review is straightforward, resulting in the inclusion of the actual investment in the RAB; or (2) the company invests more than initially forecast in the RAB, in which case the *ex post* review must include a detailed investigation into whether the investment decisions were adopted prudently. If that investigation deems some of the investments to be unnecessary, the respective assets may either not be included in the RAB or deemed to be eligible for depreciation expenses but not rate of return. The investments found to be necessary and efficient, by contrast, are included in the RAB for full recovery [2].

Another issue that must be addressed is *ex post* review timing. Three options can be considered.

- At the end of the regulatory period: assuming regulatory periods of 5 years, the incentive for efficient investment would be 5 years for facilities coming on stream in the first year of the regulatory period and 1 year for facilities commissioned in the final year.
- At the end of the next regulatory period: the incentive would range from 10 to 6 years.
- On a rolling basis: a 5-year incentive would be established for any investment irrespective of the year when it is actually made.

The longer the time lapsing between the inclusion of an asset in the RAB and the *ex post* review, the greater is companies' incentive to lower actual investment costs. Consequently, the second of the three options listed generates the most powerful incentive, followed by option 3. See Alexander and Harris [2] for numerical examples and a fuller discussion of other alternatives.

Depreciation

Assets may be depreciated by a number of methods [7], two of which are defined below.

- The annuity method: a flat annual charge is applied throughout the life of the facility to recover the capital invested plus the return on investment. The constant annuity consists of the repayment of the principal (amortisation or depreciation itself) on the one hand and the return on investment (interest or rate of return on capital) on the other. The amount of amortisation each year is computed, so that the sum of amortisation plus return on the remaining capital is constant.

Table 4.5 Depreciation methods: 10 % rate of return over 7 years [7]

	Present value	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
RAB and depreciation (annuity)								
Regulatory asset base		100.00	89.46	77.86	65.11	51.08	35.65	18.67
Depreciation		10.54	11.59	12.75	14.03	15.43	16.98	18.67
Return on assets (WACC*RAB)		10.00	8.95	7.79	6.51	5.11	3.56	1.87
Depreciation + return on assets	**100.00**	20.54	20.54	20.54	20.54	20.54	20.54	20.54
WACC	**10 %**							
RAB and depreciation (linear)								
Regulatory asset base		100.00	85.71	71.43	57.14	42.86	28.57	14.29
Depreciation		14.29	14.29	14.29	14.29	14.29	14.29	14.29
Return on assets (WACC*RAB)		10.00	8.57	7.14	5.71	4.29	2.86	1.43
Depreciation + return on assets	**100.00**	24.29	22.86	21.43	20.00	18.57	17.14	15.71
WACC	**10 %**							

Table 4.6 Accelerated depreciation: 10 % rate of return over 7 years

RAB and depreciation (accelerated)	Present value	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
Regulatory asset base		100,00	77,14	57,14	40,00	25,71	14,29	5,71
Depreciation		22,86	20,00	17,14	14,29	11,43	8,57	5,71
Return on assets (WACC*RAB)		10,00	7,71	5,71	4,00	2,57	1,43	0,57
Depreciation + return on assets	**100,00**	32,86	27,71	22,86	18,29	14,00	10,00	6,29
WACC	**10 %**							

- The straight line method: the yearly depreciation expense remains constant throughout the life of the facility, and therefore the total annual charge, including the return on investment, declines over time.

The two methods are illustrated in Table 4.5.

If correctly applied, the two methods yield the same present value of revenues to be received by the regulated company (see Table 4.5), although the impact on present and future customers differs. Straight line depreciation would require present customers to pay more than future customers, while under constant annuity arrangements the two groups of customers would be equally impacted.

Accelerated depreciation practices are also common. In accelerated depreciation either the expense is higher in the early years or the depreciation period is shortened. Like the other methods, accelerated depreciation would not affect the

present value of consumer repayments, although today's customers would pay more than tomorrow's (see Table 4.6).

Moreover, depreciation timing also affects the amount of corporation tax paid by the company, for as a deductible expense, it lowers the net operating income on which the tax liability is computed (see Sect. 4.4.5). With accelerated depreciation the company could pay less tax in the early years but more in the final years of the depreciation period. In such cases, the regulator should determine how to pass tax savings on to customers [7].

In principle, it is up to the regulator to establish both the service life of facilities for the intents and purposes of depreciation and the depreciation method (the straight line method is the most common).

4.4.3 Calculating the WACC

Since a company can use both debt and equity to finance its investments, its cost of capital is a weighted average of the interest rate on debt and the expected rate of return on equity. This average rate is known as the *weighted average cost of capital WACC*.

$$\text{WACC} = [\text{Debt}/(\text{Debt} + \text{Equity})] * R_{\text{debt}} + [\text{Equity}/(\text{Debt} + \text{Equity})] * R_{\text{equity}} \tag{4.10}$$

For example, if 40 % of a company's capital is debt and 60 % equity, and the interest paid on debt (R_{debt}) is 5 % and the cost of equity is 8 %, the WACC is 6.8 %.

As discussed above, the total rate of return set by the regulator during price or revenue revision processes is a crucial variable and has a direct bearing on revenues and expenses. The rate of return (WACC) determines both the average remuneration for the company's capital and it can also be used as the discount rate, $d = 1/(1 + \text{WACC})$, applied to find the present cost of projections when calculating allowed revenues throughout the regulatory period (see Table 4.4 in Sect. 4.4.1).

The interest rate on debt is usually lower than the rate of return on equity, since shareholders are more exposed to the financial failure of the company than the lenders. Then the above formula would appear to infer that raising the percentage of debt in the company's overall capitalisation would lower the WACC. Increasing debt above certain limits, however, also raises the likelihood of company bankruptcy and therefore the interest rate on debt, since the risk of default of the firm increases intolerably. Since risk rises with the debt/equity ratio, lenders will demand a higher rate of return.

The cost of equity can be estimated from securities market information on similar companies or industries. The company's cost of capital is thereby quantified as a whole, including different types of business and different levels of risk.

The most popular model used by regulators that quantifies the cost of capital on the basis of the risk associated with different lines of business is known as the capital asset pricing model, CAPM.

CAPM is based on the assumption that the rate of return for any activity is equal to:

- the rate of return on risk-free assets in the economy in question, regarded to be the amount received by investors placing their money in the safest financial assets, typically state bonds (this was correct before the financial crisis), computed as the average for the last few years for long-term rates, to establish a basis consistent with the life of the company's assets,
- plus a risk premium based on the degree to which the asset tracks the securities market; in other words, the specific additional risk associated with the asset over and above the average market risk (different methods are used to appraise debt and equity).

The risk premium applied to company equity is assumed to be proportional to a coefficient β, which represents the volatility of the value of the company's financial assets (shares) compared to average market volatility (see [21] for further details).

$$R_{\text{equity}} = \text{Rf} + \beta * (\text{Rm} - \text{Rf}) \tag{4.11}$$

where Rf is the risk-free rate of interest, Rm is the expected return on an efficient market portfolio.

Since these rates may be expressed in nominal or real (net of inflation) terms, the first step is to decide whether nominal or real WACC is to be used. The following expression indicates the relationship between the two.

$$(1 + \text{WACC}_{\text{nominal}}) = (1 + \text{WACC}_{\text{real}}) * (1 + \text{inflation rate}) \tag{4.12}$$

In countries with no international securities exchange or which lack sufficient liquidity for the type of industries regulated, or where the regulatory or financial risk is perceived to be high, the cost of equity is adjusted upward to include a country risk premium.

In conclusion, the utility's average cost of capital is calculated as follows:

$$\begin{aligned}\text{WACC} = {} & [\text{Equity}/(\text{Debt} + \text{Equity})] \cdot (\text{Rf} + \text{B} \cdot (\text{Rm} - \text{Rf}) + \text{Rc}) \\ & + [\text{Debt}/(\text{Debt} + \text{Equity})]\, R_{\text{debt}})\end{aligned} \tag{4.13}$$

where Rc is country risk, R_{debt} is the cost of debt calculated as the interest rate on corporate bonds, including the country premium.

Finally, as discussed below, in earnings and cash flow calculations the interest paid on debt is corporation tax deductible. Consequently, the WACC value must be defined as a before tax or after tax rate. Assuming t to be the tax rate per unit, the following relationships hold:

$$\mathrm{WACC}_{\text{after tax}} = (1 - t) * \mathrm{WACC}_{\text{before tax}} \tag{4.14}$$

$$\begin{aligned}\mathrm{WACC}_{\text{after tax}} = {} & [\mathrm{Debt}/(\mathrm{Debt} + \mathrm{Equity})] * R_{\text{debt}} * (1 - t) \\ & + [\mathrm{Equity}/(\mathrm{Debt} + \mathrm{Equity})] * R_{\text{equity after tax}}\end{aligned} \tag{4.15}$$

Note that, economically speaking, this means that the company pays interest at only $1-t$ of the before tax interest rate.

4.4.4 Operating Costs and Benchmarking

An estimate of the company's operating costs is needed to be able to compute the tariffs and establish the price or revenue formula for the next regulatory period.

The point of departure in this exercise is generally the audited accounts of costs incurred in the preceding period and a business plan furnished by the company with projections for all the years in the next regulatory period. The company must also break this information down as far as possible to show the different cost items by: activities (facility maintenance, delivery of supply to new users, repair of equipment and facilities and new infrastructure); categories (labour, materials, office material and expendables, energy consumption); type of consumer by service area (residential, commercial, industrial, street lighting).

The problem facing the regulator is how to define a feasible operating cost objective for the period. This target must be able to serve as an incentive for efficiency and the reduction of present costs, as well as to maintain company medium- and long-term sustainability; i.e. the level of efficient costs the company should strive to attain. Once the regulator somehow determines the operating cost objective for the regulatory period, where growth in demand or in number of customers have been already taken into consideration, the result can be expressed by applying a productivity factor, X, to these costs throughout the regulatory period:

$$\mathrm{OPEX}_t = \mathrm{OPEX}_{t-1} \times (1 + \mathrm{RPI}_t - X_{\mathrm{OPEX}}) \tag{4.16}$$

Where

- OPEX_t is the allowed operating costs in year t,
- RPI_t is the inflation rate per unit in year t,
- X_{OPEX} is the productivity factor for the allowed operating costs.

If Eq. (4.16) has not taken market growth (number of consumers or energy delivered) induced variations in operating costs into consideration, this can be replaced by an additional term in the equation.

Benchmarking techniques to compare relative efficiency among regulated companies in the same sector or to compare actual efficiencies against a reference company are used by regulators to calculate productivity factors for operating costs under the building blocks approach or for total costs under the TOTEX approach.

Box 4.5. Benchmarking Techniques

A company is more productive or more efficient than others if it requires fewer inputs to attain the same outputs, or if it produces more outputs with the same inputs. When the company is producing at its highest ideal productivity, it is said to be at its productivity frontier [4]. Reviews on the benchmarking techniques used by regulators to assess network monopoly efficiency can be found in [1, 8].

Figure 4.4 shows inputs and outputs for three companies, A, B and two situations for company C. Companies B and C are efficient, because they have already reached their productivity frontier (the rightmost curve). Company A, however, has room to enhance its efficiency, to "catch-up" with the others: its input is the same as B's but its output lower, while its output is the same as C's but its input higher. Companies that have already reached their productivity frontier can raise their long-term productivity by adopting newer technologies or innovative processes. This is what is known as the "frontier shift" (the leftmost curve). While productivity calculations should ideally take both effects into consideration, regulators often focus on benchmarking current best practice. The distance between a company's current productivity and its projection on the frontier is a measure of its inefficiency. The higher the inefficiency, the higher is productivity factor *X*.

Regulators have a number of different methods or benchmarking techniques from which to choose. These methods involve applying statistical techniques to compare the efficiency of different companies providing the same service in the same or similar countries: electric power distribution or transmission companies, for instance. Correlation analyses are run to compare individual cost items in the various markets. The results of such analyses can be used to define an average efficiency pattern or identify the most efficient companies (best practice) as models that others should emulate. Benchmarking requires a substantial amount of duly validated information and also serves as the basis for what has been referred to in preceding sections as yardstick competition. The most popular techniques are known as "*frontier methods*": data envelopment analysis (DEA), corrected ordinary least squares (COLS) and stochastic frontier analysis (SFA).

The various benchmarking techniques are shown in the flow chart in Fig. 4.5.

One traditional, simple and practical method for comparing efficiency among companies in the same industry is based on uni- and multi-dimensional ratios (a weighted combination of uni-dimensional

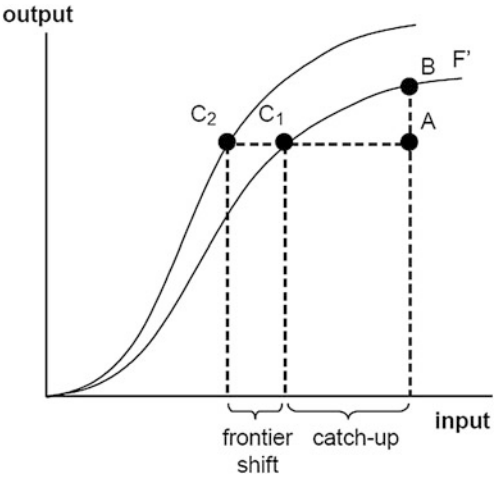


Fig. 4.4 Productive efficiency and productivity frontier *Source* [1]

ratios). It is used, for instance, to calculate costs per customer or unit of energy delivered or the number of customers served per employee.

The input–output relationship in the production process cannot be reflected in this family of methods. Total methods try to solve this problem. Index methods such as *total factor productivity* (TFP) generate a ratio based on the weighted sums of outputs and inputs. Frontier methods are an analytical method for finding the optimal weighting with which to combine outputs and inputs.

Data envelopment analysis (DEA) is a non-parametric technique that requires no functional relationship between outputs and inputs. Efficiency is defined as the ratio between the weighted sum of the outputs and the weighted sum of the inputs. Each company's efficiency factor is calculated by solving a linear optimisation problem where the weight factors and the efficiency factor are calculated for that company. The obtained efficiency factor lies between 0 and 1. See Jamasb and Pollit [8] for details on the formulation of linear programming problems. Companies with an efficiency factor of 1 determine the productivity frontier. For instance, in Fig. 4.6, four companies are compared when producing one output by using two inputs. The reported values are represented on the input1($X1$)/output(Y) and input2($X2$)/output(Y) graph. Note that companies 1 and 3 determine the productivity frontier (blue line), i.e. their efficiency factors are equal to 1. They require less input ($X1$ & $X2$) than companies 2 and 4 to produce the same output (Y). In this case, the efficiency factor can be obtained graphically. For instance, the company 2's efficiency factor is calculated as the distance OA divided by the distance O2. Note that

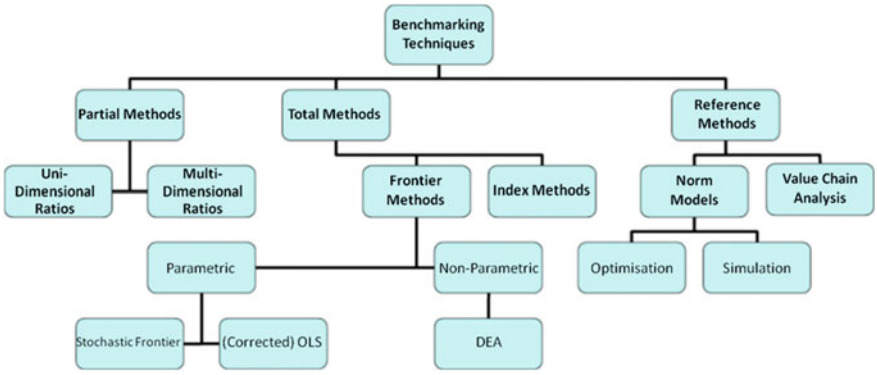


Fig. 4.5 Classification of benchmarking techniques (Source [1])

efficiency factors decrease with the distance to the productivity frontier.

Parametric methods require an understanding of production or cost functions. Figure 4.7 depicts production costs versus outputs or cost drivers (Y) for several companies. Each point represents one company. The average cost pattern is obtained by fitting a regression line, C_{OLS}, to the points using the *ordinary least squares (OLS) method*. The *corrected ordinary least squares method* (COLS) defines the efficiency frontier, C_{COLS}, in terms of the performance of the most efficient company(ies). For company B, the efficiency factor is calculated as distance EF divided by distance BF.

Stochastic frontier analysis (SFA) is similar to the COLS technique, except that it takes stochastic measurement errors into consideration when estimating the productivity frontier and efficiency factors. Each firm's distance to the frontier is explained here as the sum of a symmetrical error term (associated with relative efficiency) and a random error term to account for noise in the observations. Known probability functions for the distribution of those errors must be assumed [8].

Finally, when only a small number of companies is involved or when exogenous factors hinder inter-company comparisons, an ideal company may be constructed and taken as a reference. This is known as the *norm or reference model method* [13]. The ideal company is designed to provide the same service as the regulated companies, but at minimum cost. Due to the simplifications normally involved in modelling, the output from reference models cannot be used to directly determine the efficient cost levels that should be allowed. Adjustments must always be made to accommodate actual operating constraints that cannot be included in the model. Nonetheless, the model constitutes an objective on which the company should attempt to converge and which

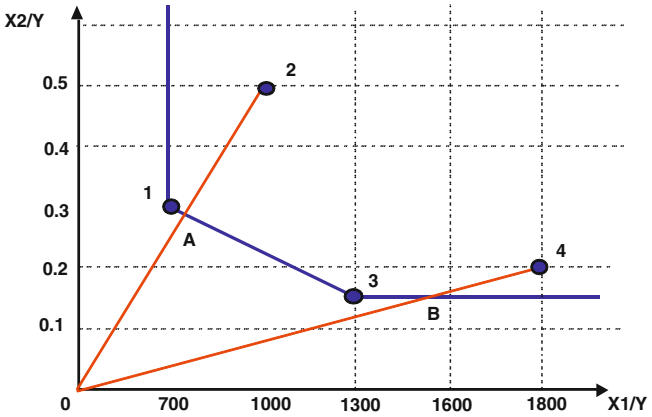


Fig. 4.6 Data envelopment analysis: two inputs, one output, four companies

may serve to compare the efficiency of a set of companies engaging in the same industry. Alternatively, it may be used to compare present and past costs for a given company based on past market growth. In a forward perspective, these models can also be used to calculate how far costs must rise from the reference year to the target year to meet expected market demand growth. Incremental costs, found by dividing the rise in costs by the increased demand met, can be then used to determine allowed revenues and design network tariffs, which for power distributors differ for each voltage level [20, 26].

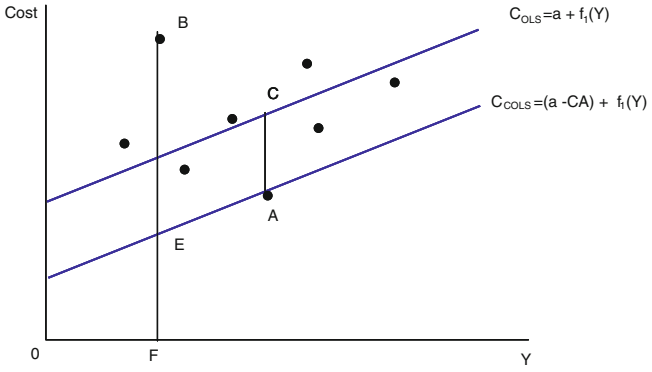


Fig. 4.7 Ordinary least squares (OLS) and corrected ordinary least squares (COLS) benchmarking methods

4.4.5 Tax, Cash Flow and Profitability

Utilities must pay a series of duties or fees and taxes levied on their earnings that affect their cash flow and business profitability. Regulators must take such expenses into account in the tariff revision process when calculating the rate of return and costs that the company is allowed to recover.

The charges to be paid by a utility are associated with the type of business conducted. Distributors must pay a series of local duties, including: municipal tax, property tax, chamber of commerce dues, business licence fees and other local charges. For instance, in Spain, such duties account for approximately 8 % of a power distribution company's gross receipts.

As tax rates cannot be controlled by the company, they should be taken into account when calculating its allowed revenues. Moreover, companies must pay the corporate tax, which for instance can amount to 30–35 % of the earnings before-tax. As it has been explained in Sect. 4.4.3 regulators take into account this tax when setting the allowed rate of return (Eqs. 4.14 and 4.15).

Box 4.6. Income Statement and Cash Flows

Both local charges and the tax on corporate profit are shown on the company's income statement, which may adopt the following form (Table 4.7).

The company's cash flow is calculated from the EBITDA for a given year by subtracting the interest and principal debt payments, the corporate tax and the investments (Table 4.8).

Table 4.7 Sample income statement

Allowed revenues
(−) Operating expenses
(−) Local and other non-corporate taxes
EBITDA (earnings before interest, corporate taxes and depreciation/amortisation)
(−) Depreciation
EBIT (Net operating income or earnings before interest and corporate taxes)
(−) Financial expenses (interest and other debt-related expenses)
EBT (Earnings before corporate taxes)
(−) Corporate tax
Net earnings

Table 4.8 Sample yearly cash flow calculation

Allowed revenues
(−) Operating expenses
(−) Local and other non-corporate taxes
EBITDA (earnings before interest, corporate taxes and depreciation/amortisation)
(−) Financial expenses (interest and other debt-related expenses)
(−) Debt amortisation (debt payment corresponding to the principal)
(−) Corporate tax
Cash flow
(−) Investment
Net cash flow

The foregoing leads to a series of conclusions on how duties and taxes affect utilities.

- Like operating expenses, duties (including local taxes and similar) must be paid out of allowed revenues. They differ from operating expenses; however, in that the company has no power to reduce or otherwise control them.
- Taxes have a direct effect on the company's earnings. Depreciation policy affects its tax liability. Depreciating more in the early stages, for instance, implies smaller earnings in those initial years and therefore lower corporation taxes, and consequently higher cash flows.
- Allowed rate of return calculations must be consistent with allowed company earnings calculations. For example, if the WACC defined by the regulator refers explicitly to before-tax income, to be comparable, the company's actual rate of return must be calculated as before tax earnings divided by the rate base. On the contrary, if the rate-of-return established by the regulator refers to the after-tax results, the actual rate of return must be calculated by dividing net earnings (i.e., net of taxes) by the rate base (see Sect. 4.4.3 for the relationship between before and after tax WACCs).

Finally, an example will illustrate how a regulated company's actual profitability for a given regulatory period is calculated after its revenues are set *ex ante* by the regulator. Table 4.9 gives the results for the regulated company described in Sect. 4.4.1 throughout the 5-year regulatory period, assuming performance to be exactly as predicted by the regulator when calculating the revenue-yield cap. In this example, as earlier, the values shown are net of inflation and company sales are flat throughout the period.

Table 4.9 shows that company investment and operational costs are as predicted (see Table 4.4), and its financial structure is divided into 60 % equity and 40 % debt. Note that the 8 % before tax WACC is computed based on that 60/40 structure, the 7 % rate of return on equity, the 3.8 % interest rate on debt and 35 % corporate tax. That before tax WACC is the rate allowed by the regulator to compute regulated revenues for each year (see Table 4.4). Net income is calculated as revenues minus operating expenditure, depreciation and interest payments.

Table 4.9 Actual rate of return of a regulated company under revenue-yield cap arrangements

Actual asset base and investment	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Opening AB		450.00	467.50	473.00	481.25	489.15
Gross capital expenditure—investment		**50.00**	**40.00**	**45.00**	**47.00**	**50.00**
Depreciation		32.50	34.50	36.75	39.10	41.60
Closing AB	**450.00**	467.50	473.00	481.25	489.15	497.55
Average AB		458.75	470.25	477.125	485.2	493.35
Actual depreciation	Life	Year 1	Year 2	Year 3	Year 4	Year 5
Existing assets (15 years)	**15**	30.00	30.00	30.00	30.00	30.00
New investment in year 1 (20 years)	**20**	2.50	2.50	2.50	2.50	2.50
New investment in year 2 (20 years)	**20**		2.00	2.00	2.00	2.00
New investment in year 3 (20 years)	**20**			2.25	2.25	2.25
New investment in year 4 (20 years)	**20**				2.35	2.35
New investment in year 5 (20 years)	**20**					2.50
Total depreciation		32.50	34.50	36.75	39.10	41.60
Financial structure & cost of capital		Year 1	Year 2	Year 3	Year 4	Year 5
Equity (%)		**60**	**60**	**60**	**60**	**60**
Debt (%)		40	40	40	40	40
Rate of return applied to equity (post-tax)(%)		**7.0**	**7.0**	**7.0**	**7.0**	**7.0**
Rate of return applied to equity (pre-tax)(%)		10.8	10.8	10.8	10.8	10.8
Rate of interest on debt (Rdebt)(%)		**3.8**	**3.8**	**3.8**	**3.8**	**3.8**
Tax rate on benefits		**35**	**35**	**35**	**35**	**35**
WACC (pre-tax)(%)		8.0	8.0	8.0	8.0	8.0
Actual operational costs (OPEX)		Year 1	Year 2	Year 3	Year 4	Year 5
Actual OPEX		**60.00**	**57.90**	**55.01**	**51.98**	**49.90**
Actual incomes set by the revenue yield cap		Year 1	Year 2	Year 3	Year 4	Year 5
Actual revenues		**131.27**	**130.55**	**129.84**	**129.12**	**128.41**
Balance sheet		Year 1	Year 2	Year 3	Year 4	Year 5
Incomes		131.27	130.55	129.84	129.12	128.41
Accounting costs		99.38	99.45	98.91	98.36	98.90
Operational costs (OPEX)		60.00	57.90	55.01	51.98	49.90
Depreciation		32.50	34.50	36.75	39.10	41.60
Debt payment (AB * Debt(%) * Rdebt)		6.88	7.05	7.16	7.28	7.40
Benefit before taxes		31.89	31.10	30.92	30.76	29.51
Taxes (Benefit * Tax rate)		11.16	10.88	10.82	10.77	10.33
Benefit after taxes		20.73	20.21	20.10	20.00	19.18
Actual remuneration of equity	Present value	Year 1	Year 2	Year 3	Year 4	Year 5
Actual equity (AB * Equity (%))	1.169.33	275.25	282.15	286.28	291.12	296.01
Actual benefit after taxes	82.29	20.73	20.21	20.10	20.00	19.18
Actual rate of return on equity after taxes (%)	7.0					
Make difference PV(equity)*ror − PV(benefit) = 0	0.00					

Lastly, the actual rate of return on equity is checked to verify that it concurs with the value allowed by the regulator, i.e. 7 %.

In the example, the actual rate of return, found as the present value of the after tax profit divided by the shareholders' equity, proves to be equal to 7 %, the value assumed for calculating the 8 % before tax WACC. Note that the present value is calculated by discounting at the unknown actual rate of return, using the Excel "goal seek" tool.

As an exercise, the reader may wish to calculate the increase in the actual rate of return assuming that the company reduces its operating costs to a constant value of 50 from the first year of the regulatory period.

4.4.6 Consumers or Companies Risks

A key issue in implementing revenue or price caps is to analyse how inflation variations, sales growth or recession would affect regulated revenues and company costs, and who, consumers or companies, would bear the risk of such deviations.

Under traditional cost-of-service regulation, the prices or tariffs set during the rate case process remain in effect until the next rate case is studied 1 or 2 years later. In this approach, the company's revenues obviously rise or fall in proportion to sales. If costs rise faster than sales, because inflation is high, for instance, or sales are lower than expected, the company shoulders all the risk and applies for a tariff revision. Conversely, when costs grow below expectations and/or sales are growing faster than initially predicted, the company benefits.

Viewed from the perspective of incentive-based regulation, *price cap regulation* is equivalent to cost-of-service regulation as far as the risks of variations in demand or sales are concerned, but with a heavier economic impact, because the regulatory period lasts for several years. Under price cap arrangements; however, the general effect of inflation on costs and revenues can be handled directly with the cap formula. In some cases, a weighted average of price indicators, i.e. a mix of general consumer price and specific industry indexes, is adopted. Although in general, inflation may not correspond to the actual evolution of the cost of the utility.

In *revenue cap regulation*, revenues can be made to grow by raising only the value of the market variables that have a direct effect on cost increases. One example would be a revenue cap formula for a distributor that takes account of larger revenues collected as a result of a rise in the number of consumers, but not as a result of higher demand for energy by the existing customers. As noted earlier, in Sect. 4.3.2, from a practical standpoint this type of revenue cap for distributors is consistent with the implementation of energy efficiency and savings programmes or the introduction of distributed generation.

Finally, the revenues allowed in revenue cap arrangements in general differ from the company's actual receipts from tariffs. Differences always arise between the expected sales taken as the grounds for the revision process and the company's

real sales during the period the revenue cap is in effect. Companies may have incentives to manipulate the sales estimates submitted to regulators:

- to persuade them of the need for higher investment in infrastructure (the company raises it sales growth predictions)
- to persuade them not to reduce prices or even to allow a price rise (the company lowers its sales growth predictions).

In revenue cap regulation revenues can be easily adjusted *ex post* to accommodate revenue deviations resulting from differences between *ex ante* predictions and the actual value of variables beyond the company's control, such as demand growth or inflation forecasts. In a nutshell, with revenue caps (and in the absence of *ex post* adjustments) consumers may bear all the risk associated with demand variations, while with price caps that risk is transferred to the company.

The *ex post* approach to acknowledging investments or otherwise for inclusion in the regulatory asset base discussed in Sect. 4.4.2 raises the risk assumed by the company, which is uncertain whether the investment cost will be acknowledged by the regulator. The *ex ante* approach, by contrast, affords companies greater certainty, transferring risk to consumers, who would pay or benefit for any deviations from actual infrastructure needs.

4.5 Quality of Service and Other Issues

As discussed in the foregoing, under cost-of-service regulation companies may have an incentive to over-invest to achieve, for instance, self-imposed quality of service levels primarily on the grounds of technical considerations or with any other excuse, simply because the capital invested earns a generous rate of return. Under incentive-based regulation the quality of service situation is exactly the opposite, however. Regulatory incentives to lower costs and enhance efficiency can lead to deterioration of the quality of service delivered. Quite obviously, both investment in infrastructure and operating costs for maintenance and repair in the event of failures have a direct impact on quality of supply.

In incentive-based regulation, performance parameters or indicators that the company must meet should be defined, along with penalties that reduce its revenues when it fails to do so. This is the alternative provided by incentive- or performance-based regulation to solve the difficult problem discussed earlier in connection with cost-of-service regulation. The solution consists not of monitoring each and every company investment to determine its technical and economic justification, but rather of monitoring the results delivered by the company in terms of both total cost and quality of supply.

Consequently, the regulator must establish a scheme to measure and monitor the company's performance indicators, i.e. the quality of service offered. Electricity distribution monitoring, for instance, might involve verifying factors such as

service restoration time after an outage, response time in meeting requests for new service connections or support for customers filing claims or complaints.

One light-handed regulatory tool to encourage companies to improve quality of service levels is public disclosure of their results. The advantage is that the regulator need not define quality targets or the economic implications associated with failure to meet them. Regulatory costs are therefore low. This type of regulation may be insufficient; however, if specific problems need to be solved or when public opinion fails to influence utility behaviour.

Other stronger mechanisms for regulating quality of service are based on penalties and incentives.

Penalties are established when the company fails to comply with the *individual quality standards* set by the regulator. Here, the company must pay a penalty to consumers for poor quality service, which may be measured in terms of the number and severity of supply outages in a year or failure to respond in a pre-established time to a request for a new connection to the grid, for instance. Such penalty schemes are based on individual user quality monitoring systems.

Furthermore, *integrated price quality regulation* has been implemented in several countries to regulate network infrastructures such as electricity distribution. A utility's yearly revenues may be increased or reduced by a certain percentage, for example, depending on the degree of compliance or non-compliance with the quality standards established for given areas or for the company as a whole. In practice, this entails inclusion of a Q-factor or a quality incentive term in the RPI formula. In this case, the regulator sets target values for specific *system level quality indices*, such as customer minutes lost or percentage of customers with a certain number of outages. The utility's measuring and monitoring systems must, moreover, be open to regulator audit.

4.6 Summary

This chapter discusses the fundamentals of and most common methods for regulating monopolies, with specific reference to their implementation in connection with network industries, in particular electricity distribution companies.

The essential ideas set out in this chapter are summarised below.

- Traditionally, the most common regulatory method is what is known as *cost-of-service* or *rate-of-return* regulation. From time to time, every year or two, the regulator analyses the company's costs, assets and investments and establishes the new revenue requirement for the following period. Such revenue requirement enables the company to cover its operating costs and asset depreciation expenses as well as to earn a rate of return, set by the regulator, on its rate base or net assets. The regulator usually also determines the tariffs for the different kind of consumers, from whose application the revenue requirement is recovered.

- Some drawbacks can be identified in cost-of-service regulation: (1) it provides no incentive for lowering company costs, since the regulator tends to acknowledge all costs incurred; (2) the rate of return is generally set high enough to constitute an incentive for companies to invest more than is economically optimal and (3) information asymmetries between the regulator and the company are more difficult to manage adequately.
- To mitigate such problems, *incentive-based regulation* methods are becoming more and more popular applied to regulate network industries. The chief characteristic of this type of regulation is that the tariffs or revenues the company is authorised to charge or receive are kept in place for longer intervals, typically 4 or 5 years. This provides an incentive for the company to lower its costs and be more efficient than under cost-of-service regulation. The resulting efficiency improvements are considered by the regulator in the following price control, and therefore the consumers also benefit. The two most commonly used methods are *price caps* and *revenue caps*.
- These methods set prices or revenues that the company may charge or receive throughout the regulatory period with a formula with a yearly adjustment factor known as *(RPI-X)*.
- The chief difference between price and revenue caps is that, under price caps, any increase in sales leads to higher revenues; i.e., costs are assumed to increase proportionally with sales. Under revenue caps, by contrast, receipts do not rise in direct proportion to sales, but only in keeping with the selected cost drivers and in the proportion established. Price and revenue caps are frequently applied to regulate transmission or distribution network companies, which are characterised by high fixed investments and whose costs do not normally depend on the intensity of use of the resulting assets.
- Incentive-based regulation that induces companies to lower costs in operation or investment may lead them to do so at the expense of *service quality*. For this reason, such regulatory mechanisms go hand in hand with what is known as performance-based regulation, which not only regulates revenues but sets objectives to be met by the company. The most important of these objectives are the quality standards that the company must meet. Regulators may establish *system level quality targets* to verify average company performance or *individual standards* to guarantee a minimum quality to each consumer. In both cases, when the company fails to comply with the established standards, it is financially penalised by a reduction in its allowed revenues.
- In each *rate case* or *price control* procedure, regulators must project future efficient costs with which to align allowed revenues. Some of the key elements of this process are:
  - Calculation of efficient operating costs: assessing company cost-efficiency on the grounds of reported costs with *benchmarking techniques* and/or *norm reference models*.

  - Inclusion of assets in the *regulatory asset base* (RAB): updating the RAB with yearly variations to accommodate asset depreciation and projected allowed investment.
  - Establishment of an allowed rate of return evaluated on the basis of the *WACC*, which takes the cost of both company debt and company equity into consideration.
  - Acknowledgement of other *non-controllable costs* such as charges, duties and taxes that should also be included when calculating allowed revenues.
  - Calculation of the X factor, as the factor that equates the net present value of allowed costs, including return on capital, to the net present value of allowed revenues during the regulatory period.
- Incentive-based regulation, because of its emphasis on efficiency, will fail in promoting innovation activities, which are risky and typically need times for maturity that are longer than the price control periods in incentive-based regulation. Addressing this issue is a current open area of research in monopolies regulation.

References

1. Ajodhia VS (2005) Regulating beyond price: integrated price-quality regulation for electricity distribution networks, PhD Thesis, Delft University of Technology
2. Alexander I, Harris C (2005) The regulation of investment in utilities: concepts and applications. The World Bank Working Paper No. 52, Washington
3. Averch H, Johnson L (1962) Behavior of the firm under regulatory constraint. Am Econ Rev 52(5):1052–1069
4. Coelli T, Prasada Rao DS, Battese GE (1998) An introduction to efficiency and productivity analysis. Kluwer Academic Publishers, Boston
5. Comnes A, Stoft S, Greene N, Hill L (1995) Performance-based ratemaking for electric utilities: a review of plans and analysis of economic and resource-planning issues, vol 1. Lawrence Berkeley National Laboratory, Berkeley (Nov): LBNL-37577
6. DOE/FERC (1973) Uniform systems of accounts prescribed for public utilities and licenses classes A, B, C and D. Federal Energy Regulatory Commision. D.O.E./FERC-0028, Washington
7. Green R, Rodriguez-Pardina M (1999) Resetting price controls for privatized utilities. a manual for regulators. The World Bank, Washington, Feb 1999
8. Jamasb T, Pollitt M (2003) International benchmarking and regulation: an application to European electricity distribution utilities. Energy Policy 31:1609–1622
9. Joskow P, Schmalensee R (1986) Incentive regulation for electric utilities. Yale J Regul 4(1):1–49
10. Joskow PL (2006) Incentive regulation in theory and practice: electricity distribution and transmission networks. Cambridge working papers in economics CWPE0607, electricity policy research group working paper EPRG 0511
11. Kahn AE (1988) The economics of regulation: principles and institutions. MIT Press, Cambridge

12. Laffont JJ, Tirole J (1993) A theory of incentives in procurement and regulation. MIT Press, Cambridge
13. Mateo Domingo C, Gómez San Roman T, Sánchez-Miralles A, Peco González JP, Candela Martínez A (2011) A reference network model for large-scale distribution planning with automatic street map generation. IEEE Trans Power Syst 26(1): 190–197
14. National Action Plan for Energy Efficiency (2007) Aligning utility incentives with investment in energy efficiency. Prepared by Val R. Jensen, ICF International. www.epa.gov/eeactionplan
15. National Audit Office (NAO) (2002) Pipes and wires. http://www.nao.gov.uk
16. Navarro P (1996) Seven basic rules for the PBR regulator. Electr J 9(3):24–30
17. OFGEM (2004) Electricity distribution price control review. Final proposals, office of gas and electricity markets. www.ofgem.gov.uk
18. Petrov K, Nunes N (2009) Analysis of insufficient regulatory incentives for investments into electric networks. An update. Final report. KEMA consulting GmbH submitted to European Copper Institute. http://www.leonardo-energy.org/
19. Posner RA (1999) Natural monopoly and its regulation, 30th Anniversary edn. CATO Institute, Washington
20. Roman J, Gómez T, Muñoz A, Peco J (1999) Regulation of distribution network business. IEEE Trans Power Deliv 14:662–669
21. Rothwell G, Gómez T (2003) Electricity economics: regulation and deregulation. IEEE-Wiley Press, Piscataway
22. Rudnick H, Raineri R (1997) Chilean distribution tariffs: incentive regulation, in (De) Regulation and Competition: the Electric Industry in Chile, Ilades-Georgetown University, pp 223–257
23. Shleifer A (1985) A theory of yardstick competition. Rand J Econ 16:314–327
24. Suelflow JE (1973) Public utility accounting: theory and application. Michigan State University—Public Utilities Studies, USA
25. The Regulatory Assistance Project (RAP) (2000) Performance-based regulation for distribution utilities. www.raponline.org/Pubs/General
26. Turvey R (2006) On network efficiency comparisons: electricity distribution. Util Policy 14:103–113
27. Viscusi WK, Harrington JE, Vernon JM (2005) Economics of regulation and antitrust, 4th edn. MIT Press, Cambridge

Chapter 5
Electricity Distribution

Tomás Gómez

A friend of mine who worked in a distribution company likened electric power generation and transmission to a bull and distribution to a bee hive. Whereas generation and transmission comprise comparatively few and very large-scale facilities, distribution involves a much larger number and wider variety of equipment and components. Both are critical to keeping the lights on.

The above difference may be what essentially distinguishes electricity transmission from distribution, the two network-infrastructure-based natural monopolies. Transmission facilities can be evaluated one by one for suitability; by contrast, any regulatory scheme for distribution must acknowledge from the outset that each and every investment made by a distributor cannot possibly be controlled and that more aggregated or global evaluation mechanisms are required.

This chapter provides a detailed analysis of the various aspects involved in the regulation of electricity distribution. As natural monopolies, i.e., the sole providers of a service in their area of influence, distributors must be regulated, essentially in two respects: service price and service quality.

When utilities are regulated as vertically integrated monopolies that generate, transmit, distribute, and bill power to consumers located in their franchise area, these activities are not separated nor, often, are the cost items associated with the different types of activity clearly distinguished. Therefore, the rate charged has to cover all the utility's costs as a whole. After restructuring and liberalisation has occurred in many countries, each activity has to be independently regulated, since separate companies now own the different parts of the business. In most countries there is usually only one national transmission grid, with, as a rule, one or at best a handful of owners, while a rather large number (two to ten, up to a few hundred) of operators distribute electrical power in different areas of the country. Since under the new circumstances separate accounts must be kept for the various transmission and distribution companies, regulatory schemes must design rates for all of these businesses on the basis of the specific conditions prevailing in each. In this

T. Gómez (✉)
Instituto de Investigación Tecnológica, Universidad Pontificia Comillas,
Alberto Aguilera, 23 28015 Madrid, Spain
e-mail: tomas.gomez@upcomillas.es

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_5,

environment, distribution must be a sustainable business in and of itself over the long term, with an efficient and sufficient remuneration scheme.

In many power systems the network distribution business continues to be closely tied to energy retail in the case of regulated-rate customers. These customers buy energy from their distributors, which in this case play a dual role: carrying power in their grids and selling energy to these customers. In this situation, companies receive an annual remuneration for the two activities. Generally speaking, distributors should be remunerated by their incurred costs, including a fair rate of return on capital, for supplying electricity to regulated customers. In this regard, distributors/retailers are normally authorised to charge rates that cover the full costs incurred in the wholesale purchase of the power consumed by their regulated customers. By contrast, in countries or regions such as the European Union [9] where the distribution and retail activities have been unbundled and electricity generation and retail have been deregulated and are presently governed by the rules of free competition, distributors are not allowed to conduct retail activities or may only do so temporarily or under special circumstances, for small commercial and residential customers. This chapter focuses on distribution regulation related to the strict grid activities made by distribution companies.

As it was analysed in the preceding chapter when dealing with regulated monopolies, regulating distribution primarily entails defining a scheme to link companies' yearly revenues to their operating costs, plus depreciation and a return on the capital invested. In electricity distribution as in transmission, cost-of-service regulation is giving way to incentive-(price or revenue cap) based regulation. Nonetheless, certain specific issues are dealt with differently in distribution than in transmission.

- In transmission, each new infrastructure is analysed individually and its suitability and inclusion in the rate base is determined accordingly. In distribution, infrastructure growth must be evaluated as a whole and linked to the growth in demand and reliability requirements by area and voltage level.
- In transmission, the price signals associated with line use may be singled out at the node or grid substation level, and congestion is analysed for each of the chief sections comprising the network. Node or area prices can be determined, together with the cost of congestion in each corridor. In distribution, by contrast, since the grid has an aggregate spatial layout separated only by voltage levels, average combined costs are determined by voltage level and consumers and lines can only be distinguished from one another by the voltage at which they connect to the grid.
- Quality of service is regulated differently in transmission and distribution: in transmission, unavailability is recorded for each individual facility, whereas in distribution, aggregate unavailability metrics in the service to end consumers are found in each zone into which the distributor's service area is divided.
- Finally, another difference between transmission and distribution, no longer valid, was that generation was basically connected to transmission networks, while there was very little generation in distribution. As discussed below, the

deployment of distributed generation based on renewables or combined heat and power production is drastically changing this paradigm.

In addition to set yearly regulated revenues, regulation must establish objective, non-discriminatory rules to ensure the principle of network access for energy transactions among the various players, while maintaining grid operator, i.e., distributor, neutrality.

Other activities traditionally performed by distributors, such as electricity meter reading, billing, handling new service connections, demand-side management and energy-saving programmes are also being reviewed in the industry's new organisational structure. Coordination between distributors and suppliers and the entry of market newcomers authorised to conduct such activities are still open issues pending a regulatory consensus in the competitive framework.

This chapter addresses the various factors included in the regulatory design for distribution companies. Following this introduction, its contents are organised as follows.

The Sect. 5.1 describes the physical structure of distribution networks by voltage level in downstream hierarchical order, from the substations that connect into the transmission system to end customer service connections. This section also examines the functions performed by distributors (grid planning, development, operation and control activities) in fulfilling their chief mission, which is to supply reliable electric power to present and future customers.

The Sect. 5.2 analyses the crucial aspects of distribution regulation. First, the requirements to be met to obtain a licence for a distribution franchise are listed; second, the ways to guarantee free access to the distribution grid by the different types of consumers, small generators or other grids are described; lastly, a section is devoted to the principles to which grid rates should be designed and the typical form that such rates should adopt. This latter issue is addressed at greater length in a further chapter which deals with tariff design.

The Sect. 5.3 discusses the most common price and revenue cap schemes that have been implemented to calculate the annual remuneration earned by distributors in a few systems where distribution has been unbundled from the other electricity industry activities and is independently regulated. The schemes used in the United Kingdom, Norway, California and Spain are described in some detail.

The Sect. 5.4 defines the three areas that comprise electricity service quality. Particular attention is devoted to the technical aspects: continuity of supply (power interruptions) and voltage quality (voltage disturbances or power quality). The theoretical principles underlying each of these problems are considered, and mechanisms are proposed for their regulation and control by distributors.

The Sect. 5.5 addresses the problem of how to encourage distributors to hold energy losses in the grid to an economically efficient level. For energy losses, as for service quality regulation, the regulator establishes targets to be met by the company, which is penalised or rewarded accordingly.

The Sect. 5.6 analyses the impact of the connection of distributed generation on the distributor's functions and revenues. Distributed generation, i.e., small-scale

generators connected directly to the distribution grid, has recently experienced substantial development. This upward trend is associated with the promotion of renewable energy and the greater energy efficiency achieved when consumers generate their own combined heat and power (CHP or co-generation).

Finally, the Sect. 5.8 introduces some relevant issues regarding rural electrification in developing countries, where the provision of universal electricity service should be one of the main aims pursued by the respective governments.

5.1 Distribution Grids and Distributor Functions

The role of electricity distribution consists of carrying energy from the connection points on the typically nationwide transmission grid to end consumers over a network known as the *distribution grid*, which is regional and local in scope. Small-scale (generally under 50 MW) generators may also connect to the distribution grid. Such generators are typically known as *distributed generation*. Moreover, frequently the distribution company's grids are connected to the grids of neighbouring companies or embedded in them.

5.1.1 Network Structure

Typically, the distribution grid is hierarchically structured into functional areas by voltage levels. These levels, in downstream order from the transmission grid to the end consumer, are listed below.

- *High voltage* (HV) distribution grids have nominal voltages of up to about 132 kV (kilovolts). In Europe, they typically carry electricity at voltage levels between 33 and 132 kV. In the US, the most commonly used voltage levels are 26, 69 and 138 kV. This is the grid that connects distribution to the transmission substations and also supplies large industrial customers requiring several megawatts (MW). This network is also called the *sub-transmission* grid.
- *High/medium-voltage (HV/MV) substations* are the points on the HV grid that feed large geographic or densely populated urban areas. MV feeders are supplied by these substations.
- The *medium voltage (MV) grid* comprises the main feeders and branch lines. In Europe, these grids typically carry electricity at voltages between 10 and 30 kV. In the US, the most widely used voltage levels are 4 and 13 kV. This grid feeds distribution transformers and end consumers (industrial, commercial and office buildings) requiring tens or hundreds of kilowatts (kW). This network is also known as the *primary distribution grid*.
- *Medium/low-voltage (MV/LV) transformers* connect into the MV grid, where they feed groups of end consumers in a relatively small area, such as a town or

city neighbourhood, for instance. Power is supplied to the LV lines to end customers' service connections by these *distribution transformers.*

- The *low voltage (LV) grid* consists of the radial lines from the MV/LV transformer to small commercial and residential end customers requiring several kW. In Europe, the nominal voltage in this grid is 400 V (volts) from phase to phase and 230 V from phase to neutral. In the US, the supply voltages are 420 and 240 V, which allows a 120–110 V supply for end consumers. This network is also known as the *secondary distribution grid.*

This hierarchy is illustrated in Fig. 5.1.

The *high voltage or sub-transmission grid*, as stated above, is connected to the nationwide transmission grid, with distribution substations to carry power to large cities or large-scale consumers. This high voltage distribution grid is regional in scope, i.e., it generally extends across several states or provinces, covering areas on the order of tens of thousands of square kilometres. The grid is usually laid out in a ring or loop configuration to enhance the reliability of supply at the demand points it feeds, namely the HV/MV substations. With such a meshed, loop or ring configuration, if one of the lines comprising the network fails due to an outage, supply is not interrupted at the load or demand points, as it can be automatically fed over an alternate path unaffected by the outage. Then it is said that the network meets the *N-1 reliability criterion.* Pure radial high voltage structures that do not meet the N-1 criterion are only used in distribution grids in rural areas.

The *MV distribution grid* structure is different in urban and rural areas. *Urban MV grids*, as their name implies, service cities and are usually run underground, with insulated electrical cables. For reasons of reliability of supply, such grids

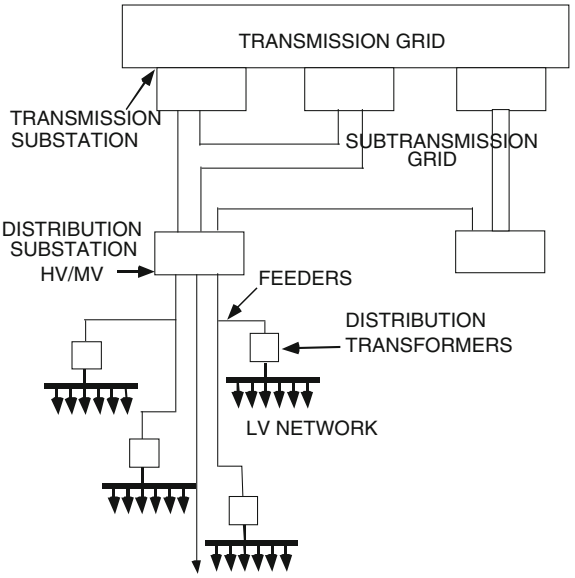


Fig. 5.1 Distribution grid: general hierarchical structure

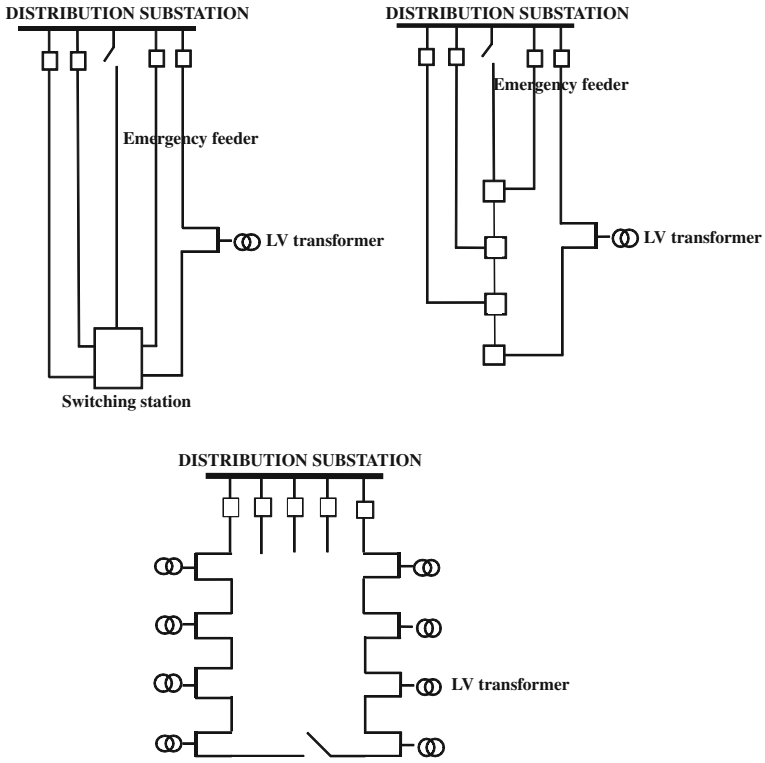


Fig. 5.2 Urban MV grid structures

have a loop structure, although they are operated radially. In the event of failure in a given cable, this loop structure ensures that the demand powered by that cable is supplied by another cable, the emergency or back-up feeder, for instance, over an alternative path. Three urban grid structures are shown in Fig. 5.2.

Rural MV grids normally consist of bare overhead conductors with a pure radial configuration. Here the supply reliability requirements are not as strict; as a rule, when a failure occurs, all the service downstream of the failure is interrupted for as long as it takes to repair the outage. Figure 5.3 shows a standard rural MV grid.

LV grids run from distribution transformers to the end consumers, typically residential and commercial customers. Larger scale industrial consumers connect directly into the MV or HV grids. The LV grid has a pure radial configuration, and each LV line supplies several consumers. The point of entry to one or several consumer premises is called the *point of connection.*

In distribution, because of the large number of facilities and equipment involved and the variety of component suppliers, the company must *standardise* facility and equipment types and coordinate purchases with its suppliers.

By way of illustration, the Tables 5.1 and 5.2 shows the size of the main companies and the number of distribution facilities in Spain.

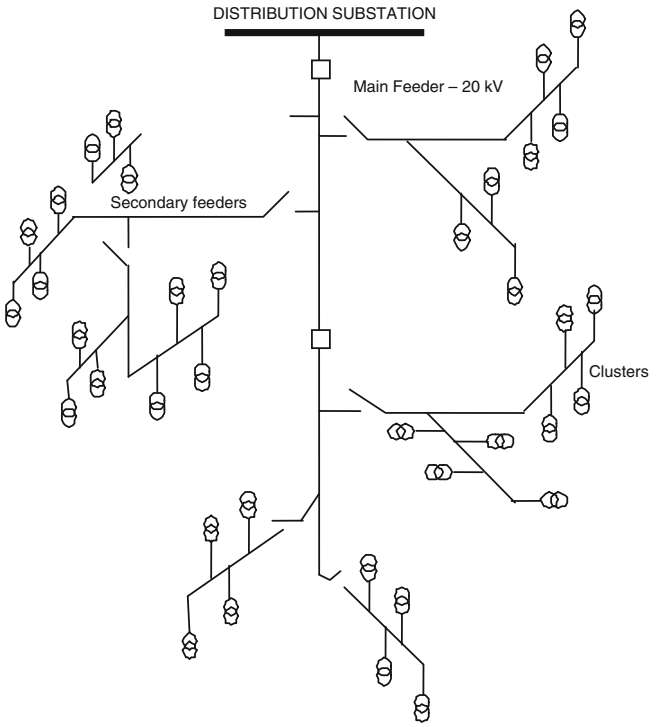


Fig. 5.3 Rural MV grid structure

Table 5.1 Distribution companies in Spain

Distribution company	Number of customers (2006)	Distributed energy (GWh) (2006)
Iberdrola Distribución Eléctrica	9,021,217	92,631
Endesa Distribución Elléctrica	9,749,480	90,177
Unión Eléctrica Fenosa Distribución	3,273,717	33,724
Cantábrico Distribución	531,748	9,536
Electra de Viesgo Distribución	536,372	5,393

Table 5.2 Distribution facilities in Spain

Distribution facilities in Spain (2004)	Total
HV lines (36 kV $<$ V $<$ 220 kV)	60,396 km
HV/MV substations (installed capacity)	90,840 MVA
MV lines (0.38 kV $<$ V $<$ 36 kV)	219,167 km
MV/LV substations (installed capacity)	49,866 MVA
LV lines	281,678 km

Compiled by the author from Spanish National Energy Commission (CNE) data

5.1.2 Distributor Functions

In addition to customer management and power sales to regulated consumers, distribution companies perform a series of grid-related technical functions falling under the following general headings:

- grid planning,
- works development, and construction
- facility and equipment operation and maintenance.

Grid *planning* begins with an estimation of the demand growth to be serviced by the distributor in the future. Existing demand, the horizon considered, natural demand growth, residential and industrial urban development plans, the impact of energy savings and efficiency plans and the possible connection of distributed generation must all be taken into account. Planning must be comprehensive and hierarchical for all of the distributor's grids in the areas served. In other words, plans should first be formulated for the HV grids, followed by MV and finally LV grid design after the location of new service connections or enlargement of the existing ones is clearly defined. The long service life of distribution facilities, typically from 30 to 40 years, must also be taken into consideration when projecting the horizon for future grid architecture. Facility planning envisages wide load margins: several positions are left vacant in substations, for instance, for possible future use. The reliability criteria to be met in covering demand loads have a decisive influence on planning, specifically in questions such as when a meshed structure is more appropriate than a dual feed design, or when emergency mains or redundant feeders should be considered for outages.

Grid facility *development and construction* entail good project and works management and expeditious handling of the respective formalities. For instance, laying cables and building substations in urban areas involve complex logistics, as the impact of open ditches on everyday life must be minimised.

Finally, grid *operation and maintenance* are in turn broken down into grid studies, grid operation and control, and predictive, preventive and corrective maintenance.

Distribution studies analyse operating results, normal operating conditions and emergency situations, together with quality of supply plans and reviews.

Operation and control are essentially performed at dispatching centres, using distribution management systems (DMS) and supervisory control and data acquisition (SCADA) systems. The grid can be reconfigured as required in the event of outages, while voltage profiles are monitored and substation and feeder load levels are controlled from these dispatching centres. This is also where maintenance tasks are programmed, adopting the safety measures required to work on facilities, whether under live line or disconnected conditions. Automated remote control and remote measurement are far less developed than in the transmission grid. Generally speaking, remote measurement and remote control are automated on the grid upstream of and including the MV feeder connections at

HV/MV substations. The number of distribution transformers and customers handled by companies has been considered too large to make individual monitoring economically efficient. However, this traditional view is currently changing under the new paradigm of *smart grids* and *smart metres*. In this new context, network automation and monitoring will be extended up to the end customers, promoting demand response, energy efficiency and the integration of distributed generation.

With regard to *maintenance*, a distinction is drawn between *preventive and predictive* tasks on the one hand and *corrective* maintenance to repair outages on the other. Distributors organise and systematise maintenance by dividing the area serviced into different geographic districts covered by respective working crews. Preventive and predictive maintenance is intended to reduce the frequency of failures and, therefore, supply interruptions. Speedy grid reconfiguration and outage repair actions, in turn, are designed to shorten the time consumer supply is interrupted. Coordination between the dispatching centre and the maintenance crews acting in areas affected by outages is instrumental to prompt location of failures and restoration of supply. Higher levels of network monitoring and control, as proposed by the *smart grid* concept, would also result in greater supply reliability by facilitating and speeding up maintenance tasks [23].

5.2 Distribution Regulation

As a natural monopoly, distribution must be regulated, as noted above. This regulation, in turn, must guarantee the distributor sufficient revenue by striking a balance between profits to ensure economic viability on the one hand and low rates for service users on the other. The distributor's costs may be classified under the following main items, broken down by distribution function:

- investment to strengthen the existing grid and build new facilities,
- grid facility operation and maintenance costs,
- cost associated with meter reading and billing for network services,
- cost of energy losses involved in transmitting and distributing power over the grid.[1]

The first issue regarding economic regulation is to assess the "efficient" costs for which distributors should be remunerated. Regulator-authorised *revenues* are collected by the distributor from network users in the form of the rates charged for the service. *Network charges* are generally separated into two different items:

[1] As explained in Sect. 2.1.2.5, network losses happen in electricity networks, but they are not network costs, since the cost of producing the electricity that is transformed in heat losses in the distribution networks took place in some generation plants. However, a sound regulation must have incentives for the distributor to try to reduce network losses to an economically justified level.

connection charges, which consist of a single payment made by the user when a new connection is required or when connecting grid facilities need to be upgraded because demand rises, and *use-of-system charges*, periodic payments made by network users to cover the total cost of the regulated service. It is especially difficult in distribution grids, given the wide variety of grids and large number of consumers, to ensure that the rate paid by each user reflects the actual costs incurred by him/her while also providing for fair discrimination in terms of space and time.

A second issue in distribution regulation is related to the particular characteristics of distribution networks as opposed to transmission, as it was explained in the introduction of this chapter. In general in a country there are several distributors, each providing service in a different area, and very likely with different types of networks and costs. Unlike the transmission grid, distribution networks have little or no effect on the wholesale power market; therefore, interactions between market and network regulation are different in nature. Lastly, distribution is of key importance in the quality of supply received by customers: the origin of approximately 90 % of the outages affecting end consumers can be traced to the distribution grid. Therefore, quality regulation is a key issue in distribution.

Finally, one last general consideration should be mentioned with respect to the regulation of distribution: it must provide for sufficient regulatory stability for the electricity distribution business to be perceived by the company as a low-risk activity. This leads to lower rates of return and greater overall economic efficiency in the long term. The foregoing does not necessarily mean that distribution should be a low-margin business, but simply that this margin should not be widened by risks associated with regulatory uncertainty.

5.2.1 Distribution Licences

Monopolistic businesses are normally conducted under State authorisation in the form of a licence or franchise stipulating the conditions to be met by the distributor in providing the public service. Unbundling the legal and ownership structure, or at least the accounts, of regulated (network and regulated retail to captive customers) and non-regulated (retail) activities is a requisite in many regulatory schemes. Some of the types of provisions included in distribution licences are listed below:

- Duration and conditions for renewal or loss of the licence enabling a company to engage in distribution.
- Geographic delimitation of the area serviced by the distributor (franchise or concession area).
- Obligation to supply all consumers, generators or other networks requesting grid connections, while maintaining the established service quality standards.
- Rules capping the revenues earned by the company during the regulatory period.

- Conditions under which the various players can access the grid for power purchase and sale transactions.
- Connection and access charges and other rates.

The justification for the distributor's obligation to supply electricity in its service area lies in the essential nature of the service in question. This obligation refers both to current consumers and future new connections. In exchange, the company charges regulated rates as well as connection costs, which must be regulated separately. For regulated consumers, the obligation refers both to the grid connection and the supply of power, therefore affecting the company's grid as well as its supply business.

In countries with non-electrified rural areas, the cost of supplying power to consumers in such areas is far larger than the average supply cost. Where an obligation to provide this service is in place, then, it must be accompanied by additional payments over the standard distribution charges for already connected customers that enable the company to recover the cost incurred. Many countries, such as Chile, Argentina and Peru among others, have acquired experience with publicly supported mechanisms in which private initiative is responsible of the provision of the service. Permanent dependence on subsidies should be avoided as much as possible. One way to achieve this is by granting them on a one-off basis after the initial infrastructure investment is made. This issue will be discussed later in the section on rural electrification.

5.2.2 Network Access

Another key issue in the regulation of network business, and specifically of electricity distribution where retail liberalisation has happened, is to guarantee *third-party access (TPA)* or *open access* for power sales and purchases with objective, transparent and non-discriminatory rules to prevent distributors from abusing their monopolistic power to impede retail competition. In return, as has been stated, all network users must pay a charge or price for the service provided.

The grid access scenarios that may arise in distribution include:

- Consumers directly connected to the distribution grid, receiving power either directly from the distributor, which acts as supplier, or from another supplier.
- Generators directly connected to the distribution grid and selling the power they produce to the distributor or some other buyer.
- Power transactions with other distributors in which the grid is used, or with eligible consumers or generators connected directly to the distribution grid.

The basic principles of third-party access are shown below.

- It is a right to which all market players are entitled.
- The costs incurred must be shared by all grid users.

- The right of access exists regardless of the identity of the specific organisation supplying electric power.

The regulator must resolve any access conflicts that arise. For instance, distributors must be prevented from abusing their dominant position to deny access to suppliers competing with a supplier in the same business group as the distributor. Another potential area of dispute is the competition between two distributors to supply a given consumer when the territorial limits of their respective franchises are not clearly delimited.

The charge to be paid for a new service connection, in turn, may also cause problems relating to design or to the rules for applying the charge. For instance, regulations must determine whether each connection should pay an individualised charge or whether a general flat rate should be established and surcharges applied only to cover exceptional circumstances.

Finally, another issue often debated in the context of third-party access is whether the actual construction of connections should be part of the monopoly franchise or whether distributors' involvement should be limited to establishing minimum design and operation requirements. In the latter case, connections would be built by other companies under competitive tendering conditions. This approach is more common for building transmission facilities than for distribution installations.

5.2.3 Network Charges

As discussed in the chapter on monopoly regulation, the monopolistic company, in this case the distributor, is allowed by the regulator to earn certain revenues during the regulatory period in question. The company receives this remuneration from the rates charged to end consumers. From the standpoint of distribution, costs are recovered under what are known as *distribution charges*. These charges are either a component of the *integral tariff* paid by regulated consumers, i.e., those who still buy their electric power at regulated rates, or form a part of the *access tariff* paid by non-regulated consumers for network services. Regulated charges should, as much as possible:

- Reflect the cost incurred to provide the consumer with the service.
- Ensure full recovery of the distributor's total acknowledged costs.

Regulated distribution charges are typically structured as follows.

- The *connection charge* is a one-off charge, paid in Euros or whatever currency, for a new grid connection or an extension of the existing grid.
- The *use-of-system charge* is typically paid periodically (monthly or bi-monthly) and, in general, it may include a component proportional to the energy demand (energy component in Euros/kWh) and another somehow proportional to the load contribution to the peak demand, or to the contracted demand, when this

Table 5.3 High voltage access rates in Spain

Voltage	Capacity component (Euros/kW-year)	Energy component (Euro cents/kWh)
MV (1 kV < U < 36 kV)	16.3 – 2.7	7.0 – 0.6
HV1 (36 kV < U < 72.5 kV)	14.0 – 2.3	2.3 – 0.2
HV2 (72.5 kV < U < 145 kV)	13.2 – 2.2	1.9 – 0.2
Transmission (145 kV < U)	9.9 – 1.6	1.0 – 0.1

Source [5]. The range of values for each component of the access rate reflects seasonal and time of use differences, with the year divided into six periods, from one (highest rate) to six (lowest rate)

exists (capacity component in Euros/kW-month). This use-of-system charge is intended to recover the reminder (i.e., not included in the connection charges) of the distribution grid costs.

- The *customer charge* is typically paid periodically (monthly or bi-monthly, in Euros/month, depending on each type of consumer) and is designed to recover costs associated with consumer management and support. In some cases, this charge is incorporated as a use-of-system charge and is, therefore, not paid separately.

The calculation of efficient network distribution charges is discussed in greater detail in the chapter devoted to tariff design later in this book. Broadly speaking, the guidelines below are followed.

- Total grid costs are broken down into cost items for each grid voltage level.
- Consumers pay the costs of the grid carrying power at the voltage level to which they are connected, plus the costs of any higher, i.e., upstream, grid levels.
- The capacity component (Euros/kW) should be calculated on the basis of consumers' contribution to local system peak demand, i.e., their demand when the distribution grid peaks.
- The energy component (Euros/kWh) takes account of the responsibility of network users in incurring in grid costs other than the need to meet peak demand.

By way of example, the Table 5.3 shows the access rates that include distribution network charges and other regulated charges to be paid as network services by consumers connected to high voltage grids (>1 kV) in Spain.

5.3 Regulated Revenues for Distribution

As noted in the discussion of monopoly regulation, a regulated company's remuneration in the form of the rate revenues collected with regulator authorisation is a key aspect of regulation. Such remuneration should ensure the company's economic and financial viability and must, at the same time, be as low as possible to contribute

to the economic efficiency of the system as a whole. In the case of distributors, regulators must enable the company to strike an optimum balance between costs associated with investment, operation and maintenance, and energy losses on the one hand, and the quality of service provided on the other. To achieve higher quality, the company must obviously incur greater costs and vice versa. As also discussed earlier, the regulated costs that the distribution company must cover are:

- grid infrastructure investment costs:
  - Substations and electric power lines.
  - Facilities and switching equipment: circuit breakers, protection relays, metering and monitoring devices and communications infrastructure.
- operating and maintenance costs:
  - Dispatching centres, maintenance crews, apprenticeship and continuing education courses.
- Other costs related to customer services and other corporate costs.

In addition, distribution companies should reduce energy losses in the grid, and for this reason they must receive an incentive to reduce them. The distributor may also be penalised or credited, according to the actual quality of service provided.

The basic principles to be taken into account by the regulator when determining a distribution company's remuneration are listed below.

- The financial viability of the distribution business must be ensured.
- The characteristics of the areas serviced that underlie the differences in costs incurred by the companies supplying the service (geographic scattering of the load supplied, underground distribution in urban areas as opposed to overhead lines in rural areas or impact of climate or terrain) must be acknowledged.
- Service quality and energy losses requirements must be specifically established for each service area.
- The remuneration associated with the efficient costs of distributing electric power must be determined, in keeping with service quality and energy losses standards.

The Fig. 5.4 plots power distribution unit costs by service areas; each area corresponds to a province in Spain. These areas differ in extension, number of customers and consumed power, ratio of rural versus urban zones and ratio of underground versus aerial grids. The distribution costs were calculated with a reference network model ([32, 39]). This is the model used by the Spanish regulator to calculate efficient, grid infrastructure-related distribution costs (investment, operation and maintenance) and energy losses, using data on the demand loads supplied and distributed generation connected in each area of service. Here the total distribution network costs have been expressed in monetary terms per unit of peak demand. The figure shows that the costs per kW of peak demand change significantly between areas. Urban areas with high load densities show lower unit costs than rural areas with low load densities and with a low level of utilisation of the grid infrastructure.

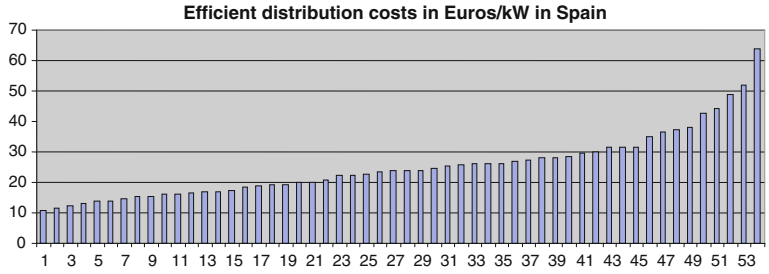


Fig. 5.4 Efficient unit cost of power distribution in Spanish service areas

As also discussed in the chapter on monopoly regulation, the two more representative methods used to determine regulated revenues for distribution companies are the traditional scheme, known as cost-of-service or rate of return and incentive-based regulation through price or revenue caps. In the US, the preferred approach is *performance-based regulation* [23, 35], as described in this book, Sect. 4.3.4.

Cost-of-service regulation is based on audits of the company's accounts and, more specifically, its expenditure and investment records. Such regulation is difficult to implement in this type of business, which involves a large number of small facilities and investments. In the end, the company has little incentive to cut operating costs and undertake the optimum investment that enables it to achieve the established quality standards. As a general rule, the tendency is toward over-investment, justified on the basis of engineering criteria whose wide safety margins and grid redundancy ultimately may lead to low levels of economic efficiency and high rates for end users.

Conversely, *incentive-based regulation,* in the form of price or revenue caps, is gradually being adopted alongside liberalisation and unbundling processes under way in countries around the world. Under price or revenue caps, the regulator establishes the highest price (price cap) or highest revenue (revenue cap) allowed during the regulatory period, which lasts a certain number of years (four or five). Thereafter, prices or revenues rise or fall with the retail price index, *RPI*, or inflation rate, less an adjustment X factor determined by the regulator (RPI-X). In the case of revenue caps, a provision for revenue growth in keeping with the growth of some cost drivers subject to a certain economies of scale factor may be included. A typical revenue cap formula is described below by way of illustration (Eq. 4.6).[2]

[2] Equation (5.1) can be seen under an interesting different perspective, when the factor X is understood as a mere adjustment factor, as explained in Sect. 4.4. Under this perspective, the growth in distributed energy and number of consumers have been already taken into account when estimating all distribution costs for the next regulatory period and the first two factors in Eq. (5.1) seem to be all that is needed. What is then the role of the third factor $(1 + \alpha \times \Delta D_t)$? Now ΔD_t represents the *deviation* (not the absolute value) of the cost driver D_t with respect to the value that was estimated in the price control exercise when the value of X was computed. This

$$R_t = R_{t-1} \times (1 + \mathrm{RPI}_t - X) \times (1 + \alpha \times \Delta D_t) \tag{5.1}$$

where

R_t is the authorised remuneration or revenues in year t, where $t = 1, 2, 3, 4$ (the years in the regulatory period),

RPI_t is the retail price index in year t, per unit,

X is an adjustment factor, per unit,

α is the economies of scale factor, ranging in value from 0 to 1, which estimates how regulated costs and therefore revenues increase in proportion to company sales or other cost drivers and

ΔD_t is the increment in year t, per unit, of the selected cost driver(s), such as units of distributed energy, number of customers, length of the network or a combination of these.

In addition, the company's yearly revenues should be adjusted to provide incentives to lower energy losses or improve quality of supply [2].

By way of example, four such regulatory schemes are discussed later in this chapter: the pioneer scheme implemented in the United Kingdom, followed by Norway, California and Spain.

5.4 Quality of Supply

Regulated distribution utilities should strike an optimal balance between their investment and operation and maintenance costs on the one hand and the quality of supply provided to consumers on the other, since, in electricity distribution, investment and maintenance costs and quality of supply are clearly related. The higher the costs the better the service quality, and vice versa [13].

The preceding two sections also showed that regulators are introducing incentive-based regulation to encourage distributors to cut costs and raise profits. Clearly, one primary source of savings is to decrease infrastructure investment and reduce the resources devoted to facility maintenance, although this entails lowering the quality of supply provided by the distributor. Obviously, then, any price or revenue cap remuneration scheme must include a mechanism to link such remuneration to company attainment of quality objectives. This mechanism normally adopts the form of financial penalties when the quality actually supplied fails to reach the regulator's targets. Conversely, the regulator may provide for financial rewards for companies that deliver quality above and beyond the established objectives.

(Footnote 2 continued)
third multiplicative factor in Eq. (5.1) is now a very convenient *ex-post* adjustment factor, which automatically updates the remuneration of the distribution utility when the actual considered cost drivers deviate from their estimated values at the beginning of the regulatory period.

From the standpoint of electric power supply, quality of service is characterised by three different properties [38].

- *Continuity of supply* is measured by the number and duration of supply interruptions.
- *Power quality or voltage quality* is measured in terms of the disturbances affecting the ideal voltage wave parameters: variations in voltage magnitude, periodic oscillations in voltage, harmonics, voltage dips and short interruptions (lasting <3 min).
- *Commercial quality* or *customer service* is measured by indicators such as the time taken to process and act on customer applications for service, time taken to respond to complaints about poor quality, or the number of bills based on estimated rather than actual readings. Some overlap may exist between the distribution and retail activities and obligations and responsibilities should be clearly established in this respect.

The rest of this chapter addresses continuity of supply and product quality in great detail and examines international experience in quality regulation. For further details on this topic, see Ref. [11].

5.4.1 Continuity of Supply

Continuity of supply or the number and duration of supply interruptions is clearly associated with the distributor's investment and maintenance policy. The use of low-cost, poor-quality materials leads to a higher equipment failure rate. The logical outcome of having fewer maintenance crews or a poorly designed network is that it takes longer to repair failures and restore service.

Electric power failures generally involve costs for consumers. Manufacturing processes, for instance, must obviously be discontinued during power outages, at a direct cost to manufacturers that varies depending on type of process and the duration of the incident. Such direct and indirect costs are generally appraised by what is known as the *cost of energy not supplied* incurred by the consumer. Under the traditional *cost-of-service* regulatory scheme, distributors maintained an appropriate quality of supply by investing in facilities as necessary and making other expenditures to keep supply outages below a certain threshold. When interruptions did occur, the company's image suffered but not its profit and loss account (except for the loss of income, minus production costs, for the energy not supplied), since customers usually received no financial compensation for poor quality. But more recently, there is a tendency to hold company decision-making responsible for poor quality and, therefore, to expect distributors to internalise the resulting costs to consumers.

In keeping with this view of the problem, the net social cost associated with delivering a certain level of quality should be minimised for society as a whole (distributors plus consumers). In other words, the objective is to minimise the sum

of the cost of investment and maintenance for distributors to reach a certain level of quality, and the cost to consumers when the power outages resulting from that level of quality occur. Figure 5.5 shows the two cost curves and the curve resulting from the sum of the two.

According to the above graph, the aim of regulation is to minimise the net social cost, which entails sending economic signals that encourage distributors to raise quality from the current to the optimum level (OQL). This is achieved by measuring *system reliability indices* and implementing an *incentive and penalty mechanism* to adjust distributor remuneration. The above cost curves obviously depend on the area serviced by the distributor and the characteristics of the consumers. As a rule, grid investment and maintenance costs to improve reliability indices are higher in *rural areas*, where quality is nonetheless always lower than in *urban areas* serviced by underground, insulated cables with a network topology with more redundancy.

In Spain, the reliability indices that are used to quantify the quality of service for a given area are known as TIEPI (Spanish acronym for installed capacity equivalent interruption time) and NIEPI (installed capacity equivalent number of interruptions) and defined as follows:

$$\text{TIEPI} = \frac{\sum_{i=1}^{n} P_i \cdot T_i}{P} \tag{5.2}$$

where

- n is the number of power interruptions in the area over the period of time considered, e.g., one year
- T_i is the duration of the power outage i
- P_i is the nominal capacity of the MV/LV transformers interrupted during the power outage i
- P is the nominal capacity of all MV/LV transformers in the area considered.

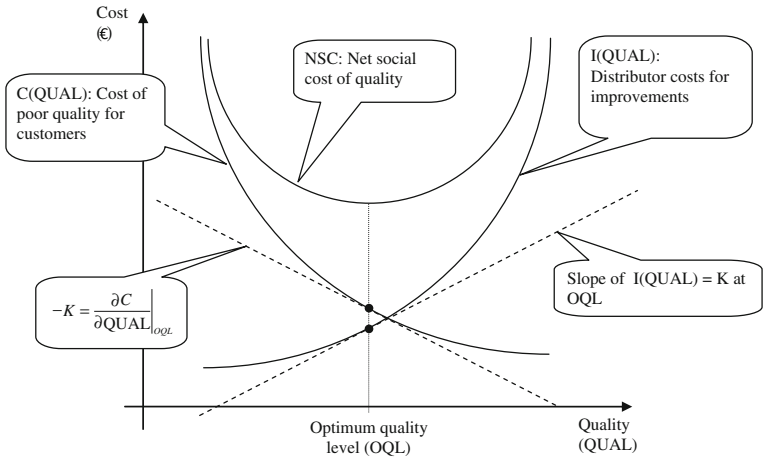


Fig. 5.5 Minimisation of net social cost in terms of quality of supply. *Source* [38]

The NIEPI for an area is defined as:

$$\text{NIEPI} = \frac{\sum_{i=1}^{n} P_i}{P} \tag{5.3}$$

The TIEPI and NIEPI indices are similar to the better-known SAIDI (system average interruption duration index) and SAIFI (system average interruption frequency index), which take the number of customers affected into consideration instead of the transformer installed capacity. For a more complete definition of reliability indices, see Ref. [15].

The incentive/penalty mechanism is formulated as follows [38]:

1. The reference or base remuneration received by the distributor is associated with a reference level of quality. This means that a reference TIEPI and a reference NIEPI must be established for each of type of area serviced by the distributor, such as rural, semi-urban or urban.
2. When the quality actually delivered in an area is better than the reference quality, the company receives a bonus, whereas if it is poorer, the operator is penalised.
3. To encourage the shift from current to optimal quality, the unit value of the penalty or bonus should be the value K in Fig. 5.5, i.e., the value of the slope of the cost curves at the optimum quality level (OQL).

The sound optimisation properties of this reward/penalty scheme can be explained by examining Fig. 5.5. The optimal quality level is set at the point where the marginal costs (for distribution companies) and marginal benefits (for customers) of quality increments are equal. For levels of quality below the optimal level, the benefits generated by an extra unit of quality are greater than the costs incurred by distribution companies for that quality increment; for levels above the optimal one, the costs of an extra unit of quality are greater than the benefits it provides to customers. Therefore, the reward/penalty should be set at the point where the marginal cost of quality equals the marginal willingness of consumers to pay for a higher quality level. Figure 5.5 shows that, for levels of quality below the optimal level, the proposed incentive regulation happens to be a win–win one for the distribution company and for the consumers.

A second, complementary objective of regulation is to guarantee all consumers certain minimum levels of individual quality. *Individual continuity indices*—the number and duration of interruptions affecting each individual customer—are used for this purpose, along with an individual penalisation mechanism. When the quality of the customer's power supply is lower than the regulator-established minimum, the customer must be financially indemnified by the distributor. Such compensation must be substantial enough to protect the customer against the damages caused by poor quality and act as an incentive for the company to correct the problem at the root. The indemnity involved is usually defined in terms of the *cost of energy not supplied* to the customer. In other words, the energy that the

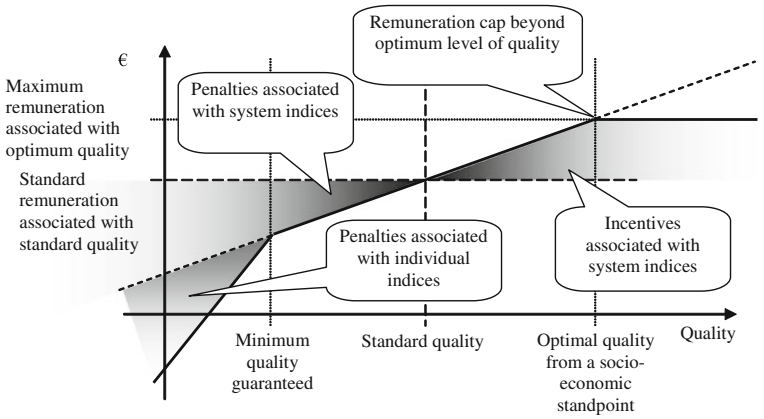


Fig. 5.6 Distributor remuneration in terms of quality delivered. *Source* adapted from [37]

customer failed to receive due to the power outage is estimated in accordance with its consumption patterns, and this amount is then multiplied by the respective penalty, such as €2/kWh not supplied, for instance, to determine the total sum.

Figure 5.6 summarises the two quality incentive/penalty mechanisms discussed here.

Implementation of these two complementary mechanisms ensures that the dual objective pursued is achieved.

1. The average probability distribution function for individual levels of quality is monitored against *system or area indices*, and the distributor's remuneration is adjusted according to the results or outcomes obtained in terms of the quality actually delivered.
2. The tails of the probability distribution for individual levels of quality are monitored against individual customer indices, thereby guaranteeing quality standards for all customers and minimum distribution function variance.

Figure 5.7 illustrates the above concepts, analysing the impact that monitoring the two types of indices, *system and individual*, has on the probability curve for power interruptions experienced by the company. The objective is to avoid the occurrence of poor quality pockets and ensure that all consumers receive a product that meets at least certain minimum quality standards.

5.4.2 Power Quality

The other technical property of electricity associated with the quality of the power supplied is known as *power quality* or *voltage quality*. The voltage wave delivered by the distributor to the customers at their supply points should meet certain ideal requirements, defined with respect to the following characteristics: *voltage*

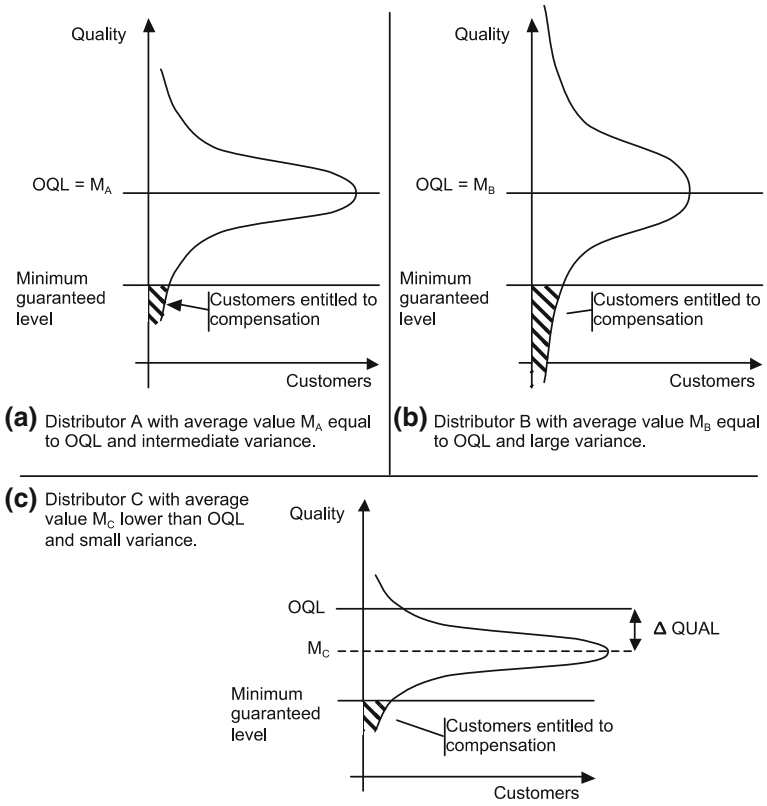


Fig. 5.7 Probability distribution function for distributor quality levels depending on whether area and/or individual reliability indices are monitored. *Source* adapted from [37]. *OQL* stands for Optimal Quality Level

magnitude relative to the nominal supply voltage, e.g., 230 V for residential consumers in Europe; *voltage frequency*, 50 Hz in Europe; concurrence of the *voltage wave form* with a sine wave and *voltage symmetry* of the three phases in the event of three-phase supply. In practice, these characteristics defining the *electricity product* are subject to disturbances that, if substantial, may cause the system or the equipment connected to it to malfunction. The most common types of electromagnetic disturbances appearing in electric power distribution grids include harmonics, periodic or non-periodic voltage oscillations, voltage dips and over voltages.

The objective pursued for power quality is known as *electromagnetic compatibility* (EMC). EMC is defined as the ability of a device, apparatus or system to satisfactorily operate in its electromagnetic environment without causing intolerable electromagnetic disturbances. In this regard, there are two sides to any device or system:

- *Emission* of voltage, currents or electromagnetic fields that may potentially cause disturbances.
- *Susceptibility* to the adverse effects of electromagnetic disturbances.

The EMC level for each type of disturbance is defined as the level of disturbance at which there is an acceptable probability of EMC. This level may be exceeded for only a small percentage of the time (usually under 5 %).

Figure 5.8 shows how, for a given type of disturbance, compatibility can be achieved between the emissions from the devices and equipment causing the disturbance (curve on the left of the Fig. 5.8) and the susceptibility levels that ensure proper system operation (curve on the right of the Fig. 5.8) by setting the EMC level to standards endorsed by international bodies, in particular the International Electrotechnical Commission or IEC [17].

The characteristics associated with power quality and voltage disturbances are, therefore, regulated at several different levels.

- Coordination committees define the EMC levels laid down in international standards.
- Product committees establish equipment susceptibility requirements to be met by manufacturers.

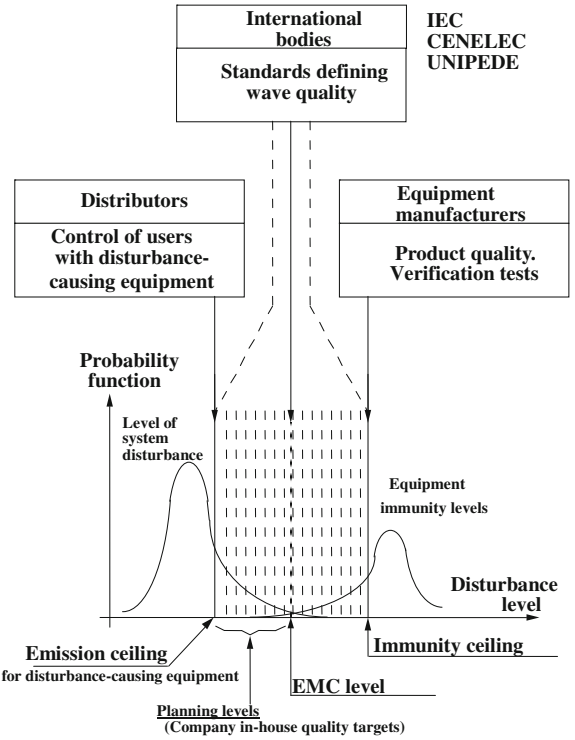


Fig. 5.8 EMC levels and emission and susceptibility limits for one type of disturbance. *Source* adapted from IEC [17]

- Distribution companies set overall emission limits and allocate them to users connecting electric facilities and appliances.
- Industrial users specify and design their facilities to comply with emission and susceptibility limits.
- Manufacturers of small electrical appliances for residential or commercial use must meet standards that limit the grid disturbance caused by this equipment.

In the European Union, CENELEC (European Committee for Electrotechnical Standardisation) standard EN-50.160, compliance with which is mandatory in Member States, is aimed at standardising the definition of the electricity product. Among other stipulations, it specifies disturbance ceilings at customer service connections. For most disturbances, such ceilings generally concur with the EMC levels established by IEC standards. Due to the random nature of disturbances, the compliance criterion is established in terms of testing or measurement time, which is usually one week. The established ceilings may not be exceeded during 95 % of that time. One important exception to this rule has to do with voltage dips and short interruptions: the standard contains mere recommendations on the acceptable frequency of this sort of disturbance in grids [8].

IEC standard 61000-4-30, in turn, specifies the procedures and instruments required to measure disturbance levels in grids and at points of supply. In general, the IEC 61000-X-XX family of standards regulates the many different aspects of *electromagnetic compatibility*.

From a regulatory standpoint, distributors are required to comply with these standards and resolve any problems that may arise within a certain time frame; otherwise, financial penalties are imposed to prevent problems from becoming perennial. In another vein, certain kinds of disturbance, such as flicker or harmonics, are directly caused by customer equipment. Distributors may impose *disturbance emission limits* in such cases and fine customers failing to observe these limits, or, if the problem persists, interrupt their power supply [38].

5.5 International Experience in Distribution Regulation

In this section several examples of distribution regulation are presented. They represent real examples of how incentive regulation and quality of service control are applied in practice. From the standpoint of quality regulation, continuity of supply (power interruptions) is clearly the chief issue for regulators because it can be associated with distributor investment and costs and, consequently, with remuneration. In this regard, the schemes that have introduced incentive regulation to reduce costs are also concerned with regulating and controlling the quality of supply delivered.

5.5.1 The United Kingdom

In Europe, the United Kingdom pioneered unbundling and the subsequent privatisation of the State-run electricity industry into several generators and distributors under the 1989 Energy Act. First, this legislation created a pool or wholesale electricity market into which all generators were to bid. It also established a programme of gradual eligibility to enable consumers to freely choose their power supplier, which reached completion in 1999. Fourteen companies, the "RECs" or Regional Electricity Companies, were created and privatised to distribute and supply electric power in their respective service areas in Great Britain. Legal separation between distribution and supply became mandatory in October 2001.

From the outset, the regulator—initially OFFER and presently OFGEM (Office of Gas and Electricity Markets)—established a price cap formula for limiting the annual remuneration received by distributors, which was to be revised every 5 years. The first revision was performed in 1994 and adjusted in 1995, when price caps were set for the period 1996–2000. Subsequent revisions have since set price caps for the periods 2000–2005 and 2005–2010. Under the *"revenue yield"* formula used, average company revenues are updated annually with the RPI-X factor. Average revenues are calculated with the number of kWh distributed (weighting low and high voltage kWh differently), with the number of consumers weighted equally. An incentive for energy loss reduction below a certain benchmark value is likewise included in the calculations. This incentive was initially valued, in 1992–1993, at a price of 3 pence per kWh of energy losses reduced. With the present arrangements, this value is also updated in keeping with the RPI-X factor. In the last review process, this incentive was set at 4.8 pence per kWh.

The United Kingdom's National Audit Office, an independent public spending audit and control body that reports to Parliament, has evaluated the regulatory experience of distribution companies very highly [27]. The 1994–1995 rate revision lowered initial prices in two steps, first by 11 to 17 %, depending on the company, and then by 10–13 %. Moreover, the productivity factor X initially established for all companies at the same value of −2 % was reduced further to −3 % per year for the period 1996–2000. From 1994 to 1998, efficiency improvements led to average cuts in operating costs of 25 %. The price review conducted in 1999–2000 lowered prices by an average of 24 % in real terms, for a 5 % reduction in the rate finally paid by electric power consumers. A 3 % yearly price reduction, in real terms, was established for the period 2000–2005. Lastly, for the period 2005–2010, the X factor has been set at zero for the first time.

In the 2004 price control [30] the average revenue cap, known as "*allowed demand revenue*", was calculated by adding up the base demand revenue, the allowed pass-through items, the amount due to incentives for loss reduction, improved quality of service and innovation funding, and finally a correction term for positive or negative deviations between allowed and actual revenues due to differences between the forecast and actual values of variables and parameters.

Base demand revenue was calculated using the building block approach. A separate assessment was established for operational expenses (OPEX) using benchmarking techniques and for capital expenditures (CAPEX) using distribution network operator (DNO) investment projections. In addition to the (RPI-X) factor, the base demand revenue for high and low voltage distribution costs was updated annually with a growth term. This term was calculated as 50 % of the annual growth in units distributed (kWh) per voltage category, weighted by a specific price for each category and 50 % of the annual growth in metering points [33].

Under this type of regulation, the determination of a "reasonable" allowed capital expenditure (CAPEX) to be paid to distributors is a critical issue. A novel approach, a sliding-scale mechanism known as the Information Quality Incentive (IQI), was introduced by OFGEM in the 2004 price control review that allowed DNOs to choose between getting a lower CAPEX allowance but a higher expected return on investment (retaining more of the cost reduction if they can better the target expenditure levels) or a higher CAPEX allowance combined with a lower expected return [30].

According to Table 5.4, a power ratio efficiency incentive is assigned to each DNO, obtained as the ratio between the CAPEX target selected by the company and the figure recommended by OFGEM's consultant (PB Power). For instance, if a DNO gets a CAPEX allowance equal to 110 % (DNO:PB Power ratio of 120 in the figure), it would get a bonus of 0.6 % of its target income. In this case, if the DNO's actual CAPEX were 70 % of the target (due to improvements in efficiency), it would get a 12.6 % increase in income as a reward. By contrast, if its actual CAPEX are over 140 % of the target, its income would be reduced by 8.4 %. This approach is a way of introducing incentives for DNOs to achieve efficiency in network investments [22].

Moreover, this mechanism is a way to prevent overestimation by DNOs when declaring future investment plans. By way of illustration, the IQI mechanism is broken down into the following three consecutive steps.

1. Each DNO sends OFGEM an ex-ante expenditure forecast in line with what it expects to spend ex-post. This is justified by inspection of each row of the matrix in Table 5.4. Note that in each row the maximum reward, identified in bold, is obtained when the forecast concurs with the expected actual expenditure, regardless of the expenditure target set by OFGEM. This removes any incentives for overestimation.
2. Once the expenditure forecasts have been established by the DNOs, OFGEM sets the target or baseline for each company and a DNO/PB power ratio is calculated for each. The DNOs are positioned in specific columns in the matrix in Table 5.4.
3. Every DNO has an incentive to spend as little as possible through efficient operation. This is justified, as explained earlier, by the figures in each column of the matrix in Table 5.4, which show that rewards increase as actual DNO expenditures decrease.

Table 5.4 DNO CAPEX-related incentives (allowed vs. actual)

DNO: PB Power ratio efficiency incentive	100 40 %	105 38 %	110 35 %	115 33 %	120 30 %	125 28 %	130 25 %	135 23 %	140 20 %
Additional income	2.5	2.1	1.6	1.1	0.6	−0.1	−0.8	−1.6	−2.4
As pre-tax rate of return (%)	0.200	0.168	0.130	0.090	0.046	−0.004	−0.062	−0.124	−0.192
Allowed exp.									
Rewards and penalties	105	106.25	107.5	108.75	110	111.25	112.5	113.75	115
Actual expenditures									
70	16.5	15.7	14.8	13.7	12.6	11.3	9.9	8.3	6.6
80	12.5	11.9	11.3	10.5	9.6	8.5	7.4	6.0	4.6
90	8.5	8.2	7.8	7.2	6.6	5.8	4.9	3.8	2.6
100	**4.5**	4.4	4.3	4.0	3.6	3.0	2.4	1.5	0.6
115	2.5	**2.6**	2.5	2.3	2.1	1.7	1.1	0.4	−0.4
110	0.5	0.7	**0.8**	0.7	0.6	0.3	−0.1	−0.7	−1.4
115	−1.5	−1.2	−1.0	**−0.9**	−0.9	−1.1	−1.4	−1.8	−2.4
120	−3.5	−3.1	−2.7	−2.5	**−2.4**	−2.5	−2.6	−3.0	−3.4
125	−5.5	−4.9	−4.5	−4.2	−3.9	**−3.8**	−3.9	−4.1	−4.4
130	−7.5	−6.8	−6.2	−5.8	−5.4	−5.2	**−5.1**	−5.2	−5.4
135	−9.5	−8.7	−8.0	−7.4	−6.9	−6.6	−6.4	**−6.3**	−6.4
140	−11.5	−10.6	−9.7	−9.0	−8.4	−8.0	−7.6	−7.5	**−7.4**

Source OFGEM [30]

Table 5.5 Example of how OFGEM sliding-scale incentives are applied

DNO forecast [M€]	125	100
DNO: PB Power ratio [%]	125	100
Efficiency incentive [%]	28	40
Additional income [%]	−0, 1	2.5
Allowed investment [M€]	111.25 % * 100 = 111.25	105 % * 100 = 105
Actual investment [M€]	100	100
Actual efficiency incentive [M€]	28 % * (111.25−100) = **3.15**	40 % * (105−100) = **2**
Additional income [M€]	(−0.1 %) * 100 = **(−0.1)**	2.5 % * 100 = **2.5**
Final income [M€]	100 + 3.15 − 0.1 = **103.05**	100 + 2 + 2.5 = **104.5**

An example is shown in Table 5.5. The first column lists the total income of a DNO that actually invested €100 million, with an initial estimate of €125 million and a DNO:PB power ratio of 125. In the second column, the total income is calculated assuming that the initial company estimate was €100 million, which concurred with the actual expenditure. Note that the figures calculated can also be found directly by using the matrix in Table 5.4 (expenditure of 100 with a ratio of 125, and an expenditure of 100 with a ratio of 100, respectively). For the same actual DNO expenditure, the total DNO income is higher in the second case than in the first. Therefore, DNOs have an incentive to adjust their investment projections to avoid overestimation.

A good review of how incentive regulation has been working in the UK since inception can be found in [21].

Finally, after a long period of 20 years of successful implementation of incentive regulation, OFGEM has consulted stakeholders about rethinking energy network regulation to address the new challenges that are anticipated. The challenge is to design a new framework for regulating networks under the different conditions required for a sustainable and low-carbon energy sector. This framework would encourage network companies to focus more on the longer term, developing network solutions for present and future consumers and increasing innovation to gradually adapt to the new needs. Some of the ideas about the new framework are: it would continue to use an *ex-ante building block* approach; it would put greater focus on the delivery of outcomes relating to safe, secure, high-quality and sustainable network services; it would promote long-term cost-cutting incentives, as at least part of the regulatory package would be extended to more than 5 years; and it would provide a separate, time-limited innovation stimulus and specific incentives for achieving a low-carbon energy sector [31]. Finally, a decision document has been released in October 2010 to implement a new regulatory framework, known as the RIIO model (revenue = incentives + innovation + outputs) in next price control reviews.

Regarding quality of service regulation, in the United Kingdom, in the early years of restructuring and privatisation, the regulator focused on controlling commercial quality indicators. Overall company indices and individual service to each customer were monitored. Penalties were established for both types of indices

Table 5.6 Incentive scheme for quality of service in Great Britain

Incentive arrangement (% of revenue)	Third price control review (2000/2001–2004/2005)	Fourth price control review (2004/2005–2009/2010)
Interruption incentive scheme		
Duration of interruptions (CML)	±1.25 %	±1.8 %
Number of interruptions (CI)	±0.5 %	±1.2 %
Service restoration after storms	−1 %	−2 %
Other performance standards	Uncapped	Uncapped
Quality of telephone response	±0.125 %	±0.05 – 0.25 %
In storm conditions	Not applicable	0 % initially to ±0.25 % for 3 years
Discretionary reward scheme	Not applicable	Up to +£1 m
Overall cap/total	−2.875 to +2 %	−4 % and no overall cap on the up side

Source [46]

in the event of distributor non-compliance. At the time, continuity of supply was regarded as acceptable, with companies implementing grid planning and development according to the same recommendations that the State-run company had followed prior to privatisation. In the two latest price controls, the regulator, OFGEM, has monitored system-level quality indices for each distributor. The two parameters monitored are SAIDI (system average interruption duration index, in number of customer minutes lost [CML] per connected consumer) and SAIFI (system average interruption frequency index, in number of customers interrupted [CI] per 100 consumers). According to the data compiled, both indices have improved over the past 10 years. The regulator has implemented procedures for measuring and controlling interruptions, and incentive and penalty mechanisms that link company remuneration to quality performance. The targets or reference values for each company are individually set by the regulator with allowance for the geographic and historical differences among the areas serviced by the various distributors. Distributors receive bonuses (penalties) if their performance is better (worse) than the specific reference values established by the regulator [46]. The Table 5.6 is an illustration of how the maximum amount of bonuses/penalties associated with quality of service, expressed as a percentage of total company revenue for different quality indicators, increased in the last price control review.

5.5.2 Norway

Norway was one of the first countries to introduce competition in generation and supply, in 1991. There were 198 regional or municipal public utilities in Norway. Through 1996, these companies were regulated under a cost-of-service scheme.

Incentive-based regulation for electricity distribution companies was introduced in 1997 and involved a revenue cap formula for the following 5 years, beginning with the period 1997–2001 [14]. The formula for paying each distributor was indexed to inflation, a productivity factor and variations in demand loads, and also included a term for considering actual energy losses. The formula applied in 1998 is shown below.

$$\mathrm{IT}_{e,98} = \left[\left(\mathrm{IT}_{e,\mathrm{w/losses97}} \cdot \left(\mathrm{KPI}_{98}/\mathrm{KPI}_{97}\right)\right) + \left(\mathrm{NT}_{\mathrm{MWh}} \cdot P_{98}\right)\right] \cdot \left[1 + 0,5 \cdot \Delta\mathrm{LE}_{a,98-97}\right] \cdot (1 - \mathrm{EFK}) \tag{5.4}$$

where

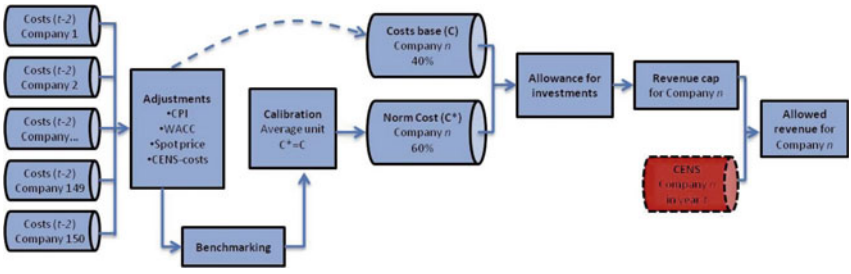
$\mathrm{IT}_{e,98}$	is the revenues allowed in 1998,
$\mathrm{IT}_{e,\mathrm{w/losses\ 97}}$	is the revenues, with no allowance for energy losses, in the preceding year, 1997,
$\mathrm{KPI}_{98}/\mathrm{KPI}_{97}$	is the adjustment for variations in the inflation index,
$\mathrm{NT}_{\mathrm{MWh}} \cdot P_{98}$	is the cost of actual energy losses incurred in the distributor's grid, i.e., energy losses multiplied by the spot price on the wholesale market during the year in question,
$\Delta\mathrm{LE}_{a,98-97}$	is the forecast yearly increase in energy distributed; 0.5 is the economies of scale factor that converts growth in energy distributed into an increase in allowable revenues and
EFK	is the productivity factor.

Norway has implemented a benchmarking technique based on *data envelopment analysis* (DEA) for comparing the efficiency levels of the various distributors. Benchmarking techniques applied to distribution companies are described in Jamasb and Pollitt [20]. The DEA model in Norway quantified a series of output variables, initially the number of customers, energy delivered, length of high voltage lines, length of low voltage lines and length of sea cables, against a series of input variables, initially the number of man-years, network energy losses, capital costs based on book or replacement value and goods and services. The most efficient distributors were the ones with the best output/input ratios, i.e., the companies with the lowest costs and the best results were awarded a 100 % efficiency factor. The productivity factors for each distributor, *EFK* in Eq. (5.4), were set year by year on the basis of the aforementioned analysis. The most efficient distributors' revenues were discounted less than those earned by the least efficient. A 2 % productivity factor was assigned to all distributors in 1998. Differences between distributors were established through the initial remuneration set for each. For more information about benchmarking in Norway, see Ref. [1].

In the 2002 revision, the increase in remuneration based on the growth of delivered energy was replaced by a scale factor that was applied to a combination of the average nationwide load increase and the infrastructure growth in the company's service area. Furthermore, the adjustments in remuneration in the second period were made by applying the formula ex-post instead of ex-ante, as was done in the first period [40].

In 2007, the revenue cap was based on a new yardstick formula to be revised after a 5-year period. The allowed revenue is 40 % of the distributor's actual cost and 60 % of an efficient cost, or norm cost, calculated by benchmarking techniques, with a 2-year lag. In addition, a correction term for adjusting company investments is included in the allowed revenue formula. The base cost is established using the sum of operating and maintenance costs, depreciation, the cost of physical losses and the cost of energy not supplied (CENS). The regulatory asset base is determined by book values. The benchmarking is based on DEA using nationwide data. The correction term for investment is designed to take into consideration network extensions and replacements that were not included in the revenue cap due to the 2-year lag [28]. The Fig. 5.9 illustrates the process described.

Regarding quality of service regulation, a penalisation scheme was introduced in Norway in 1998 to compensate transmission grid users for supply outages. Under this scheme, the price for energy not supplied was NOK 16/kWh ($2.1/kWh) for long power interruptions (i.e., over 3 min) and NOK 8/kW for short interruptions (<3 min), excluding the first short interruption of the year. The yearly compensation that could be paid to users was capped at 25 % of their payments for use of the grid. Since then, a detailed set of specific quality regulations has been developed by the regulator. Based on data reported by the companies regarding long and short interruptions, the regulator has developed a standard procedure to calculate the cost of the energy not supplied that is taken



The economic regulation from 2007 – Annual procedure

- The cost base is set on the basis of company specific costs as reported in the last available financial reports –2 year lag.
- The norm cost is set annually using comparative efficiency analyses (DEA benchmarking) where geographical differences are taken into account.
- The norm cost is calibrated such that an average efficient utility will have a norm cost equal to its cost base.
- OPEX (Operating and maintenance costs)and CENS-costs are adjusted for inflation (CPI).
- Regulatory asset base is based on book values (Historical cost –accumulated depreciations) by the end of year t-2 + 1% working capital.
- The WACC is based on the annual average of a 5 year government bond +a risk premium of app. 3,1% with a risk free nominal rate of 5%.
- The reference price for electricity is based on a weighted average of local area spot prices at Nord Pool Spot AS + NOK 11/MWh –Network losses.
- Allowance for investments based on investments in year t-2–Present value loss due to 2 year lag
- The allowance for investments applies to both new investments and reinvestments
- CENS is included in the cost base and in the cost norm as any other cost.

Fig. 5.9 Calculation of allowed revenue for distribution companies in Norway. *Source* [28]

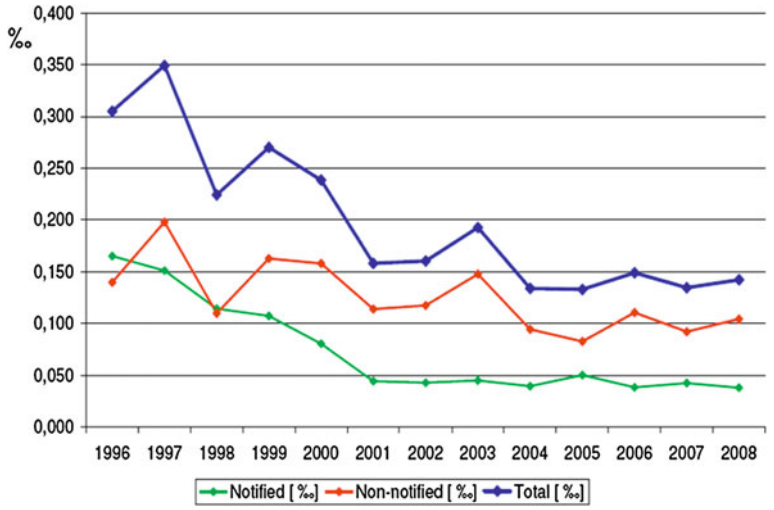


Fig. 5.10 Energy not supplied in per thousand of the energy supplied in Norway. *Source* [29]

into consideration for benchmarking, as explained in Sect. 5.4, while setting revenues for each company. Therefore, this is another way to encourage companies to improve their quality of service indices. Figure 5.10 shows the changes in energy not supplied, distinguishing interruptions that were planned and notified in advance from those that were not.

5.5.3 California

Although competition was introduced in generation and supply in California in 1998, the three large, formerly vertically integrated utilities had been regulated under *performance-based ratemaking* (PBR) mechanisms since 1992 [41]. When the three utilities (Pacific Gas and Electric or PG&E, Southern California Edison or SCE and San Diego Gas and Electric or SDG&E) were restructured, distribution continued to be regulated by the state regulator, the California Public Utilities Commission (CPUC). In 1997, for instance, SCE agreed with the CPUC to renew the PBR plan for its transmission and distribution activities for 5 years. This formula called for price caps with productivity factors that started off at 1.2 % in 1997, rising to 1.4 % in 1998 and 1.6 % from 1999 to 2001. A progressive symmetric mechanism for *sharing earnings and losses* with ratepayers was likewise introduced. This scheme established three bands, as follows.

- The inner band comprised 50 basis points around the benchmark rate of return (defined in terms of a range above or below a benchmark ROR: for instance, if the benchmark rate is 8 % and the cost reductions obtained increase the rate to

10 %, the distributor "earns" 200 basis points). The distributor received all net revenues retaining 100 % of the earnings or losses.
- The middle bands comprised 50–300 basis points. The distributor's share of earnings or losses rose stepwise from 25 to 100 %.
- The outer band comprises 300–600 basis points. The distributor retained all earnings or losses.

Beyond 600 points, the plan and, therefore, the remuneration formula were subject to revision [26].

In addition, the SCE PBR plan for 2006–2008 included a reliability investment incentive mechanism (RIIM) for reliability-related capital expenditures and capital additions specifically identified as preserving long-term electric service reliability for customers. The following expenditures fall under this mechanism: distribution infrastructure replacement, preventive maintenance, substation infrastructure replacement, load growth infrastructure and capital additions. At the end of the period, an ex-post check is performed of the differences between the forecast and the actual expenditures, with a provision for refunds to customers if actual expenditures were lower than authorised revenues [43].

Lastly, PBR plans typically include a provision for accounts devoted to specific aims or programmes. The actual and forecast deviations in these accounts are tracked by the regulator in much the same way as explained for the RIIM. Examples of the accounts in the SCE plan are advanced metering infrastructure, demand response programme, procurement energy efficiency, California solar initiative programme and public purpose programmes [43].

Regarding quality of service regulation in California, quality indicators are used to measure service reliability, company employee safety and consumer satisfaction. The real levels of service provided are measured in accordance with each indicator and compared to a benchmark value. In the SCE PBR plan, for instance, service reliability is assessed with three indices: system average interruption duration index (SAIDI) for long (sustained) interruptions of more than 5 min; system average interruption frequency index (SAIFI) for long interruptions; and momentary average interruption frequency index (MAIFI) for short interruptions. The indicator used to measure employee safety is based on reportable lost-time and non-lost-time injuries and illnesses in relation to total employee working hours. Consumer satisfaction is evaluated through surveys addressing four areas of company service: field service and meter reading activities, in-person services, telephone centre operations and service planning activities. A customer satisfaction rating is obtained as the percent of customer responses in the top two of the six response categories in the survey. The incentives and penalties associated with compliance with the SAIDI target are shown in the Table 5.7, by way of illustration.

Table 5.7 Incentives for improving SAIDI in the SCE PBR plan

Benchmark	56 min
Deadband	±6 min
Liveband	±9 min
Unit of change	1 min
Incentive per unit	$2 million
Maximum incentive/penalty	±$18 million

Source [43]

5.5.4 Spain

Wholesale competition was introduced in Spain in 1998, as well as retail competition but only for large (eligible) consumers. Initially, the eligibility threshold set was >15 GWh of energy consumed per year. Beginning in 2003, all consumers became eligible. The remuneration formula for distributors, including both the distribution business and supply to regulated customers, was based on a revenue cap for nationwide distribution as a whole, indexed to yearly inflation, a productivity factor and adjustments for increases in demand growth [40]. The *revenue cap* formula was as follows:

$$\mathrm{DR}_t = \mathrm{DR}_{t-1} \times (1 + (\mathrm{IPC}_t - 1)/100) \times (1 + (\Delta D_t \times \mathrm{Fe})) \tag{5.5}$$

where

DR_{t-1} is the remuneration received by distributors in year $t-1$,
IPC_t is the inflation rate in present year t in percent,
ΔD_t is the demand load increase in year t in per unit and
Fe is the scale factor relating increases in distributed power to increases in allowable remuneration. This factor was adjusted to a value of 0.3.

As Eq. (5.5) shows, productivity was set at a rate of 1 % for distribution as a whole in the regulatory period.

The total remuneration was divided among the five major distributors. Their shares were computed from a combination of the percentages in effect under the earlier legislation framework and the output from a "benchmark network" model. This benchmark network model was used to plan and design the optimally efficient grids that companies should have in place in each Spanish province.

In 2008, a new regulation for distribution was enacted [4]. A revenue cap formula was established for each distributor individually, instead of a single figure for distribution as a whole to be subsequently apportioned out among the companies involved. The key steps in the revenue cap formula are the determination of the baseline for each distributor and, then, the update of the remuneration on a yearly basis, taking inflation, an adjustment factor and an increase in network connections into consideration. The new cap formula applied to each distribution company i is as follows:

$$
\begin{aligned}
R_0^i &= R_{\text{base}}^i \cdot (1 + IA_0) \\
R_1^i &= R_0^i \cdot (1 + IA_1) + Y_0^i + Q_0^i + P_0^i \\
R_k^i &= \left(R_{k-1}^i - Q_{k-2}^i + P_{k-2}^i\right) \cdot (1 + IA_k) + Y_{k-1}^i + Q_{k-1}^i + P_{k-1}^i;\ k = 2, 3, 4.
\end{aligned}
\tag{5.6}
$$

where

R_{base}^i is the reference remuneration for company i established by the regulator, including remuneration for (i) investment, linear depreciation of existing assets and a rate of return applied to the net assets, (ii) operation and maintenance costs and (iii) other costs such as metering, billing and new connections,

R_0^i is the reference remuneration for company i updated to the year of calculation,

R_k^i is the remuneration for company i in year k of the regulatory period,

Y_{k-1}^i is the change in the allowed remuneration for company i due to the variation in distribution activity in year $k-1$; this is computed by using a reference network model, which determines the variation in distribution costs, CAPEX and OPEX due to demand growth and the connection of new network users, loads and distributed generators,

Q_{k-1}^i is the change in the allowed remuneration for company i due to the level of quality of service performance reported in year $k-1$; this term may be positive (incentive) or negative (penalty), depending on the level of compliance with the quality targets established by the regulator,

P_{k-1}^i is the change in the allowed remuneration for company i depending on the amount of energy losses reported in year $k-1$; this term may also be an incentive or penalty, depending on actual energy losses compared to energy loss targets set by the regulator and

IA_k is the adjustment factor in year k, calculated as a weighted average of the retail price index (the acronym used in Spain is IPC) and the industrial price index, *IPRI*, according to the following expression:

$$\text{IA}_k = 0.2 \cdot (\text{IPC}_{k-1} - x) + 0.8 \cdot (\text{IPRI}_{k-1} - y) \tag{5.7}$$

where the efficiency or productivity factors x and y were adjusted to 80 and 40 basis points, respectively, for the period 2009–2012.

Two essential tools are used by the regulator to implement this revenue cap formula: *regulatory accounting* and *reference network models.*

Regulatory cost accounting for facilities of over 1 kV, together with an equipment inventory, helps determine the rate base and its development over time, a key issue in determining the return on capital and depreciation costs.

Reference network models calculate a "benchmark" for each service area and determine the efficient costs for the respective distribution business. Turvey [44] provides an insightful discussion on how to incorporate network models in

distribution regulation. These large-scale distribution planning models build an optimum grid, minimising investment and operating costs while meeting appropriate reliability and quality of supply standards and factoring in such considerations as the location of consumer demand and distributed generators in each service area. The output from this type of model is used by the regulator to calculate operation and maintenance costs when computing the base remuneration, and to calculate incremental distribution costs year by year due to changes in demand and new customer connections. With these models inter-company efficiency can be compared while making allowance for the peculiarities of the different service areas [32, 39].

Regarding quality of service regulation, in Spain, Royal Decree 1955/2000 set benchmark quality levels for the duration and number of long interruptions reported in distribution grids, distinguishing among urban (towns and cities with over 20,000 points of supply and provincial capitals), semi-urban (towns with from 2,000 to 20,000 points of supply), concentrated rural (towns with from 200 to 2,000 points of supply) and scattered rural (towns with fewer than 200 points of supply, and points of supply located outside the town centre) areas. The reliability or continuity of supply indicators measured and monitored area by area are, among others: installed capacity equivalent interruption time (Spanish acronym TIEPI) and installed capacity equivalent number of interruptions (NIEPI). When results dip below the benchmark values, distributors are required to implement improvement plans [3].

Under Royal Decree 222/2008, explicit incentives (or penalties) as a percentage of total remuneration are established for distributors whose performance is better (worse) than the specified TIEPI and NIEPI reference values [4]. In addition, the number of interruptions affecting individual users, as well as their total cumulative duration per year, are likewise monitored. For instance, the total cumulative interruption time affecting a consumer connected to the low-voltage grid in an urban, semi-urban, concentrated rural or scattered rural area may not exceed 5, 9, 14 or 19 h, respectively. When this limit is exceeded, the company must compensate the customer by paying an amount equivalent to five times the price of the *energy not supplied*. Moreover, quality aspects associated with voltage disturbances and customer services are also regulated. To ensure that companies respond to customer requests and complaints within certain deadlines, customer support standards are established [3].

5.6 Technical and Commercial Losses

Technical losses are defined as energy losses in grids as a result of the operation of their electrical components, primarily lines and transformers. Part of these losses, independent of the power flowing across the grid, are essentially fixed magnetisation losses in transformers. The remainder, which are proportional to the square of the power flow, i.e., the current circulating in transformers and lines, are known as ohmic losses. *Commercial* or *non-technical losses*, by contrast, are due to

electricity theft or consumer payment default. The responsibility for technical and commercial losses is incumbent upon the distributor, which must hold them within acceptable bounds, as defined by the regulator. With the unbundling of distribution and retail, a sound regulation should make suppliers responsible for commercial losses because they are in charge of billing to end customers, while distributors should be responsible for technical losses associated with grid operation and investment.

Strictly speaking, losses correspond to generation rather than to distribution, since the energy lost and not billed is produced by generators. In the final analysis, generators are the ones to incur the cost of producing the lost energy, although this cost is ultimately passed on to consumers. Nonetheless, grid management by distributors is obviously instrumental in controlling the level of losses and, therefore, enhancing system efficiency as a whole, and the corresponding incentive mechanisms, should be created for this purpose. Distributors may design networks and invest in facilities with lower loss levels or design lines with an optimal length and cross-section to reduce losses to efficient values, in accordance with power flow forecasts, or undertake action to reduce theft in their networks.

In conclusion, distributor remuneration formulas usually include a system of economic incentives to lower losses, with bonuses when they do decline and penalties if loss levels exceed the established target values.

Such an incentive and penalty mechanism may be designed along lines similar to the scheme discussed above for continuity of supply. For instance, the distributor's base remuneration should be associated with certain reference loss levels in each type of distribution area. The company should pay or be paid for differences between the cost of energy losses actually incurred and the reference or benchmark value. If the losses actually incurred are smaller than the benchmark value, the distributor earns extra revenue, but is fined if they are larger. It should be noted that, under this incentive scheme, consumers only pay for reference or benchmark losses, generators receive the payment for actual losses and the distributor pays or is credited for the difference between the two.

Benchmark loss values must be periodically updated by the regulator, at the end of each regulatory period, for instance. As benchmark losses are gradually reduced, consumers benefit from the distributors' efficiency gains.[3]

As stated above in Sect. 5.5.1, in the United Kingdom, the revenue cap formula itself includes an explicit incentive to reduce distributor losses at a fixed price established by the regulator. In Norway, on the contrary, the revenue cap formula acknowledges the losses actually incurred by the distributor, but includes them in the benchmarking for the calculation of efficiency factors X. In Spain, a market

[3] The reader is invited to design a regulatory instrument that incentivizes distribution companies to optimise losses, while sharing the obtained benefits with the consumers. This regulatory instrument should be similar to the one described in Fig. 5.5. It is possible to do better than the incentive scheme described above.

procedure for settling energy purchases was formerly in place whereby end customers and suppliers paid the distributor for the energy consumed, multiplied by the benchmark loss coefficient established by the regulator for each voltage level. The distributor, in turn, purchased energy entering its grid at the market price. Therefore, the distributor was paying or being credited the difference between actual and reference losses. Royal Decree 222/2008, currently in force, sets reference values for losses for each distribution company and establishes incentives to reduce actual losses below these values.

The standard loss coefficients in place in Spain for energy consumptions at different voltage levels are shown in the Table 5.8.

A position paper issued by the European Regulators Group for Electricity and Gas (ERGEG) compares the mechanisms implemented in various European countries to incentivise loss reduction by network operators [10].

Lastly, attention should be drawn to the specific issue of commercial losses. In developing countries and developed countries with pockets of social marginalisation, commercial losses due to theft may be a serious problem. This problem may become particularly significant from a financial standpoint when it has an impact on the distribution company's profit and loss account instead of being just another State subsidy. This was the situation encountered, for example, by the new owners of distribution companies privatised in Greater Buenos Aires, Argentina, in the 1990s. The companies concerned had to book non-technical losses of over 20 % against their income. Thanks to cooperation with the regulator and the city authorities and a progressive action plan on the part of the distributors, the problem was largely solved in a matter of a few years' time. Company action included improvement of the quality of service and installing legal individual meters for all clandestine users, at no cost initially and then at a subsidised rate. The required investments were partly covered by a city tax exemption. The companies have managed to reduce their commercial losses to acceptable levels, with certain small groups of customers receiving public subsidies. Deterrents have also been found to prevent theft from recurring, including extra-high poles, coaxial service connections or remote supply interruption systems.

Figure 5.11 shows how energy losses have been reduced in Peru since the restructuring and privatisation process started in 1994.

Table 5.8 Benchmark energy loss coefficients in Spain

Voltage level	K (%)
LV (Low voltage)	13.81
MV (U < 36 kV)	6.00
HV (U < 72.5 kV)	4.00
HV (U < 145 kV)	3.00
HV (U > 145 kV)	1.62

Source [5]

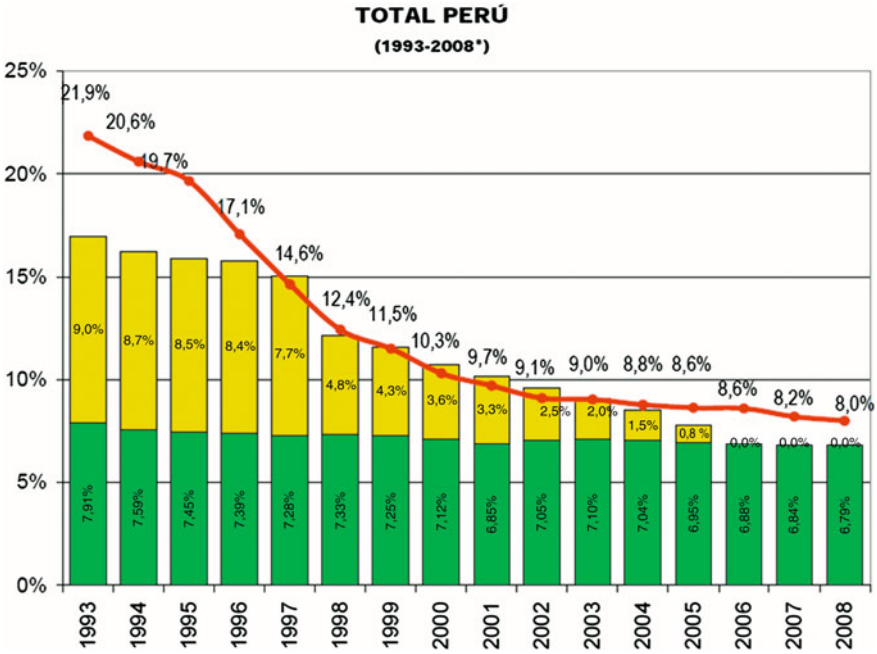


Fig. 5.11 Energy losses as a percent of total energy production in Peru. *Upper line* is for actual losses, *upper rectangle* for allowed commercial losses, *lower rectangle* for technical losses. *Source* [25]

5.7 Impact of Distributed Generation on Distribution Networks

The reasons for the recent development of *distributed generation* must be sought in energy-planning policies essentially geared to reducing the environmental impact of electricity generation or dependence on imported fossil fuel, or increasing the energy efficiency of combined heat and power production processes. So far, for the most part distributed generation has been subsidised. However, some of the involved technologies keep reducing their costs and might become competitive with conventional technologies in the near future. This may increase their rate of penetration even more.

There is no single generally accepted definition of *distributed generation* (DG). Overall, it comprises small-scale generators (in Spain, less than 50 MW)—as opposed to traditional large power plants—directly connected to the distribution grid or to the consumer's own facilities.

Based on the primary energy used, distributed generation can be divided into:

- Non-renewable DG
  - internal combustion
  - gas turbines (mini- and micro-turbines)

  - fuel cells.

- Renewable DG
  - solar (photovoltaic and thermal)
  - wind (wind and mini-wind)
  - mini-hydroelectric
  - biomass.

As noted above, renewable generation technology is closely associated with governmental promotion and incentive policies, because of the low environmental impact it involves and the present need for regulatory support to be attractive for private investors, and become competitive with other technologies. In Spain, for instance, wind generation has developed at a brisk pace over the last few years, with over 20 GW of installed capacity at the end of 2010. The energy produced from renewable sources and combined heat and power (CHP) units under the so-called "special production regime" in Spain accounted for 33 % of all electric power consumed in 2010. In Spain, 54 % of the total installed capacity of wind farms, because of their size and location, are connected directly to the transmission grid. Electricity production at that scale is not, obviously, regarded as distributed generation. However, other technologies as photovoltaic and CHP are mostly, 98 % and 88 %, respectively, connected to distribution grids [7].

The development of non-renewable electric power generation, in turn, is closely related to the use of co-generation (CHP) in processes calling for both heat and power. Such arrangements involve a significant increase in energy efficiency and, therefore, are also the object of government incentives. Other non-renewable technologies, such as gas micro-turbines and fuel cells, are still in the research and development stage, in which the basic challenge is to lower costs and increase efficiency. Future electricity grid design, planning and development may be directly affected by these very promising, highly energy efficient, modular technologies, which can also adapt to each consumer's specific needs. Scheepers et al. [42] provide a good review of DG deployment and network integration in Europe, while [23] is mostly focused on the US.

The items listed below are some of the more prominent ways that DG would affect distribution grids and the business conducted by distribution companies, the subject of the present chapter.

- *Impact on investment in the grid.* DG connection to the grid also entails investment in new network facilities. It may also involve reinforcing existing facilities located upstream of the connection point so they can accommodate the power generated and cope with the potential reverse flows. Conversely, if generation is located near consumption, the net demand for energy will be lower at certain times of day, reducing the load on existing facilities and the need to reinforce them in the future to meet the needs of natural growth. Therefore, DG may have both beneficial and adverse effects on grid investment.
- *Impact on grid operation.* Distribution grids are essentially designed to meet the demand for power by using a one-directional, radial flow from substations to

loads. The connection and operation of DG near consumers implies that the natural direction of the power flow may be inverted, i.e., from demand upstream to the substation. The immediate effect of this new situation is to add to the complexity of determining possible overloads on conductors as well as holding grid voltage within allowable margins. A second consequence is the greater intricacy of the logic underlying automatic protection systems, since new directional criteria are required to identify and locate problems. Thirdly, as a rule, DG connection raises short-circuit power in the grid, creating the need to revise the design criteria for grid protection equipment (switches and circuit breakers) and substation busbars.

- *Impact on quality of supply.* DG can have an impact on continuity of supply and voltage quality that should be appropriately assessed. In terms of continuity of supply, DG failure rates may affect other consumers connected to the same grid and, therefore, increase the distributor's unavailability. Conversely, if DG could be controlled by the distributor, it could be used to provide support power for nearby consumers in the event of a grid failure, either in *islanded operation* or through an *islanded emergency power system.* In this case, the distributor's continuity of supply indices would logically be improved. The effect of DG on voltage quality, in turn, varies widely in such areas as voltage oscillation or disturbance emission (flicker or harmonics) and essentially depends on the technology and grid interface used. Wind power produced by asynchronous generators, for instance, may cause flicker due to the oscillations produced by variations in wind speed/direction, or voltage problems associated with the inability to control the reactive power consumed. The consequences of both types of incidents are less severe if asynchronous generators with dual-feed rotors are used. Photovoltaic cells, in turn, pose other kinds of considerations. Power converters are required to change the power generated by such facilities from direct to alternating current for delivery to the grid. These converters need to be equipped with filters to prevent the emission of voltage wave harmonics that would otherwise find their way into the distribution network.
- *Impact on energy losses in the grid.* As noted earlier, the distributor must be given incentives to maintain the energy loss level in the grid within certain limits established by the regulator. The production of energy by distributed generators may induce variations in power flows in the grid and, therefore, the associated losses, which are dependent upon the square of the power flow. In general, when the DG's level of penetration in the grid is not very high and generation is located close to consumption, the power flow demanded from upstream substations declines. This has a beneficial effect on the distributor's energy losses. By contrast, if generation reaches very high levels or is connected to dedicated lines isolated from consumers, the power flow is inverted, with the appearance of new flows or larger flows than prior to DG connection. In this case, the distributor's losses rise. Methods to evaluate these situations are necessary, therefore, and the associated economic impact for both the distributor and the DG must be acknowledged [24].

- *Sunk or stranded costs for existing facilities*. As discussed in preceding sections, distributor remuneration must be duly regulated. Incentive regulation, namely, price and revenue caps, was analysed above. Generally speaking, the distributor's remuneration depends on the number of kWh distributed, directly in the case of price caps, or through a proportionality coefficient, as appropriate, under revenue cap arrangements. When DG is embedded in the customer premises and net metering is applied this may have an adverse effect on the distributor's revenues in the short- and medium-term. In next price control reviews, when setting the base remuneration, this effect should be taken into account and the facilities already installed by the company to supply demand, as was forecasted without DG, must continue to be remunerated and not affected by the decline in net circulating current. In the long term, this effect would yield grid investment needs much more closely in line with the new technology as well as a new grid design in which DG would be more fully integrated. Reference network models like the ones discussed in the section on Spanish regulation can help regulators with this task.
- *Distributor operation and maintenance personnel safety*. This last item is by no means the least significant. On the contrary, it is of utmost importance in distribution company dispatching centres, where all grid maintenance and repair activities are planned and where the first priority is to ensure that the facilities to be worked on are duly disconnected and grounded. The existence of a generator on the grid that is not directly operated by the distributor or under its control is a safety concern. For these reasons, generators are required to have an automatic protection system that disconnects the facility from the grid when upstream voltage is lacking. Much still has to be done to develop an operating and control system that fully guarantees the safety of the people working on the grid at any given time while not jeopardising the generator's interests. IEEE Standard 1547 provides a good description of the technical issues relating to the interconnection of distributed energy resources to electricity networks [16].

By way of summary, DG connection to distribution grids entails both advantages and drawbacks in nearly all the areas discussed above. Further work is needed in matters such as analytical tools, planning, operation and control procedures and regulatory design to resolve the new problems posed by distributed generation.

From the standpoint of energy supply to regulated customers and the mandatory purchase of power generated by renewable DG or CHP producers connected to distribution grids, two other issues concern distributors.

- One problem is the effect of the volume of power generated by intermittent and particularly variable sources on the distributor's aggregate energy demand forecast, when the distribution and retailing functions are not unbundled. These sources can be ranked in descending order of variability and uncertainty as follows: wind, photovoltaic and CHP associated with industrial processes. The implication is that this generation may trigger deviations from the net power

consumption forecasts on which the distributor bases its electricity market purchases, causing it to incur the concomitant costs.
- The second issue is compliance with the requirements and obligations imposed on distributors with regard to the provision of and demand for ancillary services in the electricity market. In Spain, for instance, the system operator requires distributors to maintain certain power factors at consumption points on the transmission grid; in other words, the consumption of reactive power, expressed in terms of active power, must be held within a given range of values. DG connection usually decreases the net active power consumed, as noted earlier, and barely affects or even increases the net reactive power consumed. The immediate consequence is an increase in the distributor's likelihood of failure to comply with power factor requirements, thus incurring the resulting fine. Another issue is the need to increase system regulation and reserve capacity due to the uncertainty introduced by renewable (wind and photovoltaic) generation. Because of the intermittent nature of this type of generation and possible unforeseen variations in power over a short period of time, the system operator needs a larger margin of spinning and regulation reserves. This raises system operation costs because such additional reserves must be paid for by one of the parties concerned, logically the DG responsible for the costs and not the distributor.

In short, a series of technical, economic and regulatory barriers will have to be surmounted in the near future to facilitate greater DG integration in electric power grids and markets. This is indispensable if the policy objectives on renewable energy and energy efficiency set out by the European Union and most countries in the world are to be achieved by the end of this decade and beyond, as planned. Some of the issues that need to be addressed and solved are:

- Legislation regulating DG connection to grids, in accordance with objective and transparent technical procedures.
- Evaluation of the actual and future impact of DG on grids, identifying the benefits for or extra costs to the system and obliging DG to share in such benefits or costs.
- Changeover to active management of distribution grid operation (flow and voltage control), establishing the requirements for controlled DG dispatching within the capabilities of each DG technology.
- Provision for the effect of DG on investment in the scheme for remunerating distributors.
- Regulation of DG access to the generation market and the provision of other ancillary services.
- Guarantee of free DG access to distribution grids, with due regulation of the interaction between the generation and grid businesses modelled on the arrangements in place for large-scale generators and transmission grids, and creation of independent distribution system operators as recommended by the European Electricity Directive [9].

Some of the preceding issues relating to DG integration and network regulation are discussed from a European perspective in Cossent et al. [6].

5.8 Rural Electrification

Access to modern forms of energy is a key element for the development of human societies. United Nations Energy [45] argues how this access is key for achieving the Millennium Development Goals. International Energy Agency [18] highlights electricity as the most critical energy carrier for development and evaluates the costs and impacts of achieving universal access to electricity and other modern forms of energy by 2030. In 2008 electrification rates (percentage of households with access to electricity according to the World Bank's definition) amounted to 99.8 % in transition and OECD countries, but to only 72 % in developing countries. Among these countries, low electrification rates are concentrated in rural areas (electrification rate of 58.4 %, versus 90 % in urban areas), where 55 % of the population lives in the less developed regions. In addition, in absence of vigorous policies, in 2030 the same number of 1.3 billion people in the world will still live without access to electricity.

Therefore, the need to foster electricity access in rural areas in developing countries seems urgent. However, this task is very complex; rural areas in developing countries are usually very poor and their inhabitants' per capita energy consumption is (as a cause and as a consequence) very low. Thus, the benefits of electrifying these areas would be low and risky for private companies. In addition, households tend to be dispersed over remote and inaccessible areas, and the low consumption levels do not allow for taking advantage of the economies of scale normally present in the electricity sector. Thus, electrification costs are very high. This combination makes rural electrification activities (network expansion and operation, as well as possible investments in new generation capacity or isolated systems) very unattractive for private investors. This is one of the major underlying causes of low electrification levels in rural areas in developing countries.

On the other hand, access to modern forms of energy is in many countries a constitutional right, which makes government the subsidiary authority in charge of making sure that this right is fulfilled. This, added to the above-mentioned advantages for economic and social development, has led many governments to propose large investments in rural electrification, although it is difficult for them to cover the usually high costs. Therefore, it is necessary to involve private initiatives in the process; not only large multinational energy companies, but also small private arrangements such as cooperatives, communities or other actors.

The diverse sustainability aspects of these investments are a key aspect to be considered. Rural electrification programs should be based on a solid economic

regime that provides economic sustainability for the installations, with adequate maintenance and development of local capability being key elements. Environmental concerns should also be taken into account, as well as ensuring, through participatory instruments, social sustainability. Universal access strategies must have priority in the international agenda of energy policy, which is also compelled by the fight against climate change and might yield opportunities as well as contradictory measures. International Energy Agency [18] contains detailed discussions and data on these issues.

With a view to providing the population in developing countries with this essential service and ensuring political determination and support, an appropriate regulatory design is required so that investment in such infrastructure can be made in keeping with a model of adequate long-term viability and sustainability [12, 19, 36].

In Latin America, for instance, a number of very successful electrification experiences, such as in Chile and Guatemala, have highlighted the importance of a specific regulatory design. The Table 5.9 compares the countries in the Latin American region in terms of the levels of electrification achieved in rural and

Table 5.9 Electrification coefficients and programmes in Latin America

Country	Electrification coefficients (%)	Programmes and legislation to develop rural electrification
Argentina	90 (total) 70 (rural)	PERMER Project (National Government and Secretariat of Energy)
Bolivia	65 (total) 87 (urban) 33 (rural)	Electricity Act (State) ERTIC Programme (Department of Electricity)
Brazil	99 (urban) 75 (rural)	Act No. 10762 (ANEEL [National Electrical Energy Agency] and utilities)
Chile	93.5 (rural)	Government programmes In 15 years, the coefficient rose from 53 % to 93.5 %
Colombia	92 (total)	Act No. 188 National Royalties Fund
Ecuador	Not available	FERUM Fund
El Salvador	97 (urban) 75 (rural)	PROERES Programme (FINET funds, utilities and municipalities)
Guatemala	85 (total)	PER Plan (INDE [National Electrification Institute]–Government) EURO-SOLAR Programme (EU and Government)
Panama	98 (urban) 58 (rural)	Electricity Act (Rural Electrification Office—utilities)
Peru	80 (total) 30 (rural)	Government policies with regional and local governments (Act No. 28749)

Source [25]

urban areas. The electrification coefficient is expressed as the percent of the population located in areas with electricity compared to the total population. The specific programmes and legislation developed to increase the level of electrification in each country are also included in the table.

In general, market-oriented regulatory reforms, with little State involvement and no specific programme to supply energy to communities with no access to this service, have not improved the situation of such communities. The key to success in rural electrification depends on a number of factors.

(i) A firm political commitment to this task must be made by all levels of government: national, regional and local. The starting point should be the acknowledgment that the access to this modern form of energy must be a human right.
(ii) Funds and subsidies must be allocated exclusively ("ring fenced") for this purpose. The level of investment needed is usually much higher than standard network expansion development, and is unaffordable for the target population. In this respect, the technical and economic requirements to electrify remote and isolated rural areas may differ substantially from the requirements for expanding the current interconnected system to closer areas. Special attention must be paid to isolated rural communities that are too far away from the main grid and with loads so disperse that it is not technically or economically viable to connect to the existing distribution grid. Real Academia de Ingeniería de España [34] examines in detail the appropriate technological approaches to provide essential services (water, energy, communication) to isolated rural communities in developing countries.
(iii) The programme must focus on the lowest-income population, according to a sustainable system design valid not only for the initial installation but also for ongoing operation and maintenance, to include the specific needs, opinions and support of the population involved. In general, electrification should be accompanied by social and economic actions to develop these poor areas.
(iv) A clear legal framework must be implemented for all parties involved, including distributors and other service providers in remote areas, including, where appropriate, the explicit obligation to provide service in their concession areas and with specifically designed electricity rates and subsidies to ensure long-term sustainability. Private investment will be needed in most cases, given the volume of the investment that is needed to attain universal access. Therefore, the major regulatory challenge is to provide an attractive business model under the circumstances explained above.

Box 5.1. Solar PV for off-grid applications in developing countries. Regulatory issues (The case of Guatemala)[41]

Guatemala is the most populated country in Central America and at the same time the largest economy in the area. Nonetheless 57 % of the population lives in poverty, 21.7 % in extreme poverty. Seventy-four percent is concentrated in rural areas and 76 % is indigenous population. The electrification rate rose from 37 % in 1990 to 84 % in 2002. The major part of the electrification has been achieved via extension of the national electricity grid. Rural and mountainous areas have been left apart and are nowadays isolated. These areas are at the same time those with the highest poverty indices. The characterisation of demand for housing, schools or medical centres was taken from Rafael Landívar University, CIEMAT and own estimates. The proposed regulatory framework would provide electricity to 700,000 people (6 % of the Guatemalan population). This corresponds to the electrification of 137,470 households in 3,722 communities.

The project characterised the isolated areas on one hand because of their high geographic dispersion, being far from densely populated nucleus, having weak transport and communication infrastructures, in many cases complex orography, and in high value natural environments, and on the other hand because of their low consumption, low income and low growth perspectives.

These characteristics led to the development of a decentralised model for electricity supply, not only referring to generation but also for management, funding, control, etcetera.

REGEZRA provided the Guatemalan Electrical Authorities with the legal, technical, administrative, financial and business framework needed to supply this population with a basic electricity service for the coming years. Based on the principles enumerated in this section, the framework for off-grid rural electrification is a service-based model, rather than an investment-based one, described by the following features.

[4] Between 2007 and 2009 the Fundación Energía Sin Fronteras (Energy Without Borders) together with Comillas University—Instituto de Investigación Tecnológica IIT (Spain), the Guatemalan Instituto Nacional de Electricidad and Comisión Nacional de Energía Eléctrica, the Guatemalan Universities of San Carlos, Rafael Landívar and Universidad del Valle, the Spanish Comisión Nacional de la Energía along with Mercados EMI, Fundación Solar, NRECA Internacional and Universidad Politécnica de Madrid developed the REGEZRA Project (Regulation for the Electrification of Isolated Rural Zones in Guatemala), funded by the Agencia Española de Cooperación Internacional. The project developed a specific regulatory framework proposal for electricity access in these zones of Guatemala, complementary and coherent with the existing regulation for electrification in rural areas through network extension. This contribution has been provided by Andrés González, researcher at Comillas University and one of the main participants in the REGEZRA project.

Promotion of private initiative and competition

This would be achieved by a competitive tendering process, by which private investors would compete for the subsidies available for the electrification of the rural areas previously identified in a National Rural Electrification Plan. These subsidies, which should cover the gap between the costs incurred by the investor and the income received from consumers, would be released by the public administration according to the correct installation and operation of the equipment.
Under this scheme, a potential supplier must bid the minimum subsidy to be received for each connection point.

Type of developers

It is recommended that the type of developers who should carry out the electrification projects would be local ventures and communities, who should be incentivised to participate in the tenders and in the maintenance of the installations, given their crucial role in the sustainability of the project.

Financial regime

Given that income will usually be lower than costs, subsidies will be necessary. These subsidies may come from different sources: other energy consumers, national budgets, advanced financing mechanisms like the Kyoto Protocol's Clean Development Mechanisms, or national, regional or international development agencies. However, in order to guarantee their availability, and also to decouple funding agencies or sources from investors, REGEZRA proposed the creation of a dedicated fund, which on the one hand aggregates the different sources, and on the other hand, guarantees its exclusive use for rural electrification.

In order to achieve the sustainability of the projects, subsidies must be released upon the provision of the service, and not associated with the investments. Therefore, subsidies will be paid to investors during the lifetime of the project, to deter "build-and-run" behaviours. This should be governed by a contract signed between the electricity provider and the public administration managing the subsidies. The disadvantage of this proposal is that, by deferring the grant, the contractor will need more funding, which means that only those agents who have borrowing capacity could engage in this type of competition. This aspect should, therefore, be carefully planned. The payment of the subsidies must be subject to the verification of the continuity and quality of the electricity service.

Electricity rates

Electricity rates must be calculated in reference to the existing social tariff for grid customers, and it is recommended that they should not be above it.

However, they must cover at least maintenance costs to ensure the financial viability of the project. Different rates may be set depending on the quality of service.

Ownership of the equipment

Since this is a service-based model, the achievement of rural electrification should be measured in terms of the quality of the electricity service provided, rather than on the number of installations. This results in that the ownership of the generation equipment belongs to the supplier, rather than to the final users. This in turn places the responsibility for maintenance on the suppliers, which usually have expert personnel, instead of on the final users.

Other elements promoting sustainability

The following elements are introduced to ensure the sustainability of the project, in addition to those previously described:

- The temporal scope for the regulation and the financial regime must always go beyond the investment phase.
- The costs to be recovered must include not only investment ones, but also replacement, operation and maintenance costs during the lifetime of the installation.
- Making users pay involves them in the scheme, makes them conscious of the cost of electricity, and makes them require a certain quality for it.
- Local administrations become the monitoring agents for the technical and economic terms of the electricity service, thus involving local communities and decentralising the administrative process.
- A fraction of the dedicated fund must be devoted to training and education for electricity users.

5.9 Summary

This chapter has analysed the chief technical and economic characteristics of the electricity distribution business and the mechanisms most commonly used to regulate distributors as network-infrastructure monopolies.

The essential ideas set out in this chapter are summarised below.

- Electricity distribution is a regulated network business whose main purpose is to invest in and maintain the grids that carry electric energy from the transmission grid to end consumers. Retailing or commercialisation of electricity to end consumers is a different business, although in most non-European Union countries, both activities are performed by the same company. In the European

Union, the Electricity Directive calls for the full unbundling of distribution and retail in all Member States [9].

- Electricity distribution is a natural monopoly, since the competitive installation and operation of power lines in any given geographic area would be highly inefficient.
- Since electricity is considered an essential service, distributors are normally obliged to connect to the grid any user requesting supply in their service area.
- Distributors must ensure open third-party access to consumers, generators and other grids to enable all players to freely engage in power transactions. Conflicts of interest may arise in this respect—and should be supervised by the regulator—when a distributor with holdings in a supplier attempts to favour its investee or obstruct access or discriminate against other competitor suppliers.
- Distributors receive yearly regulated revenues according to some ratemaking scheme (cost-of-service, price or revenue cap). Such remuneration must be sufficient to cover their operating expenses, infrastructure depreciation and a rate of return on the capital invested, all subject to certain guidelines or technical and economic efficiency targets established by the regulator.
- Distributors' annual remuneration inflows as allowed by the regulator come in the form of network distribution charges, one of the items comprising access tariffs, or integral tariffs for still-regulated customers.
- Incentive regulation (price or revenue caps) is gradually replacing traditional cost-of-service regulation in power systems where the electricity industry has been restructured, and distribution, or distribution plus retailing, has been unbundled.
- Incentives to cut costs may lead to a progressive decline in the quality of supply delivered by distributors, as a result of corner-cutting. Therefore, regulators should impose quality objectives that distributors must meet under performance-based ratemaking (PBR) arrangements or explicit reward/penalty schemes. Improvements on established targets are rewarded, and failure to reach them is penalised.
- Quality targets must be established on the basis of area—fairly distinguishing among rural, semi-rural and urban areas—and individual supply point, so that each and every consumer receives a product that meets at least certain minimum quality standards that are adapted to his/her characteristics.
- International experience shows that distributors are being increasingly penalised when they fail to meet quality objectives.
- Benchmark or reference energy losses should also be established for distributors. When their actual losses are below this benchmark, they are rewarded, and when the level is exceeded, they are financially penalised.
- Governmental promotion of renewable energy and CHP is leading to the increasing development of distributed generation, i.e., small-scale generators directly connected to the distribution grid. The impact of this generation on grids and the distribution business must be carefully evaluated. All the respective technical and economic issues must be solved and suitably regulated. Such regulation should guarantee the distributor's neutral role and identify the

benefits and drawbacks for all the players involved, allocating responsibilities accordingly. An open topic in distribution regulation is how to encourage innovation in network design and operation, in anticipation of the incoming challenges of distributed generation, increased demand response, distributed storage or electric vehicles.

- In developing countries, the issue of rural electrification is a key aspect of distribution regulation. Together with political determination and support, a specific regulatory design involving distributors is needed to ensure the long-term sustainability of newly electrified areas.

References

1. Agrell P, Bogetoft P, Tind J (2005) DEA and dynamic yardstick competition in Scandinavian electricity distribution. J Prod Anal 23:173–201
2. Ajodhia VS (2005) Regulating beyond price: integrated price-quality regulation for electricity distribution networks. Ph.D. Thesis. Delft University of Technology, Delft
3. BOE (Boletín Oficial del Estado [Official State Gazette]) (2000) Real Decreto 1955/2000, de 1 de Diciembre, por el que se regulan las actividades de transporte, distribución, comercialización y suministro y los procedimientos de autorización de instalaciones de energía eléctrica (Royal decree on the transmission, distribution, retail sale and supply of electric power). Madrid, Spain
4. BOE (Boletin Oficial del Estado) (2008) Real Decreto 222/2008, de 15 de Febrero, por el que se establece el régimen retributivo de la actividad de distribución de energía eléctrica. (Royal decree on the remuneration of the electricity distribution activity). Madrid, Spain
5. BOE (Boletin Oficial del Estado) (2009) Orden ITC/3519/2009, de 28 de Diciembre, por la que se revisan los peajes de acceso a partir de 1 de Enero de 2010. (Ministerial Order on electricity network access rates as of 1 January 2010). Madrid, Spain
6. Cossent R, Gómez T, Frías P (2009) Towards a future with large penetration of distributed generation: Is the current regulation of electricity distribution ready? Regulatory recommendations under a European perspective. Energy Policy 37:1145–1155
7. Cossent R, Gómez T, Olmos L (2011) Large-scale integration of renewable and distributed generation of electricity in Spain: current situation and future needs. Energy Policy 39(12):8078–8087
8. European Committee for Electrotechnical Standardisation (CENELEC) (1994) Voltage characteristics of electricity supplied by public distribution systems. European Norm EN 50160, Nov
9. European Communities (EC) (2009) Directive 2009/72/EC concerning common rules for the internal market in electricity and repealing directive 2003/54/EC
10. European Regulators Group for Electricity and Gas (ERGEG) (2009) Treatment of electricity losses by network operators. ERGEG position paper: conclusions paper. Ref: E08-ENM-04-03c. 19 Feb. Brussels (Belgium)
11. Fumagalli E, Lo Schiavo L, Delestre F (2007) Service quality regulation in electricity distribution and retail. Springer-Verlag, Berlin
12. Global Network on Energy for Sustainable Development (GNESD) (2004) Energy access theme results: assessment of energy reforms in Latin America and the Caribbean. April version. Available on-line at http://www.gnesd.org/
13. Gomez T, Rivier J (2000) Distribution and power quality regulation under electricity competition. A comparative study. Proceedings of the IEEE ninth international conference on harmonics and quality of power, vol 2, pp 462–468

14. Grasto K (1997) Incentive-based regulation of electricity monopolies in Norway. Background, principles and directives, implementation and control system. Publication 23/1997, Norwegian Water Resources and Energy Administration, POB 5091, 0301 Oslo, Norway
15. Institute of Electrical and Electronics Engineers (IEEE) (2001) IEEE guide for electric power distribution reliability indices, IEEE Std 1366, 2001 edn
16. Institute of Electrical and Electronics Engineers (IEEE) (2003) IEEE standard for distributed resources interconnected with electric power systems, IEEE Std 1547TM
17. International Electrotechnical Commission (IEC) (1990) Standard IEC 61000-2-1: electromagnetic compatibility. Part 2: Environment, Section 1: Description of the environment
18. International Energy Agency (IEA) (2010 and 2011) World energy outlook. Available on line at http://www.iea.org/weo/electricity.asp
19. International Energy Agency (IEA) (2011) World energy outlook 2011. Available on-line at http://www.iea.org/
20. Jamasb T, Pollitt M (2003) International benchmarking and regulation: an application to European electricity distribution utilities. Energy Policy 31:1609–1622
21. Jamasb T, Pollitt M (2007) Incentive regulation of electricity distribution networks: lessons of experience from Britain. Energy Policy 35(12):6163–6187
22. Joskow PL (2006) Incentive regulation in theory and practice: electricity distribution and transmission networks. Cambridge Working Papers in Economics CWPE0607, Electricity Policy Research Group Working Paper EPRG 0511
23. Massachusetts Institute of Technology (MIT) (2011) The future of the electric grid. http://web.mit.edu/mitei/research/studies/the-electric-grid-2011.shtml
24. Mendez VH, Rivier J, Gomez T (2006) Assessment of energy distribution losses for increasing penetration of distributed generation. IEEE Trans Power Syst 21:533–540
25. Mercados Energéticos Consultores—Instituto de Investigación Tecnológica (ME-IIT) (2009) Análisis del Marco Regulatorio para la determinación del valor agregado de la distribución en Perú. (Analysis of the regulatory framework for determining the aggregate value of electricity distribution in Peru). Prepared for Organismo Supervisor de la Inversión en Energía y Minería de Perú (OSINERGMIN). Lima
26. Myers R, Strain LL (2000) Electric and gas utility performance based ratemaking mechanisms. Prepared for the California public utilities commission. San Francisco (updated Sep 2000)
27. National Audit Office (NAO) (2002) Pipes and Wires. Available on-line at www.nao.gov.uk
28. Norwegian Water Resources and Energy Directorate (NVE) (2008) Annual report on regulation and the electricity market in Norway to the European commission. June
29. Norwegian Water Resources and Energy Directorate (NVE) (2009) Annual report on regulation and the electricity market in Norway to the European commission. June
30. Office of the Gas and Electricity Markets (OFGEM) (2004) Electricity distribution price control review. Final proposals, office of gas and electricity markets. Available on-line at www.ofgem.gov.uk
31. Office of the Gas and Electricity Markets (OFGEM) (2010) Regulating energy networks for the future: RPI-X@20. Emerging thinking. Available on-line at http://www.ofgem.gov.uk
32. Peco J (2004) A reference network model: the PECO model. Working paper IIT-04-029. Instituto de Investigación Tecnológica, Universidad Pontificia Comillas, Madrid, Spain. June. Available on-line at http://www.iit.upco.es/docs/IIT-04-029A.pdf
33. Petrov K, Nunes N (2009) Analysis of insufficient regulatory incentives for investments into electric networks. An update. Final report. KEMA consulting GmbH submitted to European Copper Institute. Available on-line at http://www.leonardo-energy.org/
34. Real Academia de Ingeniería de España (RAI) (2011) Tecnologías para el Desarrollo Humano de las Comunidades Rurales Aisladas. Pérez-Arriaga I, Moreno A (eds). http://www.raing.es (in Spanish)

35. Regulatory Assistance Project (RAP) (2000) Performance-based regulation for distribution utilities. Available on-line at www.raponline.org/Pubs/General
36. Reiche K, Tenenbaum B, Torres C (2006) Electrification and regulation: principles and a model law. Energy and mining sector board discussion paper, paper no. 18, The World Bank, Washington DC. Available on-line at http://siteresources.worldbank.org/
37. Rivier J (1999) Quality of service: investment regulation and optimization. Ph.D. Thesis. Comillas Pontifical University, Madrid, Spain (*in Spanish*)
38. Rivier J, Gomez T (2000) A conceptual framework for power quality regulation. Proceedings of the IEEE ninth international conference on harmonics and quality of power, vol 2, pp 469–474
39. Roman J, Gomez T, Muñoz A, Peco J (1999) Regulation of distribution network business. IEEE Trans Power Delivery 14:662–669
40. Rothwell G, Gomez T (2003) Electricity economics: regulation and deregulation. IEEE-Wiley Press, Piscataway
41. Sappington DEM, Pfeifenberger JP, Hanser P, Basheda, GN (2001) The state of performance-based regulation in the U.S. electric utility industry. Electr J 14(8):71–79
42. Scheepers MJ, Bauknecht D, Jansen JC, de Jode J, Gomez T, Pudijanto D, Ropenus S, Strbac G, (2007) Regulatory improvements for effective integration of distributed generation into electricity distribution networks. DG-GRID Final Report. Energy Research Centre of the Netherlands (ECN). Report ECN-E–07-083. Nov
43. Southern California Edison (SCE) (2010) Regulatory information—SCE tariff books. Preliminary statements. Available on-line at http://www.sce.com/AboutSCE/Regulatory/tariffbooks/
44. Turvey R (2006) On network efficiency comparisons: electricity distribution. Utilities Policy 14:103–113
45. United Nations Energy (2005) The energy challenge for achieving the millennium development goals energy and the millennium development goals. New York
46. Yu W, Jamasb T, Pollitt M (2009) Willingness-to-pay for quality of service: an application to efficiency analysis of the UK electricity distribution utilities. Energy J 30(4):1–48

Chapter 6
Electricity Transmission

Michel Rivier, Ignacio J. Pérez-Arriaga and Luis Olmos

Those who understand transmission regulation have understood power sector regulation.

Ignacio Pérez-Arriaga

In traditional electric utilities, where generation, transmission, distribution, supply and system operation are vertically integrated, the role of transmission tends to go unnoticed. Its remuneration is based on cost of service and its economic impact is limited to a comparatively modest contribution to the total cost of electric power paid by consumers, typically from 5 to 10 % in systems with no major geographical imbalances between generation and demand. Moreover, no access disputes arise because only one player is involved. In the new free market context, on the contrary, transmission has become the meeting point for the various players interacting on the wholesale market. The huge capacity of today's transmission grids enlarges the effective size of markets enormously, facilitating competition. But this calls for more sophisticated rules of the game. The amount to be paid by each player for using the network or the benefit obtained from its use must be determined, and this charge affects each actor's competitive position. Rules of priority must be established with respect to network access when conflicts arise around limited capacity. Effective administrative or market mechanisms must also be established to ensure that transmission network expansion takes place in accordance with system needs, seeking to maximise the aggregated social welfare, in a context where each transmission reinforcement has direct implications on the individual benefits and losses of the market agents.

For most of a country's transport infrastructure (gas, railway, road and electricity networks), it makes no social, environmental or especially economic sense to develop parallel networks that would compete to provide the respective service.

M. Rivier (✉) · I. J. Pérez-Arriaga · L. Olmos
Universidad Pontificia Comillas, Instituto de Investigación Tecnológica,
Alberto Aguilera 25, 28015 Madrid, Spain
e-mail: michel.rivier@iit.upcomillas.es

I. J. Pérez-Arriaga
e-mail: ipa@MIT.EDU

L. Olmos
e-mail: luis.olmos@iit.upcomillas.es

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_6,

In principle, all these activities must be treated as regulated monopolies, although exceptions may exist under special circumstances. Like distribution, the other electricity network service, electric power transmission is typically centrally planned and paid its cost of service, by contrast to other industry businesses that can be liberalised and conducted on competitive markets.

In light of the technical characteristics of the electricity to be transported (see Chap. 1), and in particular the difficulties in storing it, the electric power network is a particularly crucial link in the system. Its viability is instrumental to system security and its technical features are essential to sustaining the system. If a power line becomes congested, it is very difficult to switch transmission to other lines, such as may be done with road traffic, for instance. An outage in any given element of the transmission system may significantly change the power flows in many lines instantaneously, and may affect the secure operation of the system as a whole, if the protection relays that are supposed to isolate the faulty component fail to work as they should.

For the conditions that make transmission a natural monopoly and for its key role as the meeting point for supply and demand, in liberalised electricity markets transmission network access and use must be regulated in a strictly equitable and non-discriminatory manner, and the costs incurred in building, maintaining and operating the grid must be fairly distributed. Furthermore, regulation must guarantee suitable network development to establish the conditions essential to an efficient marketplace by minimising the barriers to system entry for producers and consumers alike. The transmission network is the key enabler of competition in electricity markets.

Regulation of the transmission network should provide answers, based on the above criteria, to the following questions:

- Who reinforces the network when needed?
- Who can connect to the network? And what happens when the network becomes congested?
- How should the network costs be allocated?
- Who pays for the power losses that take place in the network?

The growing interconnectivity of different countries' networks and the incentives that exist in a number of geographic areas leading to the creation of regional markets are generating new challenges for electric power network regulation. The anticipated enormous growth of generation from renewable sources—wind and solar, in particular—frequently located far from load centres and with strongly variable production patterns is pushing the current paradigm of transmission regulation to its limits.

The present chapter addresses these subjects. The first half of the chapter provides the background that is needed to approach the main regulatory topics. Section 6.1 describes the network service from both the technical and economic viewpoints and identifies the role of the grid in new liberalised contexts, along with the basis for its structural organisation. Section 6.2 defines and explains the basic properties of nodal prices.

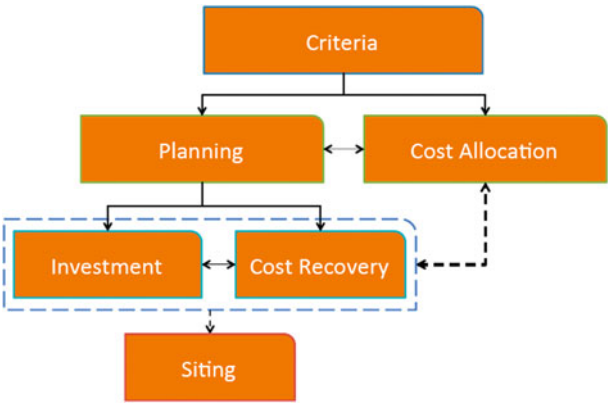


Fig. 6.1 The regulatory framework for electricity transmission

The second part of the chapter is devoted to introducing the major regulatory topics in electricity transmission network regulation. These topics and their interrelation are depicted in Fig. 6.1. ([44], Chap. 4) is a good review of regulatory issues in electricity transmission, mostly focused on the US power sector.

Transmission planning seeks to identify the most suitable reinforcements of the network, according to some prescribed criteria. The cost of these reinforcements must be borne by those for whom, or because of whom, they were built. Therefore, there is a close relationship between network planning and network cost allocation. One or more business models must be available for investors to decide to finance projects for the expansion of the grid. The mechanisms for cost recovery are an essential part of the business model and they are directly related to the cost allocation method. Siting is typically a difficult problem, because of the generalised opposition to the presence of transmission lines, and it is more of a social, environmental and political nature. The operation of the network can be considered a separate topic (not represented in Fig. 6.1), with the management of access to the necessarily limited capacity of the grid being the major regulatory issue.

We have grouped these regulatory issues into three major topics, which are covered in the three final sections of the chapter. Network investment, with planning and the associated business models plus siting, is addressed in Sect. 6.3, network cost allocation in Sect. 6.4 and finally the problems associated with network access in Sect. 6.5.

The issues arising around regional or multinational markets and regional interconnections can be found in Chap. 10 of this book.

6.1 Characterisation of the Transmission Activity

The key ideas to proper regulation can hardly be understood without first identifying the features that characterise electric power transmission. The technical and economic realities of this business, as well as the role it is to assume in the new

liberalised context prevailing in the industry, condition and mark regulatory development. What aspects need to be regulated, how they should be regulated and where the chief priorities lie are basic questions that cannot be answered without giving at least some thought to the characteristics and role of this activity.

This section summarises the key technical and economic features of this business. To that end, it draws attention to its impact on electric power production and consumption, analyses the role played by transmission and the structural organisation of its various sub-businesses and identifies the aspects that need to be regulated and the general guidelines that should inform such regulation.

This first section also serves as an introduction to the other sections of this chapter where the areas outlined here are discussed in greater depth. Chapter 1 of this book, which highlights some points of greatest relevance to this chapter, makes for useful supplementary reading.

6.1.1 The Role of Transmission in Liberalised Markets

The transmission grid has acquired particular relevance in the new regulatory framework open to competition, since it is the meeting point for market players and the element that makes the wholesale market possible. Furthermore, the development of domestic and international transmission network connectivity and capacity has paved the way for nation-wide, region-wide or international electricity markets.

Transmission services may be generically defined to be activities with economic value conducted by the transmission network to the benefit of grid users. A distinction may be drawn between:

- primary service: transmission of electricity from production to consumer hubs, and
- secondary services: contribution to system security, in particular regarding voltage control, from generation and (to a lesser extent) demand.

Grid activities include investment planning, construction, maintenance planning, maintenance and operation. Investment planning is the process that determines the commissioning date, location, capacity and other characteristics of new grid assets. Maintenance planning is the process of scheduling line outages for repairs and the tasks required to keep the grid operational to a suitable level of reliability. Construction and maintenance are activities that can be conducted by specialist firms that need not necessarily be electric companies. Grid operation comprises energy flow management through direct action that physically affects transmission installations and must be co-ordinated with the production and consumption facilities. Grids may also participate in the provision of certain ancillary services such as voltage regulation, which are usually managed with specifically designed methods.

Since transmission network reinforcement and maintenance planning impacts the co-ordination of activities that in turn affect the electric energy market, the independence of the organisation responsible for conducting these activities, typically the system operator, must be scrupulously guaranteed. In addition, some of the system operator's decisions also affect the transmission business, since they render certain grid activities, such as line maintenance or the settlement of potential conflicts in the use of generation or transmission, more or less cumbersome or expensive. This may support the case for the separation of system operation and transmission, as in the Independent System Operator (ISO) model used primarily in the US. Having these two activities conducted by separate firms, however, may rule out benefitting from the synergies to be had when they are jointly performed by a Transmission System Operator (TSO), as in most EU systems.

6.1.2 Differences Between the Transmission and Distribution Networks Regarding Regulatory Issues

As noted in Chap. 1, transmission and distribution grids fulfil entirely different purposes. Despite their obvious physical similarities (with components including lines, substations, transformers, protection and operation switchgear, as well as metering equipment, at a range of voltages), their expansion planning decisions, operating methods and asset volumes vary, suggesting a need for differentiated and specialised regulation. They are consequently dealt with in separate chapters in this book. The key issue determining the need for differentiated regulations is the volume of the assets: while in transmission the small number of significant new assets per year makes it possible for regulators to examine each one individually, this is simply not possible with the huge volume of components in the medium and low-voltage distribution grid.

Some ambiguity always exists in each individual system around the role of lines with voltage values in the lower range for transmission and the higher range for distribution. The transmission grid (which operates at very high voltage) is designed to offset the deficits and surpluses between generation and demand in different areas of a country or between neighbouring countries and paves the way for establishing national or even international markets. This ensures that the most efficient generation is dispatched, globally speaking, nearly irrespective of demand location. Distribution grids, by contrast, carry electric power to most consumers, i.e. the consumers not directly connected to the transmission grid. They transform electricity from their high-voltage (HV) distribution lines, which are densely meshed to be able to receive energy from several nearby nodes, to medium (MV) and low (LV) voltages, where the distribution layout becomes more radial. This arrangement is designed to strike the most suitable balance between grid investment costs, energy losses and, especially lately, environmental impact. The layout

and voltage level used for these grids vary with the size of the region covered, the density of electric power consumption and the distance between the main consumer hubs and generating sites. In most countries, the transmission grid consists of lines and other facilities (such as transformers or circuit breakers), typically with phase-to-phase voltages of 220 kV or higher, although in some electrically small countries they may be lower: 132 or even 66 kV. The phase-to-phase voltage in distribution grids varies more widely. In Spain, for instance, voltages of 132, 66, 45, 20, 6, and 1 kV and 380 V are commonly used.

Transmission grid regulation should focus on the questions and issues described in the introduction to this Chapter and summarised in Fig. 6.1: network planning, the definition of suitable business models, siting issues and fair assignment procedures of network access and cost allocation, with the proper performance of the wholesale electricity market as the primary concern. In distribution grids, in turn, where the vast majority of end consumers are connected, service quality is an issue of particular regulatory importance. Furthermore, as indicated before, since the large number of distribution grid facilities rules out individually calculating the regulated remuneration for each, global, simplified procedures need to be deployed, whereas no such procedures are required for transmission grid regulation. Lastly, as discussed below, in competitive electricity markets very strict requirements must be instituted to separate transmission and system operation from generation at the wholesale level, because of the need to guarantee an unbiased market operation. Such measures have not been necessary for the distribution grid, although the presence of distributed generation is also forcing a change in the traditional regulatory paradigm.

6.1.3 The Transmission Grid: Technical Considerations

In network industries, and particularly in electric power transmission, regulation cannot be suitably designed if the view of the underlying physical phenomena is overly simplistic. The physical components of transmission grids and the main technical factors that affect their performance and use are described in Chap. 1. The following discussion deals in greater detail with the factors that condition regulation most directly.

Electric power transport is not transport in the usual meaning of the term, which generally implies the physical shipment of a product from manufacturing plant to consumer. Electric current or electrons travel over the grid at a very modest speed, on the order of a few centimetres per second and have little relevance for the transmission of electricity. The entire network, however, is energised nearly instantaneously (at the speed of light) by connecting generation sources to it and drawing electric power from any point on it. Conductors *guide* electric energy across the network (in the space external to the wires, in overhead lines) from the points where the energy is physically injected to the points where loads are physically located, but the physical phenomenon simultaneously involves all

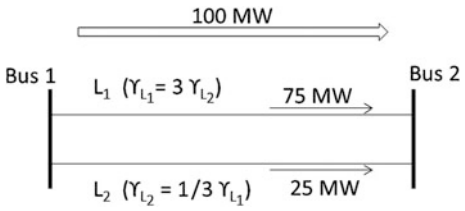


Fig. 6.2 Distribution of the energy flow across two parallel lines, L_1 and L_2 with different admittance values, γ_{L_1} and γ_{L_2}

generators and loads. The overall phenomenon is very complex and cannot be reduced to the simplistic image of the movement of a fluid in a pipeline network. Therefore, there is no unambiguous way of attributing grid flows to system agents, despite the many "reasonable" attempts that can be found in the technical literature.[1] The flows of power in the different transmission lines obey some mathematical equations known as Kirchhoff's laws and depend on the *impedance* of the lines, a physical property of the lines themselves (which for non specialists in power systems could be understood as a measure of the "resistance or opposition" that the line presents to the flow of electricity), as well as on the injections and withdrawals of power at the different nodes and the specific grid topology.

Since the grid is generally meshed, energy uses multiple paths to reach the loads from the generation nodes. Contrary to the situation in the vast majority of other grid infrastructures; however, energy cannot be steered at will through any one or any series of such paths. Energy flows across all of them in proportion to their *admittance*. This physical law is known as Kirchhoff's second (or voltage) law. Admittance, a physical parameter that is the inverse of impedance, characterises each grid line, transformer or other grid component. A simple example is given in Fig. 6.2. Assume that 100 MW of power (if the duration is an hour, the energy measure would be 100 MWh) need to be transmitted from node 1 to node 2. These two nodes are physically connected by two parallel lines, which have different physical specifications (length or thickness of the wires or distance between them, for instance), and consequently different admittance values. Supposing that admittance in line L_1 is triple the admittance in line L_2, the flow of energy (power) is distributed between them as shown in the figure: 75 MW across line L_1 and 25 MW across line L_2.

This circumstance has a number of fairly significant consequences, four of which are relevant to the subject at hand.

Impact on flows of the addition of new lines

Adding new electric lines to an existing grid modifies the way the flow is distributed across the remaining lines in accordance with the new admittance ratios between paths. No consumer can choose to use a new line or otherwise, because

[1] What can be unmistakably computed is the incremental effect on the flows in the transmission grid of an additional bilateral transaction of a specific amount of power between a generation node and a consumption node, assuming a certain pre-existing operating level in the power system.

that cannot be controlled.[2] In the example in Fig. 6.2, if only line L_1 existed, the hypothetical generator located at node 1 would use only that line to transmit its energy to node 2. When line L_2 is installed, the energy produced by the generator is automatically distributed across the two lines in proportion to their admittance values. The generator cannot choose the path or the facilities it wants to use. On a much larger scale, energy exchanged between Germany and Italy, for instance, would not flow only across Switzerland, which lies between the two; rather, part would flow through the Netherlands, Belgium, France, Austria, Slovenia and even Croatia. This is what is generally known in the technical literature as loop flow.

Impact on network capacity of the addition of new lines

Adding electric lines or new electric paths to an existing grid may (although this is not usually the case) detract from transmission capacity and therefore raise system operating costs. This can be readily seen from the example in Fig. 6.2. Imagine that the maximum transmission capacity of line L_1, given its physical specifications, is 100 MW. If only that line existed, the interconnection capacity between the two system nodes would therefore be 100 MW. If a second line, L_2, is installed and its capacity is 20 MW, its presence not only fails to raise total capacity, but in fact reduces the transmission capacity between nodes 1 and 2. Given the three-to-one ratio between the admittance for the two lines, further to Kirchhoff's voltage law, when 20 MW (maximum capacity) flow through L_2, three times that amount or 60 MW flow through L_1. The maximum transmission capacity would consequently decline to 80 MW (60 + 20 MW). Any attempt to transmit more power (energy) across the two lines would require raising the flow through L_2, which would not be possible since it would already be operating at its maximum capacity. Raising the amount of current flowing through that line would constitute a safety hazard, since the conductors in an overhead line will distend, sag and eventually make contact with the ground or vegetation. The overload protection systems installed on the line would isolate the faulted circuit element to prevent such consequences.

Impact of the failure of a line

If a line should fail, the energy flowing across it prior to the incident is redistributed instantaneously along alternate paths. This may entail overloading some other component whose protection relays would break the respective circuits as well, obviously setting off a domino effect and causing major supply outages. The prevention of such situations calls for coordinated and centralised (generation and grid) system operation, applying operating and grid design criteria able to ensure that the chance failure of grid components can be safely handled.

[2] See the comment on the capabilities of FACTS devices later in this section.

Independence of flows from commercial transactions

Energy flows and the use of electric lines do not depend on commercial agreements between market actors (generators, retailers, consumers). Since all these agents may be assumed to behave rationally in pursuit of their maximum benefit, the physical injection and withdrawal of energy at each node on the grid is essentially predetermined, regardless of the commercial agreements concluded between them. Why? Even where a generator commits under a commercial agreement to deliver energy to a consumer at a given point on the grid, if less expensive power (than its operating costs) is available on the market, it will purchase that energy, so that what is ultimately dispatched is the least expensive energy possible. Line use depends exclusively on the energy injected and withdrawn at each node on the grid and the prevailing physical parameters. The inference is that the grid tariff charged to agents for using the transmission grid should depend on their physical location and production and consumption patterns, but not on their (often private) commercial commitments.

The electricity system may well be the sole network business with a grid infrastructure whose technical characteristics impose such peculiar behaviour. At the same time, the decisive role these factors play when designing transmission regulation is readily visible. Given these technical specifications, non-centralised, non-regulated grid expansion planning and operation is difficult to imagine, even in the absence of equally decisive economic factors. Economic issues are discussed in the following section, given their impact on transmission grid users and agents as a whole. The allocation of the costs of the grid use must obviously take these physical conditioning factors into consideration.

In any event, thanks to the progress in power electronics, FACTS hardware (see Chap. 1) is now available that provides for some control over flow distribution, although because of the very high cost of such facilities their use continues to be marginal and sporadic. Another word of clarification is in order in this regard: the physical behaviour described in the foregoing refers to alternating current systems. As described in Chap. 1, direct current energy flows can be controlled, and therefore the planning and use of connections (or lines) with this technology require modifications in the customary regulatory criteria. Finally, the above considerations are less problematic in distribution, particularly in low-voltage grids, where the configuration is typically radial and the flow has always been "outward" or towards end consumers. The simplicity of that layout is also disappearing, however, with significant increases in distributed generation.

Other technical considerations that should be borne in mind for regulatory purposes are ohmic losses and grid constraints. The practical consequences of these technical specifications are analysed in a later section of this chapter.

Table 6.1 Standard costs for AC overhead transmission lines in Spain

	Investment costs (k€/km)		O&M annual costs (k€/km per circuit)
	1 circ.	2 circ.	
Overhead 220 kV	283	436	2.7
Overhead 400 kV	317	488	4
Underground 220 kV[a]	2306	2271	2.1

Source ITC/368/2011 Ministry of Industry, Tourism and Commerce Order of 19th February 2011
[a] The values provided correspond to copper cables of 2.000 mm^2

6.1.4 The Economics of Transmission

The chief economic characteristics of the transmission network can be summarised as set out below.

- Since grid operation and maintenance costs are roughly proportional to the volume of grid assets, total transmission costs can be considered to be driven by investment costs.
- Transmission costs are highly subject to economies of scale, a characteristic feature of natural monopolies.
- The relative economic weight of the transmission network with respect to all the activities involved in the supply of electricity varies widely depending on a country's size and the scattering or clustering of its production and consumption centres. It is significantly lower than generation and distribution, and it typically contributes about 5–10 % to the total cost of electricity.

Tables 6.1 and 6.2 show, by way of example, the standard costs *per unit of length* acknowledged by the Spanish regulator for investment plus operation and maintenance for 220- and 400-kV substations and transmission lines with one or several circuits. Obviously, such costs may vary substantially depending on the characteristics of the terrain and labour costs in each country, among other factors, but they provide an order of magnitude of such costs, which rises with facility voltage.[3]

Note that in Tables 6.1 and 6.2 the investment costs are total overnight costs,[4] while the O&M costs are annual values. In order to be able to compare them on the same basis, if we assume a WACC of 7 % and an economic life of 40 years, the previous tables become Tables 6.3 and 6.4.

[3] A good source of cost and technical data for transmission equipment is provided by the research RealiseGrid Project of the European Commission, see http://realisegrid.rse-web.it/.

[4] Overnight cost is the cost of a construction project if no interest was incurred during construction, as if the project was completed "overnight". An alternate definition is: the present value cost that would have to be paid as a lump sum up front to completely pay for a construction project. *Source* Wikipedia.

Table 6.2 Standard costs for AC substations in Spain, updated to 2008

	Investment costs (k€/km)		O&M annual costs	
	Substations (k€/bay[a])	Transformers (k€/MVA)	Substations (k€/bay)	Transformers (€/MVA)
Conventional 220 kV[b]	877		68	
Conventional 400 kV[c]	1,214		83	
Gas insulated 220 kV[d]	1,401		43	
Gas insulated 400 kV[e]	2,602		53	
Transformers[f]		10		246

[a] *Source* ITC/368/2011 Ministry of Industry, Tourism and Commerce Order of 19th February 2011
A bay of a substation is a part of a substation containing extra-high (or high) voltage switching devices and connections of a power line, a power transformer, etc., to the substation busbar system(s) as well as protection, control, and measurement devices for the power line, the power transformer, etc. Normally, a substation contains a number of line and transformer bays and also other bays
[b] 40 kA per bay
[c] 50 kA per bay
[d] 50 kA per bay. May rise up to 1,751 k€/bay for SF6 gas-insulated ones
[e] 63 kA per bay. May rise up to 3,252 k€/bay for SF6 gas-insulated ones
[f] 400/220/132 kV three-phase transformer

Table 6.3 Standard annual costs for AC overhead transmission lines in Spain

	Annual investment costs (k€/km)		O&M annual costs (k€/km per circuit)
	1 circ.	2 circ.	
Overhead 220 kV	21.2	32.7	2.7
Overhead 400 kV	23.8	36.6	4
Underground 220 kV	173.0	170.3	2.1

Table 6.4 Standard annual costs for AC substations in Spain

	Investment costs (k€/km)		O&M annual costs	
	Substations (k€/bay)	Transformers (k€/MVA)	Substations (k€/bay)	Transformers (€/MVA)
Conventional 220 kV	65.8		68	
Conventional 400 kV	91.1		83	
Gas insulated 220 kV	105.1		43	
Gas insulated 400 kV	195.2		53	
Transformers		0.75		246

The important indicator to ascertain the presence of economies of scale in transmission is the cost of 1 km of line per unit of transmission capacity, and not the cost of 1 km of line. Figure 6.3 illustrates the significance of economies of scale for the transmission business. The solid circles in the figure represent the

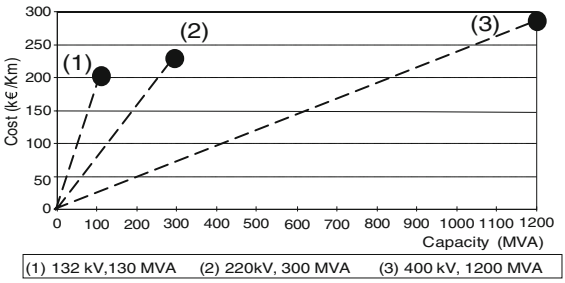


Fig. 6.3 Economies of scale in transmission: unit costs and maximum transmission capacity for transmission power lines

Table 6.5 Cost needed to deliver 2,400 MW for different line technologies [25]

	765 kV	500 kV	345 kV
Capacity per single-circuit line (MW)	2400	910	390
Required number of lines	1	3	6
Average investment cost (M€/km)	1.6	4.3	5.6

transmission capacity and cost (per km) for three line types, which differ essentially in the voltage level at which they operate.

The numerical values are merely indicative, since the transmission capacity of a power line depends not only on its voltage and physical configuration, but also on its length, ambient temperature, grid topology, system stability and short-circuit conditions.

The total cost per km of higher voltage lines obviously exceeds the cost of lower voltage lines. However, as Fig. 6.3 shows, the transmission capacity of a line grows approximately with voltage squared, while costs rise more or less linearly. Hence, the cost of one km of line per MW of transmission capacity is much smaller for a 400-kV than for a 220-kV line. Since this value is equal to the slope of the dashed lines in the figure, the differences between the three types of line are readily visible.

Table 6.5 clearly shows the huge amplitude of the economies of scale. If we assume that 2,400 MW of transmission capacity is needed in a corridor, the next table compares the investment cost per km required if 765, 500 and 345 kV line technologies are selected (these voltages are typical in the US grid). It is also important to notice that the higher utilisation of right-of-way with higher voltage lines greatly minimises the landscape footprint (from 66 m wide for a 765 single circuit kV line to 56 m for each 345 kV single circuit line).

In electricity generation, the economies of scale vanish when the power systems are so large that the total demand to be supplied is many times larger than the optimal size (from economic and technical considerations) of the generating units of the technologies of interest. This is not the case in transmission since, regardless basically of the system size, a transmission line is needed to connect two nodes in the network and the amount of power to be transported between two nodes falls well within the capacity of a single line of the appropriate voltage in all cases.

The use of several lines for this purpose is non-economical. Only reliability considerations may justify the use of more than one line to connect two given nodes. Exceptions to this could be very singular projects that might involve very large volumes of new generation far from demand.

The considerable magnitude of the economies of scale has extraordinarily relevant implications. It is hardly rational for several smaller networks to compete against one another when power can be transmitted over a single large line much more cheaply, provided that a significant part of the line capacity is used. For instance, if the transmission needs between two nodes come to 900 MW, it is more cost-effective to build a 400-kV line than three 220-kV lines (and far more than nine 132-kV lines). This service could not be effectively provided by several competing transmission companies. Line construction could be assigned via a competitive auction, but once it is built only the company that owns the facility can provide transmission service between the two nodes, i.e. it is a natural monopoly.

Another interesting feature of transmission lines is that they are forever. Transmission networks are reinforced and extended, but transmission lines are almost never dismantled. The insulators, conductors and pylons are eventually replaced, but the line and more importantly, the right of way, remains. Once the economic life of the line is exhausted (i.e. the initial investment is fully amortised), the regulator has to figure out the remuneration corresponding to the "permanent life extension costs".

Finally, the weight of transmission grid investment, operation and maintenance costs in the overall electricity system varies substantially with a country's size and geography, mix of electricity generation technologies, as well as the location of some of its primary energy sources (river basins, mines, ports) with respect to its large consumer hubs. Such costs commonly account for around 5 % and at most 10 % of the total in most EU countries (densely populated countries, where distances between generation and demand are fairly short), but up to 20 % in countries where the distances from the best generation sites to the main load centres are long, such as Chile or Brazil. Establishing complex and precise transmission regulation in matters such as grid cost allocation, ohmic losses and management of grid constraints is therefore much more logical in the latter type of countries. In the former, by contrast, much coarser and simpler regulatory solutions may be justified by the greater electricity market transparency and operational streamlining involved. The less efficient use of the grid has been so far of only minor importance because of the scant economic impact of transmission. This situation, however, is rapidly changing in many countries with the numerous requests for network access by wind and solar farms, which require efficient locational signals to guide the siting process. And the proposal of megaprojects involving large deployment of renewable generation—in Europe the off-shore wind from the North Sea or solar from Southern European countries and Northern Africa, and in the USA the on-shore wind from the Great Plains or solar from the South West deserts—and the corresponding transmission developments are bringing to the fore the costs of transmission and the rules for their allocation.

6.1.5 Transmission as a Natural Monopoly

As discussed above, the existence of substantial economies of scale is a major factor in characterising electricity transmission as a natural monopoly. Other factors also contribute to such status; however, which to some extent also characterise distribution grids (particularly when operating at higher voltages). From a regulatory standpoint, this conclusion is of major importance, for it infers that transmission management and regulatory design must be in keeping with that status.

The main factors that justify regarding electricity transmission as a natural monopoly are summarised below.

- The duplication, or multiplication, of transmission or distribution grids is presently unthinkable in terms of cost, land use and environmental impact. End consumers should be supplied through a single transmission and a single distribution grid.
- The preceding conclusion is reinforced by the fact that there are large economies of scale in electric power network construction. As shown above, the unit cost of transmitting electric energy (cost of one km of line per unit of transmission capacity) declines steeply, all other things being equal, with a line's total transmission capacity. It makes more economic sense, then, to have a single grid with the transmission capacity that can do the job, rather than a multiplicity of grids with each one only meeting part of the total need for transmission capacity.
- The capacity of transmission facilities, power lines in particular, cannot be freely designed. Rather, only two or three standard voltage levels and a few configurations (such as single or double circuit or simplex or duplex conductors) are feasible. Therefore, only a limited number of discrete values of new transmission capacity can be installed. This, in conjunction with the existence of economies of scale, typically results in a volume of investments that—initially at least—amply exceeds the transmission capacity that would be strictly needed under ideal conditions (continuous investments and no economies of scale).
- Unregulated ownership, operation and pricing of transmission facilities in a competitive market environment would bestow strong market power at the wholesale level. This would be very detrimental to the system as a whole.
- For the specific technical characteristics discussed earlier, the transmission network must operate and be operated as a whole.

It might be wrongly concluded from the foregoing that there should be a single transmission owner and also a single distribution grid owner in each region. This is not the case (although some power systems may choose this solution), since facility ownership may in fact be fairly widely shared, providing all installations operate in a co-ordinated fashion under the direction of the respective system operator and remuneration of the transmission facilities is subject to regulation. Exceptions to this principle are possible (the so-called "merchant lines") and are discussed later in this chapter.

In summary, transmission and distribution companies have enormous market power, since each is indispensable for the supply of electricity. Consequently, their remuneration must be regulated and they must be obliged to provide open and non-discriminatory access to their facilities when the system operates under a competitive framework. To avoid conflicts of interest, these network companies should ideally be independent of the market players. Such independence is imperative for system operators.

6.1.6 *Network Effects of Transmission on System Operation Costs*

The producer–consumer connections provided by electric power networks are not perfect; their imperfections can be divided into three main categories: network losses, grid-imposed operating constraints and adverse consequences for quality of service. These effects naturally have an impact on transmission grid investment and operating decisions, as well as on operating costs and system generation dispatching. Due to these effects, both the total operating costs and marginal generation costs incurred to meet system demand are subject to (possibly significant) change. For the same reason, the location of generating plants and consumers on the grid may become a relevant cost factor, and effective grid regulation may need to include efficient locational signals.

Hence, the importance of fully understanding the nature of these additional costs and their behaviour and impact on dispatching is clear. The present section attempts to clarify these issues. By way of illustration, the explanations are supplemented with the simple conceptual example shown in Fig. 6.4, which consists of two nodes (buses) connected by an electric line. The demand at both nodes at a given time is 50 MW. The generating unit located at node 1 has a production capacity of 200 MW and a unit production cost of €50/MWh. The unit at node 2 also has a production capacity of 200 MW, but a unit production cost of €70/MWh.

If the line were ideal, i.e. if it had zero losses and infinite capacity, economic dispatching (the least expensive way to meet the 100-MW total demand with the available generating capacity) would involve producing the 100 MW with the cheapest generating unit (unit 1), 50 MW of which would be used to meet the demand in node or bus 1 and the remaining 50 MW would be sent by the line to

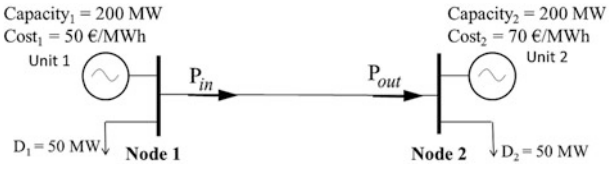


Fig. 6.4 Two-node system

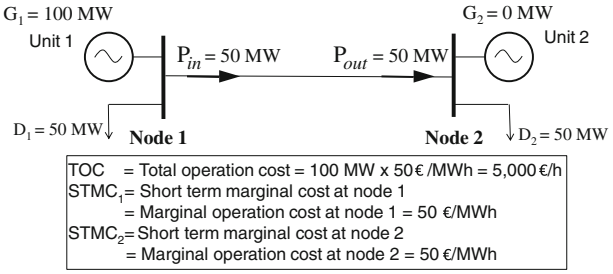


Fig. 6.5 Optimal dispatch for ideal line behaviour

node 2 to meet its demand. In that case, as shown in Fig. 6.5, the amount of energy flowing through the line would come to 50 MW and all the power flowing out of node 1 would reach node 2. The dispatching cost would be the cost of producing 100 MW at €50/MWh or €5,000/h. If the demand at node 1 rises by one unit, the most inexpensive way to meet that rise would be to produce one additional MW with unit 1, for an extra cost of €50/MWh. Consequently, the system marginal cost at node 1 (the additional cost of meeting an increase in demand at node 1 as cheaply as possible) would be €50/MWh. The marginal system cost at node 2 may be readily deduced to also be €50/MWh.

The impact of ohmic losses and grid constraints on optimal dispatch costs and marginal costs is analysed below.

Ohmic losses

Most of the energy losses in electric power grids are due to the resistance of conductors to the circulation of electric current flows, a phenomenon known as the Joule effect or ohmic losses. Other losses may be due to the so-called corona effect, induced by electrical discharges in the air surrounding high-voltage line conductors, and yet others to internal losses in a variety of grid elements such as transformers, reactors and capacitors. The most significant effect of losses is that consumers receive less energy than generators produce.

Transmission network losses result in additional system costs, because more energy has to be produced than is required by consumers. These costs constitute additional generation costs that arise because of grid characteristics. They are not grid costs per se, although they are induced by transmission facilities and are impacted by investment in transmission and the system operator's decisions. Regulatory instruments should therefore be sought that incentivise grid loss cost reductions.

Ohmic losses are nearly proportional to the square of the flow of energy circulating in the line (actually, they are proportional to the square of the current circulating in the conductors). This means that, for a system operating under certain conditions and therefore with certain grid flows, the average value of losses is not equal to the marginal value of such losses. Rather, the average value is less than the marginal value. In other words, if the system load increases by one unit

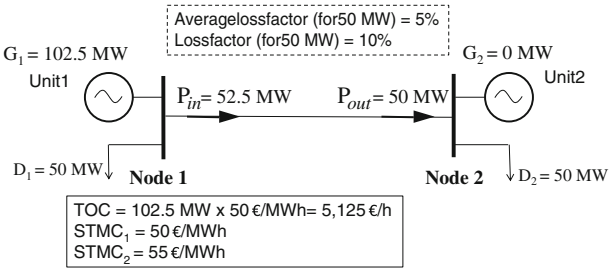


Fig. 6.6 Effect of transmission ohmic losses on dispatch costs and short-term marginal costs

(marginal increase), losses rise by a value proportionally greater than average system losses (total losses/total system loading). Consequently, the marginal value of the extra cost due to transmission network losses (cost increase/increase in system load) differs from the average value (cost/total system load).

Figure 6.6 provides a simple example of the effect of grid ohmic losses on both dispatch costs and short-term marginal energy costs (defined below as nodal prices). Here the electric line is not ideal. The electric line has resistance, and this resistance induces ohmic losses. Assuming that in this example the losses in the line are expressed as Losses = 1/1,000 × flow^2, a flow of 50 MW results in a loss of 2.5 MW (5 % of the value of the flow). The ratio losses/flow, defined hereafter as the average loss factor, differs for any other flow since losses are proportional to the square of the flow. In this example, the average loss factor = losses/flow = 1/1,000 × flow = 0.05 MW/MW. Given these values, the least expensive way to meet the 100 MW demand is to produce 102.5 MW of power with generating unit 1 (100 MW to cover demand, plus 2.5 MW of ohmic losses). The total dispatching cost comes to 102.5 MW × €50/MWh = €5,125/h, for an extra cost of €125/h.

Under these operating conditions, given the quadratic relationship between losses and power flows, losses vary with unit fluctuations in flow at a rate of 0.1 MW/MW. Indeed, ∂Losses/∂flow = 2 × 1/1,000 × flow = 2 × Losses/flow = 0.1 MW/MW. This, as the figure shows, is generally called the loss factor (this term should be used with care, since the same expression can be found in the literature with other meanings). One additional MW of demand at node 1 would require producing 1 MW with generating unit 1, for an additional cost of €50, whereas the cost of covering one additional MW of demand at node 2 would require producing 1.1 MW more with generating unit 1 (1 MW for the demand plus 0.1 MW in additional line losses), for an additional cost of €55. Stated in economic terms, the short-run marginal cost at node 1 (the cost of meeting a marginal increase in demand at node 1) is €50/MWh, while the short-run marginal cost at node 2 is €55/MWh.

In other words, since ohmic losses create geographic differences in the marginal cost of supplying electric energy, the marginal cost of covering a marginal increase in demand can only be correctly assessed if the exact node involved is specified.

If generating unit 2 had operating costs of 52€/MWh, for instance, it would have been more economical to use it to meet a fraction of the demand at node 2. Why? With no flow in the line, and therefore no losses, it is cheaper for the system to bring power at 50€/MWh than to produce locally with generator 2 at 52€/MWh. However, as the flow increases to meet demand at node 2 with production at node 1, losses also increase in a quadratic way and imports to node 2 will increase the cost of losses. Therefore, each additional increment of import to node 2 from node 1 will cost higher and higher. An equilibrium exists between the increased cost of importing to node 2 (increased costs due to producing with generator at node 1 plus increased costs due to losses) and the increased costs due to producing with the generator at node 2. Once the increased cost from importing exceeds the increased cost of producing with the generator at node 2, all remaining demand can be met with the node 2 generator, and the marginal price at node 2 will not exceed 52€/MWh. Consequently, depending on grid location and given ohmic losses, it may be more economical to dispatch units with higher, instead of units with lower, operating costs. The system operator takes the grid losses incurred by each unit into consideration when computing its operating cost; depending on the location of the generating unit, more than 1 MW[5] must generally be produced to meet 1 MW of demand, since the rest is lost during transmission. Each unit's dispatching merit order is in fact altered by the loss coefficients[6] associated with its grid location.

Grid constraints

Grids limit the operation of the electricity system in many ways. The most typical limitation is congestion, which occurs when the maximum current that can be handled by a line or other facility is reached, thus determining the amount of electric power that can flow through the element in question. The underlying cause for the limitation may be thermal, and therefore dependent upon the physical characteristics of the facility. It may also be related to the characteristics of system operation as a whole; for instance, provisions to guarantee security in the system's dynamic response to disturbances or to stability-related problems that usually increase with line length. Another typical grid constraint is the need to maintain voltages within certain limits at all nodes, which may call for connecting generating units near the node experiencing problems. The maximum allowable short-circuit power established may also limit grid configuration. Generally speaking, the main effect of grid constraints is to condition system operation and in so doing to cause deviations from economically optimum operation. The most common constraints in distribution grids are related to voltage and maximum line capacity.

Just as in the case of ohmic losses, the mere existence of the transmission grid constraints adds to system costs by requiring operation with more costly

[5] Less than 1 MW may also be required, if the generator's specific position induces counter flows (with smaller flows and consequently smaller losses) in the system.

[6] As explained above, these factors depend on the operating conditions prevailing at any given time.

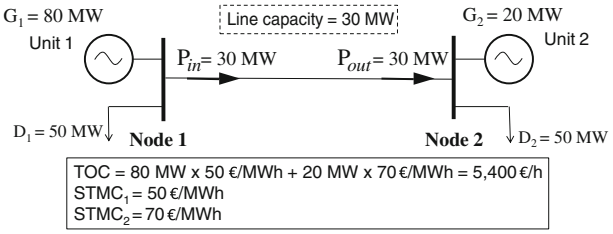


Fig. 6.7 Effect of transmission congestion on dispatch and short-term marginal costs

generating units to accommodate the physical limitations entailed. This is not to imply in any way that grid design or development is flawed, since investment of the sort required to ensure the total absence of constraints in the system would probably not be economically justified. Constraints may be justified from an economic standpoint, in particular when they are only sporadically active or when the cost of removing them is very high.

Grid constraint costs, like the cost of ohmic losses, constitute additional generation costs occasioned by grid characteristics, but are not grid costs per se. The points discussed above with respect to losses also apply to grid constraints.

Figure 6.7 provides a simple example of the effect of grid constraints on both dispatching and short-term marginal energy costs (defined below as nodal prices). Here the effect of ohmic losses is ignored, although the line capacity limitation is considered, with an assumed maximum of 30 MW. Under such circumstances it may be readily deduced that, even seeking the least expensive way to cover system demand, the cost is clearly higher than in a scenario where the line is subject to no capacity limit. Indeed, in this case the generating unit at node 1 would produce 80 MW at a price of €50/MWh, 50 of which would be used to cover the demand at its node and 30 would be exported over the line to node 2. Generating unit 2 would produce the remaining 20 MW needed to cover the total demand at node 2, but at a much higher cost (€70/MWh). The total dispatching cost amounts to 80 MW × €50/MWh + 20 MW × €70/MWh = €5,400/h, or an additional €400/h attributable to the technical limitations affecting the transmission grid. The generating units have to be re-dispatched to accommodate the grid constraint. As in the case of ohmic losses, this effect impacts the marginal costs of each node (nodal price). A marginal increase in demand at node 1 can be covered with a marginal increase in generating unit 1 output at a cost of €50/MWh, whereas a marginal increase in demand at node 2 must necessarily be covered by a marginal increase in generating unit 2 production at a cost of €70/MWh.

As shown, ohmic losses and grid constraints induce changes in the economics of system operation. The merit order in which generating units start to produce depends not only on their operating costs but also on their grid location and impact on ohmic losses and grid constraints. Furthermore, extra system costs appear that may need to be allocated, and the cost of meeting demand is not uniform across the system, but depends on unit location on the grid.

Nodal prices are economic energy purchase signals that efficiently internalise all grid-related effects. In light of their importance, nodal prices are specifically addressed in Sect. 6.2.

Quality of service
The third important way that real networks affect electricity system operation is related to quality of service. In countries with a well-developed electricity system, interruptions in consumer service can seldom be attributed to insufficient generation; in a tiny percentage of cases their origin lies in joint generating and transmission security failures (although the consequences of such events are usually very severe and affect large areas in the system). They are in fact practically always due to local distribution grid failures. Distribution business regulation should seek a balance between increases in the investment needed to improve grids and the resulting enhancement of end consumer quality of service. This subject has been addressed in Chap. 5 of this book.

6.1.7 Major Regulatory Issues in Transmission

In light of the previous discussion, the new regulation of electric power grids can be reduced to three main aspects: investment, access, and pricing or network cost allocation, as discussed below. The aims of such regulation are:

- to contribute to efficient network expansion,
- to ensure the economic viability of the transmission business through suitable remuneration,
- to further economic efficiency for network users both in the short term (attainment of optimal operation) and the long term (sending correct locational signals for future network users, whether they be generators or consumers),
- and to contribute to the efficient operation of the transmission network and the appropriate maintenance of its facilities.

6.2 Transmission and Locational Energy Prices: Nodal Pricing

The presence of the transmission grid, with its ohmic losses and potential grid constraints, has been shown to cause energy prices to differ from one transmission node to another, thereby sending location-related economic signals to grid users. Depending on how these signals are implemented in practice in each specific power system, energy prices may also entail partial recovery of total transmission costs, but only a small fraction of the total amount in general, as it will be shown later.

Locational energy prices send the right economic signals to the market players, enabling the market to operate properly in the short term (with respect to losses and possible grid congestion), as well as in the long term, encouraging future players, be they producers or consumers, to choose their locations accordingly.

Short-term locational energy prices apply to all the MWh injected into or withdrawn from the transmission grid. Short-term locational signals, i.e. energy prices that vary hourly in day-ahead markets or even shorter intervals, such as several minutes, in real-time markets, are needed to secure system efficiency. The term "efficiency" refers to ensuring that the generators with the lowest variable costs are dispatched as much as possible, and that demand is able to respond to the actual costs of supplying energy at each system location taking network effects into account. Short-term signals also have a long-term impact, since expectations around future values of energy prices at different locations affect players' long-term decisions, in particular with respect to the siting of new generation or demand facilities.

In a hypothetical electricity system with no technical or capacity constraints or energy losses in its transmission facilities, the locational component of energy prices would be nil. In actual power systems energy prices vary in space and time. The most sophisticated and efficient expression of these signals is to be found in nodal pricing.

This section introduces the basics of nodal prices and a series of examples of practical application in electricity markets. Whether and how these prices are implemented in electricity markets affect the regulatory problems around transmission discussed in the following Sects. 6.3, 6.4 and 6.5: investment, access and use of the grid whose interconnection capacity is limited, and transmission remuneration and cost allocation.

6.2.1 Definition of Nodal Prices

Marginal system cost, defined in earlier chapters as the cost of meeting a differential rise in demand as cheaply as possible, is the economic signal that encourages efficient generator and demand-side market behaviour. As discussed above, when transmission grid effects are taken into account, system marginal cost is not a single value, but varies from node to node. The nodal price at system node k is defined as the system's short-term marginal operating cost of meeting an increment in demand at node k as economically as possible and within the constraints imposed by the system. Nodal prices are also called spot prices and locational marginal prices. The concept of nodal prices and its potential applications was elaborated and thoroughly developed by Professor Fred Schweppe and colleagues at MIT in the early 1980s [67].

When actual system realities are taken into account, then, system marginal cost is an understandable concept only if it is differentiated by node. The notion of a single or uniform marginal cost mentioned in earlier chapters and used in practice

in several electricity markets (in most EU countries, for instance) is a mere approximation in which the geographic differentiation imposed by the transmission grid is ignored.

Strictly speaking, a nodal price can be determined for both active and reactive energy, although the latter has seldom been used in electricity markets.

System marginal cost may, moreover, be determined for the long and the short term. The definition, differences and equivalences of this concept are discussed in Chap. 2, where grid effects are ignored. Short- and long-term nodal marginal prices can also be determined. Long-term marginal prices assume that both generating and grid investments can respond marginally and optimally to meet demand at any given node, while the calculation of short-term marginal costs assumes fixed investments. The relationship between the two costs and their behaviour follows the same pattern as described when no account is taken of grid effects.

6.2.2 Properties of Nodal Prices

Nodal electricity pricing can be shown to precisely reflect the impact of the transmission network on the electricity system, from the technical and economic standpoints, including both line losses and grid constraints. The nodal price at any given time depends on system operating conditions: available generating and transmission facilities, load level at each node, each generator's variable costs, active generation and grid losses and constraints. The most prominent features of nodal pricing are discussed below.

Property 1: Efficient short-term energy prices

The most outstanding feature of nodal pricing is that it sends out efficient short-term locational signals for generation and consumption, since it internalises all the grid effects in a single value (in monetary units per kWh) that depends on each system node. As shown earlier, optimal dispatch depends on network effects such as losses and constraints. The nodal price, as the purchase price of the energy produced and consumed at every single system node, sends the appropriate economic signal to each and every network agent to induce more efficient operation and maximise the social benefit produced by the system.

The following example illustrates this point. The layout of a nine-node system is depicted in Fig. 6.8. Generation units are defined by two values, their maximum available capacity (in MW) and their variable operating costs (in €/MWh). Several units may be connected to the same node. The total load connected to each node is also shown (in MW). Technical data of the system required to compute the optimal load flow of this system (resulting energy prices and production levels per node) are provided in Annex A to this chapter.

Figure 6.9 shows the result of optimal generation dispatching (the values alongside the generation unit symbols give the total power output for each system

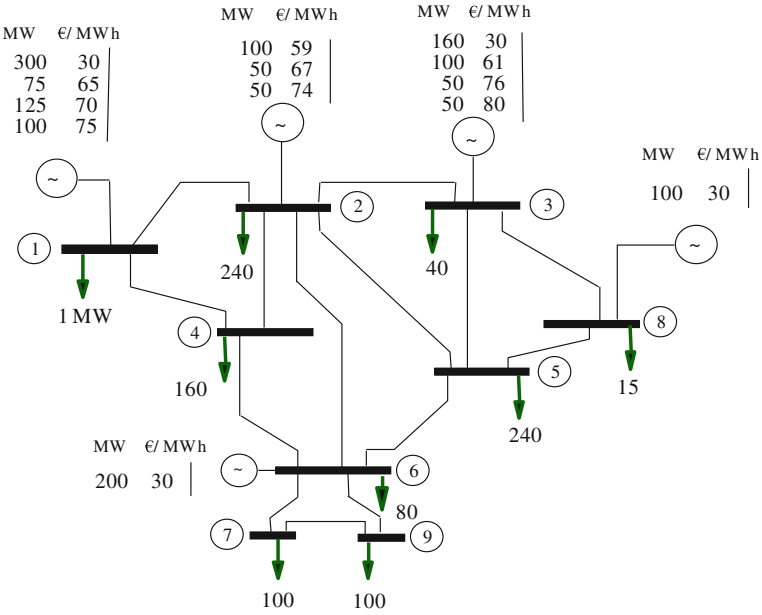


Fig. 6.8 Nine-node example of nodal prices: configuration and data

node) and the resulting nodal prices (in €/MWh), when the effect of ohmic losses is ignored and the line capacity is assumed to be high enough not to condition optimal dispatching. This, of course, is tantamount to completely ignoring the grid (the so-called single bus model). Generation units are dispatched in strict accordance with their own operating cost, regardless of their location on the grid.

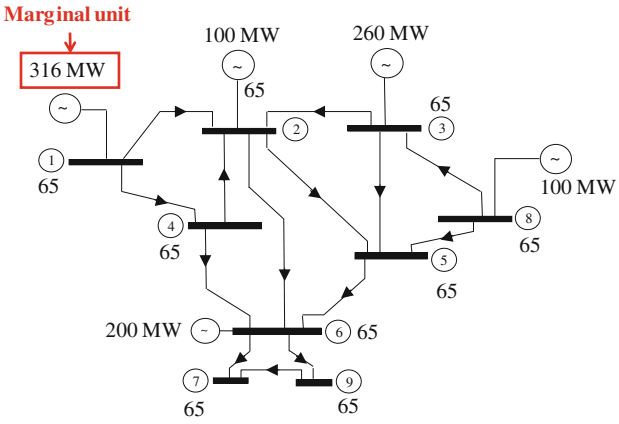


Fig. 6.9 Dispatch results and nodal prices ignoring grid losses and constraints. In this figure and in the following, all prices are expressed in EUR/MWh

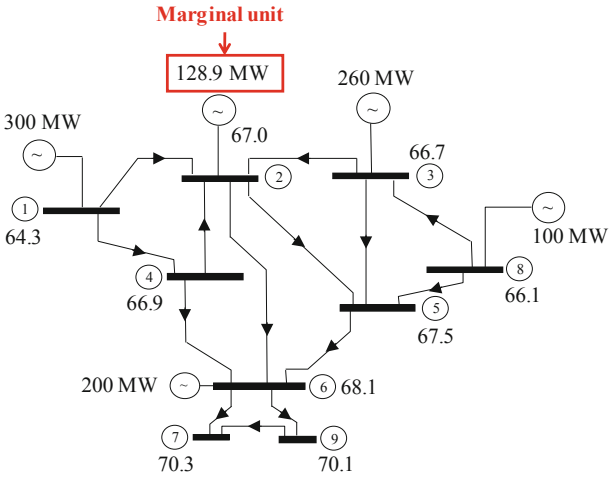


Fig. 6.10 Dispatch results and nodal prices considering grid ohmic losses

Moreover, since nodal prices are same across the entire system, the cost of meeting a unit increase in load is exactly the same, irrespective of the node. When losses and constraints are ignored, the unit increase of a load, wherever it occurs, is met by the marginal unit, i.e. the available unit that is able to produce electricity at the least expensive variable cost (in this case, the second unit at bus 1), whose operating cost is €65/MWh. Since this is the efficient energy price, any unit with a lower operating cost than the price at the node where it is connected operates at maximum capacity: as it is paid more than its operating cost, it produces as much as it can. By the same token, any unit with a higher operating cost than the price at the node where it is connected operates at minimum capacity: as it is paid less than its operating cost, it produces as little as it can.[7] The marginal unit's operating cost is equal to the price of energy at the node where it is connected.

Figure 6.10 shows the dispatching results and nodal prices obtained when line ohmic losses are taken into consideration. An analysis of the findings shows that the overall output would be larger than in the earlier discussion, for not only demand but the losses originating in the lines (which total 12.9 MW) must be covered. In addition, and significantly, different units are dispatched. Indeed, when only the costs of the unit itself are considered, the additional 12.9 MW needed are met with the 12.9 MW still available in the unit with an operating cost of €65/MWh (located at node 1). Nonetheless, that arrangement does not lead to optimal dispatching. Based on the flows created and consequently the ohmic losses incurred, the identity of the system node where each generating unit delivers its

[7] For the sake of simplicity, all generating units were assigned zero minimum capacity output. In real systems, thermal units that are connected and in operation cannot produce below a technical minimum output.

energy is very material. Minimisation of the operation cost in fact calls for dispatching the second unit at node 2, with an operating cost of €67/MWh, instead of the second unit at node 1, with an operating cost of €65/MWh. Therefore, the second unit at node 2 now becomes the marginal unit of the system. The node 2 units clearly create lower ohmic losses than the node 1 units, and the second unit at node 2 would therefore have dispatching priority over the second unit in node 1, despite its slightly higher operating costs. Notice, therefore, that considering losses will change the merit order dispatch and may lead some units to decrease their production because, even if cheaper, they are poorly located regarding ohmic losses in the system.

Nodal prices are different at each system node and higher than in the preceding case. When demand rises at any of the nodes, the least expensive unit is the marginal unit, which in this case is the second unit at node 2, with an operating cost of €67/MWh. Here, however, to meet 1 MW of additional demand at any node except node 2, more than 1 MW must be produced with the marginal unit to compensate for the loss of part of the energy during transmission to the demand location. To meet an additional 1 MW of demand at node 9, for instance, 1.0463 MW would have to be produced with the marginal unit, 0.0463 MW of which would be lost during transmission. The marginal cost at node 9, therefore, would be €67/MWh × 1.0463 MW/MW = € 70.1/MWh. Since flow distribution follows the Kirchhoff laws mentioned above, energy transmission from the marginal unit node to each of the other nodes creates different flow changes and consequently different losses. Consequently, each node has a different nodal price.

Notice that at nodes 1, 3, 4 and 8, nodal prices are lower than the cost of the marginal unit. This indicates that to meet 1 MW of additional demand at any of those nodes, less than 1 MW must be produced with the marginal unit. Indeed, a look at the direction of the line flows in this case shows that power is flowing from those nodes to the node where the marginal unit is located; therefore an increase of demand at those nodes together with an increase of generation of the marginal unit will cause a reduction of the flows in those lines and therefore a decrease of the ohmic losses. For instance, to meet an additional 1 MW of demand at node 8, only 0.9865 MW would have to be produced with the marginal unit, because 0.0135 MW of ohmic losses would be saved in the transmission network. The marginal cost at node 8, therefore, would be €67/MWh × 0.9865 MW/MW = €66.1/MWh.

Nodal prices send efficient signals for optimal system operation. If the prices paid to the generating units are the nodal prices specified, the result of these signals would, then, be the optimal dispatching arrangement shown in Fig. 6.10.

Finally, Fig. 6.11 shows dispatching and nodal prices when the system is affected by a grid constraint. In this case, with a lower transmission capacity in the line connecting nodes 3 and 5, system dispatching must be modified to comply with that constraint. Output by the units sited at node 3 must be reduced and compensated by increasing output in the units at nodes 1 and 2. This gives rise to two marginal units, whose nodal prices reflect the grid constraint. At the nodes where generation has to be reduced, releasing cheap generating capacity, the

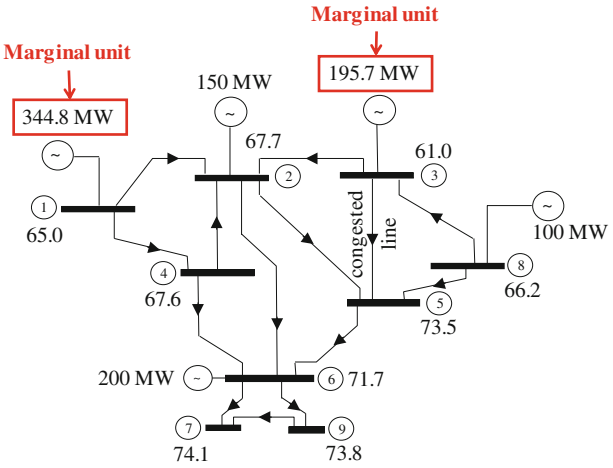


Fig. 6.11 Dispatch results and nodal prices considering both grid losses and constraints

marginal price is even lower than when no constraints are in place.[8] At the rest of the nodes, however, more expensive units have to be dispatched and the grid constraint prevents meeting the demand at those nodes with the inexpensive generating capacity available at node 3. Kirchhoff law-mediated flow distribution results not only in an impact on node 3 and 5 prices, which are directly affected by the constraint, but also on prices at the rest of the nodes.

In other words, nodal prices also send efficient signals for optimal system operation in the presence of grid constraints. If the prices paid to the generating units are the nodal prices specified, any unit with an operating cost higher than the nodal price at the node where it is connected operates at its minimum output (zero in this case). By the same token, any unit with an operating cost lower than the nodal price at the node to which it is connected operates at its maximum output. The result is the dispatching arrangement shown in Fig. 6.11, which makes allowance for the maximum capacity constraint at line 3–5.

Any other type of grid constraint could be analysed in this same way, including any manner of inter-area exchange limitation (due to stability problems or commercial considerations, for instance).

Property 2: Efficient allocation of network losses, associated costs and dispatch costs due to network constraints

Nodal pricing efficiently allocates the extra costs associated with network ohmic losses and constraints. Demand at nodes where delivery causes higher losses must pay a higher nodal price than demand at nodes where supply lowers losses.

[8] In extreme cases, some nodal prices may even turn negative, because the increased demand at each node alleviates a nearby constraint, with the concomitant change from more to less expensive generation.

Generating units poorly positioned in terms of losses are paid lower nodal prices, in turn, than units located more rationally (in that respect).

The most graphic example may be found in a system with a large generating plant that produces cheap electricity located at a long distance from consumers (a real example might be the Comahue region in Argentina that supplies greater Buenos Aires with electricity over a grid nearly 2,000 km long). Due to the losses involved, the nodal prices in the area where power is generated and exported are lower than in the importing, consumer region. A generator located in the former area is "poorly" positioned in terms of ohmic losses, for its output must be carried to the consumer hubs with significant losses along the way. This generator is paid a lower nodal price for the energy generated than if it were located at the import node, where the nodal price is higher. Under this arrangement, the cost of the losses is allocated to the generator primarily responsible for them. Conversely, demand located in the consumer region causes the greatest ohmic loss in the system, since its energy must be transmitted from where it is generated. Demand located in the export or generating area, by contrast, causes no system losses, since its energy is supplied by local generators and need not circulate on any line. Demand in the import area has a higher nodal price as a result of the allocation of more of the extra cost associated with ohmic losses.

A similar situation arises in connection with grid constraints. Demand or generating units that cause grid constraints and consequently the need for more expensive re-dispatching are subject respectively to higher or lower nodal prices, whereas demand and generating units that alleviate the constraint benefit from lower nodal prices. Hence, the extra cost associated with grid constraints is allocated to the agents actually responsible for them. Since several grid constraints naturally tend to co-exist, locating the responsible agents is no easy task, particularly in view of the fact that electricity flows are governed by Kirchhoff's laws. Nodal prices are able to make such complex allocations efficiently.

Nodal prices consequently send out an initial location signal associated with system operation. Generating units and demand can raise their revenues and lower their costs, respectively, if they connect to the most suitable grid nodes (nodes that cause the lowest ohmic losses and constraints in the system).

Property 3: Failure of network revenues from nodal pricing to recover investment and maintenance costs

Once nodal prices are in place, it makes sense to argue that the economic value of the transmission network consists of purchasing electricity at a price equal to the nodal price in those nodes where electric energy is injected in the network, and selling this energy (minus any losses that may occur), priced at nodal prices, at the nodes where it is withdrawn from the network. The network revenue would be the profit that the transmission network would earn if energy were purchased from generators at their nodal price and sold to consumers at theirs. However, as explained before, the transmission network should not be allowed to freely establish a price for its use, but must rather be treated like a regulated monopoly

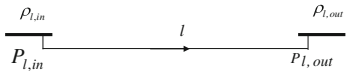


Fig. 6.12 Line revenues from nodal pricing

with pre-established remuneration. Exceptions may be justified for individual lines under special circumstances, as discussed later in this chapter.

In any case, the application of nodal prices, whereby each generator is remunerated at the nodal price prevailing at its node and each consumer pays the nodal price at its node, gives rise to some network revenues (NR), also known as the variable transmission revenues (VTR):

$$\mathrm{VTR}_t = \mathrm{NR}_t = \sum\nolimits_n \left(\rho_{n,t} \cdot g_{n,t} - \rho_{n,t} \cdot d_{n,t}\right) \tag{6.1}$$

where

- NR_t transmission network revenue at time t for the whole system
- $\rho_{n,t}$ nodal price at node n at time t
- $g_{n,t}$ generation output connected to node n at time t
- $d_{n,t}$ demand consumption at node n at time t.

The annual variable transmission revenues are

$$\mathrm{AVTR} = \mathrm{NR} = \sum\nolimits_{t \text{ for one year}} \mathrm{NR}_t \tag{6.2}$$

These network revenues for the whole system can be also computed line by line, as in Fig. 6.12 [47]. Each line may be regarded as buying energy at the node where the flow enters the line and then selling that energy at the other node. Since the flow into the line at the former node and the flow out of the line at the latter differ because of ohmic losses, and since nodal prices at both line end nodes also differ, this economic transaction has net value.[9]

$$\mathrm{VTR}_t = \mathrm{NR}_t = \sum\nolimits_l \mathrm{NR}_{l,t} = \sum\nolimits_l \left(\rho_{l,\mathrm{out},t} \cdot P_{l,\mathrm{out},t} - \rho_{l,\mathrm{in},t} \cdot P_{l,\mathrm{in},t}\right) \tag{6.3}$$

where

- $\mathrm{NR}_{l,t}$ network revenues for line l at time t
- $P_{l,\mathrm{in},t}$ power flow at the line l in node "in" at time t
- $P_{l,\mathrm{out},t}$ power flow at the line l in node "out" at time t
- $\rho_{l,\mathrm{in},t}$ marginal price at node "in" at time t
- $\rho_{l,\mathrm{out},t}$ marginal price at the node "out" at time t.

The NR values in Eqs. 6.1 and 6.3 are equivalent. Again, network revenues are due to the nodal price differences induced by transmission losses and transmission congestion. Network revenues due to congestion are also known as *congestion rents*.

[9] Very exceptionally, such "network revenues" may be negative for some lines, when lines experience large losses due to corona discharge, as it has been the case in certain lines in Peru.

Network revenues must be taken into account when defining the regulatory mechanism for remunerating the transmission network, as discussed below.

The net revenues calculated by applying nodal prices depend critically on how fully the transmission grid is developed. As Eq. 6.3 shows, the greater the differences between system nodal prices, the greater are the resulting revenues. Furthermore, as seen earlier, the geographic differentiation in prices is the outcome of ohmic losses and grid constraints. Consequently, highly developed and densely meshed grids have a very small ohmic loss volume (flows are distributed along more lines, lowering losses, which are quadratically dependent upon flows) and cause scant constraints in the system. Prices differ less from node to node, leading to low grid revenues. By contrast, less developed grids with high loss levels and substantial constraints have much larger grid revenues because nodal prices tend to diverge widely.

Pérez-Arriaga et al. [56] showed that, under certain ideal grid investment conditions discussed below, in an optimally developed grid[10] the resulting grid revenues would recover 100 % of the total transmission grid costs (grid investment costs). However, this equality of revenues from nodal prices and total network costs only holds under ideal grid investment conditions, which differ substantially from actual circumstances. In practice, as it has been repeatedly shown in numerous power systems, network revenues only recover a fraction of total network costs, typically on the order of 20 % or less [56]. Rubio-Odériz [64] studied the reasons for such a huge shortfall in detail. Briefly, the primary reasons identified were as follows:

- economies of scale,
- the discrete nature of transmission investment (lumpiness effect),
- reliability investment criteria that are more stringent than the criteria actually applied during the actual operation of the power system,
- discrepancies between static and dynamic grid expansion plans, and
- planning errors and deviations that, unavoidably, happen in practice.

An intuitive explanation of the causes for the revenue shortfall is in order. The primary cause of cost under-recovery lies in the large economies of scale combined with the discrete nature of investments. Because of these two factors, optimal grid reinforcement decisions result in overinvestment with respect to the ideal, in which no economies of scale are in place and where the exact grid capacity required (with continuous and non-discrete variables) can be attained (with continuous variables). The consequences of an over-developed grid for the revenue-cost balance were discussed earlier. Opting for lines with higher capacity

[10] As discussed later in this chapter, an optimally developed grid is understood to mean a grid in which the expected sum of investment, operation and maintenance costs is minimised, or a grid in a liberalised environment that maximises expected net generation and demand-side benefits. These are net benefits, and the grid investment costs, which must be jointly paid by these agents, are given negative values.

is always more economical, even when their full capacity is not strictly needed—initially, at least—, compared to the alternative of building several lines with smaller capacity, which might adjust more tightly to capacity needs.

The existence of economies of scale, together with the lumpiness of the available network expansion options tilts the network expansion plans towards lines with more capacity than optimally needed in a continuous world, but which are optimal in the real world. This adds to the discrepancy between static and dynamic expansions, i.e. thinking of the future, the transmission planner invests in lines that may be over-dimensioned during the early part of their economic lives. Planning errors (i.e. lines that were built and are not used as much as the planner thought) reinforce the previous effects. Lines are never removed, and if the error was in the opposite direction the planner always can fix it by adding reinforcements, but not the other way around.

The second major family of causes of cost under-recovery is related to the fact that nodal prices are short-term (operating) signals, and there will be discrepancies between the short-term reality and the long-term estimates of the network planner. This is due to the failure of the critical situations supporting the reliability criteria adopted in the expansion decision phase to materialise in everyday real system operation. The transmission network planner designs a network that can cope with many possible adverse situations. At the time of real operation only rarely these difficult situations materialise, or the system operator is able to get around them by other means, and differences in nodal prices are not as large as expected according to the planner estimation.

There is an obvious regulatory implication of this empirical property of nodal prices: The use of nodal prices, in those power systems that have adopted them, does not suffice to recover the total costs of transmission. Therefore, transmission grid charges must be instituted to recover the rest of the total revenues allowed to the regulated transmission company. This charge is termed the *complementary charge* (CC) and is analysed in Sect. 6.5. In those (more numerous) power systems where nodal prices are not used, the complementary charge must cover the totality of the remuneration of the network.

It must be very clear by now that it would be a very serious regulatory blunder to leave the remuneration of the transmission activity to just the income from the nodal prices. This rule would generate a perverse incentive for the grid owner with respect to investment, maintenance and other tasks related to improving grid quality, since grid degeneration prompts a direct increase in network revenues. This is only true if the income from nodal prices is the only income of the transmission business or if it is independent of the rest of incomes. No such perverse incentive exists if total transmission income is determined a priori and nodal pricing is just a means of collecting a fraction of that total amount.

The annual complementary charge CC for the entire transmission network is thus defined as:

$$\mathrm{CC} = \mathrm{CA} - \mathrm{NR} \tag{6.4}$$

where

- CA is the total annual allowed revenue for the network,
- NR, or AVTR is the annual variable network revenue, and
- CC is the yearly additional network revenue to be recovered through transmission charges (see Sect. 6.5).

Other properties of nodal prices
Two other major properties of nodal prices, which may be useful in the evaluation and design of algorithms for the allocation of network costs and responsibilities to the users of the grid, are described in detail in Annex B of this chapter.

6.2.3 Computation of Nodal Prices

Nodal prices can be readily obtained as by-products of the models generally available either for optimising short-term generation dispatching by taking transmission grid effects into consideration, or for market clearing from generation and demand bids.

These models can be made as complex as developers wish, entering constraints at their discretion. Annex C in this chapter illustrates, in the simplest possible case, how nodal prices can be calculated both under centralised dispatching in an electricity market in which grid concerns are taken into consideration.

6.2.4 Nodal Energy Prices in Electricity Markets

Just summarising what has been already presented, a nodal price is the price paid or received for the energy consumed or generated at a given transmission node. Nodal prices implicitly include the effect of grid losses and transmission congestion, internalising both effects in a single value (monetary unit per kWh) that depends on each system node. They are, therefore, perfectly efficient signals for economic decisions concerning the short-term operation of generation and demand, since they correctly convey the economic impact of losses and constraints at all producer and consumer locations.

Examples of nodal pricing can be found in most South and Central American countries (including Chile—since 1981—, Argentina, Peru and the Central American regional electricity market, in some cases with strong simplifications), New Zealand and several regional transmission organisations (RTOs) or independent system operators (ISOs) in the USA, such as Pennsylvania, New Jersey, Maryland (PJM), New England, New York, ERCOT (in Texas) or California.

Many electricity markets do not use nodal prices. Single energy pricing is in place in almost all European countries, as well as in Colombia for example. In densely meshed transmission grids with no systematic or structural congestion, a

single energy price for the whole system may be preferred, and the transmission effects (losses and constraints) are addressed with a different method. The convenience of relying on a single electricity price is being increasingly challenged by the massive anticipated presence of renewable-based generation, whose siting should be guided by efficient locational economic signals).

Even when a single market price is adopted, there is no need to sacrifice the short-term loss and congestion signals that contribute to economically efficient system operation. The losses attributable to each player can be included as corrective factors in supply side bids, or preferably in the sums actually produced or generated, so that players internalise losses in the prices they bid. Where congestion does occur, limited capacity may be allotted by constraint management mechanisms that convey the appropriate economic signals. A number of options are available in this regard, including re-dispatching, countertrading and explicit or implicit transmission capacity auctions. These mechanisms are described in detail in Sect. 6.4.

An intermediate or hybrid approach, *zonal pricing*, consists of using a single market price except where significant grid constraints arise frequently between well-defined zones of the power system. While energy price differentials may be applied between zones, the same price prevails at all nodes within a given zone. This scheme can be viewed as a single price system to which zonal prices are added when needed for the purpose of grid constraint management. This method, also called "market splitting", is used in Italy, Nordel (Norway, Finland, Sweden, Denmark and Iceland) and MIBEL (Portugal and Spain) and also in the past in California.

Nodal pricing and network remuneration

As discussed before, the net revenues obtained from the application of nodal prices correspond to the economic value produced by the transmission grid by transporting power from nodes where electricity has a lower value (price) to those where its value is higher. Thus, the natural application of these net revenues is to cover a fraction of the regulated revenues to be earned by grid owner(s), even if they fall very short of recovering total costs. This is the usual practice in those power systems with nodal pricing.

As in nodal pricing, "zonal pricing" revenues should be allocated to cover allowed regulated grid revenues, since they are merely a simplified version of the nodal pricing scheme.

Authorities must bear in mind that the *total* revenues of transmission companies or the system operator should not depend on revenues resulting from the application of nodal prices. Otherwise, these companies will have a perverse incentive not to invest in further development and maintenance of the grid so as to increase nodal price differences and therefore their revenues. The remuneration of transmission service providers should generally be regulated (therefore not dependent on nodal price revenues), though any revenues resulting from the application of nodal pricing should probably be devoted to finance part of the payments to these companies. If this is the case, all revenues obtained for losses and constraints should be used to pay a fraction of the regulated transmission company's allowed

revenues, and transmission charges should be instituted to recover the total allowed transmission revenues, after discounting any loss or constraint management revenues.

6.3 Transmission Network Investment

This section starts the second part of the chapter, where the major regulatory topics in electricity transmission regulation will be presented. The first topic to be addressed is transmission network investment, which comprises network planning, investment decisions, business models for investors and siting issues.

6.3.1 Transmission Planning

The conclusion to be drawn from the restructuring experience accumulated in numerous markets is that transmission grid expansion depends essentially on the regulatory paradigm adopted for this business. The design of a framework for grid expansion entails the designation of the entity or entities responsible for planning the new grid investments, for authorising such investments and for building these new facilities and operating them. At the same time, a scheme must be devised for remunerating such investment, and economic signals must be established to encourage the entities responsible for these tasks to conduct them efficiently.

Transmission planning criteria

Regulation of transmission investment aims to guarantee that all network facilities that meet some specified set of criteria (intended to maximise social welfare, which includes economic and quality of service, among other considerations) are built at the right time and are duly operated and maintained at minimum cost.

Under the simplified viewpoint that all planning criteria could be somehow represented by some cost function, under traditional regulation, the basic optimal utility criterion for grid investment would be that "investments should be made to reduce electricity system costs, but only if the additional investment cost is lower than the additional savings".[11] This criterion is equivalent to the respective criterion in place in ideal competitive markets: "to maximise net aggregate benefits or social surplus, including the respective grid charges, for all players, generators and consumers." In ideal situations, the two objective functions converge, leading to optimal grid expansion [47].[12] A simple proof of this statement

[11] In general, transmission costs should include, besides the customary costs of investment and operation and maintenance, any additional costs derived from any negative environmental impact that transmission activities may cause.

[12] This is true if not only the positive but also the negative benefits, or losses, caused to some of the market agents by line construction are taken into consideration when deciding when such reinforcement is efficient or otherwise. Whether this should be so is somewhat controversial.

can be found in Annex D of this chapter. The estimation of costs and benefits is obviously subject to much uncertainty, in particular given the long economic lives of network assets.[13]

There is one important difference in transmission planning under traditional regulation and a competitive environment. In traditional regulation, generation expansion is also subject to centralised planning and it is usually considered to be an input to transmission expansion. Therefore, the power system costs to be reduced by the new transmission investments are just the generation operation costs. In a competitive market, the transmission planner does not have accurate information of the investment plans of the generation investors and must make decisions under more uncertainty. In addition, the decisions for new transmission will have in general some impact on future generation investments. This makes it more difficult to assess actual "generator benefits".

The direct economic benefits of a transmission investment, i.e. reduction of losses and of constraint-related generation costs (under traditional regulation), or increased operation margins for generators and lower payments by consumers (under market-based regulation), are well understood. Improvements of security of supply are more difficult to quantify. And transmission planners include more criteria when evaluating a transmission project [15, 28, 61]: market integration and increment in market competitiveness, emission savings, better exploitation of renewable energy sources and sustainability improvements, environmental impact, time to build, potential for social opposition, general feasibility of the project, legal aspects such as siting issues, better controllability of power flows, avoidance or postponement of other investments, improvement in the system dynamic behaviour, facilitation of distributed generation integration and more efficient reserve management and frequency regulation.

The so-called policy lines, whose main justification is to facilitate policy objectives, such as the integration of renewable sources, deserve particular attention [21] since they will probably dominate the transmission investments in the next decades [15]. It is reasonable to consider that energy policy criteria have priority over the rest. Consequently, transmission planning will seek to find the network expansion plan that maximises the benefit to cost ratio, while meeting any established policy criteria.

Transmission planning methodology

The phrase "transmission planning" refers to a recursive process of generation and evaluation of potential transmission expansion plans in search for a preferred solution that best meets the prescribed set of criteria. The high dimensionality of the search space, its high uncertainty, the lumpiness and longevity of the decision variables and the multiplicity of criteria typically characterise this process. In simple terms, the objective of transmission planning is to determine when and

[13] With the usual values of money discount rates and the growth of uncertainty with time, only the first decade or so has any quantitative importance.

where new transmission facilities should be built so that any prescribed criteria are met.

More precisely, rather than having to define a complete optimal plan, the objective of transmission planners is to define the transmission facilities that should be built now to create a robust system going forward, in the face of the strong prevailing uncertainty. This is why planning has to combine roughly two complementary approaches or time scales, see [11]: "strategic", that is, the exploration of how the future grid will look like in the long run—20 years from now, for instance,—and "tactical", where the interest is to identify the reinforcements that are consistent with the strategic plan and whose implementation process—environmental permits, acquisition of rights of way, etc.—must start immediately.

A realistic representation of the problem imposes exacting modelling requirements in several dimensions:

- A correct representation of the facilities in the interconnected system with a significant transport function.
- The complete geographical area that is relevant for the study must be considered. Note that, the larger the geographical footprint of the plan, the more opportunities that can be captured for efficient operation and resource utilisation.[14]
- The search for the optimal solution must combine somehow the bottom-up (incorporation of proposals made by local planning entities or stakeholders) and the top-down (fully integrated view) perspectives.
- Non-transmission solutions, such as storage or demand side management, should also be contemplated.
- Uncertainty in generation expansion, demand growth, fuel prices or policy measures must be adequately represented.
- Finally, the model should allow the evaluation of a "figure of merit"—either a scalar or multidimensional—that captures the desired set of criteria for the plan.

The search for the preferred plan can be formally posed as a mathematical optimisation problem, see [38]. However, the most frequent industry practice, see RealiseGrid [61], is trial and error, with evaluation of individual reinforcements or suites of lines for a prescribed ensemble of scenarios.

The present state-of-the-art of transmission planning has been able to address medium size power systems with moderate uncertainty; for example, a US regional transmission organisation (RTO) or a large European country. However, the state-of-the-art tools in transmission planning currently cannot cope with the entire EU system or the Eastern Interconnection in the US and the large

[14] For instance, the anticipated massive deployment of wind and solar generation in specific regions of North America, Europe or Northern Africa strongly requires the consideration of transmission expansion planning with Interconnection or EU-wide geographical scope, respectively.

uncertainty involved. There are preliminary efforts under way, both in the EU and the US: see [13, 15, 61]. To date, these efforts have been only able to gather bottom-up plans or to hypothesise and evaluate a few suites of lines. A more ambitious project, E-Highway 2050, has been launched by ENTSO-E in collaboration with several academic and industrial organisations to explore transmission expansion in Europe in the very long run.[15] A similar effort is under way for the US Western Interconnection.[16]

6.3.2 Transmission Network Business Models

A sound transmission policy should ensure that all beneficial lines are built. This requires that some company decide to build these transmission facilities with the expectation of receiving an attractive remuneration. This section takes the perspective of the investor and examines different business models. We shall focus on who makes the decision to build new transmission and how the incurred costs are recovered.

Transmission costs include investment costs (depreciation as well as a return on net fixed assets), costs of operation and maintenance of the network facilities and other administrative and corporate costs. Line losses and additional generating costs due to grid constraints are generating unit costs incurred for grid-related reasons, but they are not grid costs per se, as indicated before. Ancillary services are essentially provided by the generation business and should be viewed accordingly.

How transmission investors will recover their incurred costs depends directly on the adopted regulatory framework, which defines the corresponding business model. There are several basic approaches, with many possible variations on each one:

- Centralised network planning performed by a specialised institution: in this scheme, regulatory authorities must authorise an expansion plan, and investors (either restricted to one or more companies or open to any interested firm) receive a regulated remuneration that is established by regulators or in a competitive auction. The remuneration itself is recovered from regulated charges levied on the network users.
- Licensed company under a more or less traditional regulated monopoly scheme: the company is responsible for transmission network expansion, subject to some minimum performance criteria, with some kind of incentive-based remuneration that is established by the regulator and recovered from regulated charges levied on the network users.
- Transmission expansion is driven by the network users, by means of proposals to the regulator or consortia of users that are willing to finance and build the

[15] https://www.entsoe.eu/system-development/2050-electricity-highways/

[16] http://www.wecc.biz/

reinforcements. The regulator must authorise the expansion plans and determine the allocation of costs to the network users.
- Merchant investors, once authorised by the regulator, finance and build a transmission facility and charge to its users according to pre-established contributions or per actual use of the facility.

A more detailed discussion of the major regulatory approaches to transmission grid capacity expansion planning and investment follows [10, 54, 61].

Supervised centralised planning and regulated remuneration

The most common regulatory approach is supervised centralised planning, in which the respective governmental authority or regulatory body periodically calls upon a specialised agency (the vertically integrated utility in traditional regulation, the system operator under market-oriented regulation[17]) to perform this task. Pre-established criteria for selecting the best alternatives must be followed, although the proposals put forward by market players have to be taken also into consideration (and denials justified). Regulatory bodies examine the plan and authorise new facility construction, operation and maintenance. These new facilities are included in the regulatory asset base and remunerated at a specific rate of return and depreciated throughout their service life, as in traditional cost of service regulation. Nevertheless, incentives can also be defined to enhance transmission company efficiency in terms of operating costs and commercial services.

Under centralised planning, it is possible to organise a *competitive tendering* process to build and operate the new installations required, once the respective investment decisions have been adopted. Then, each new transmission installation is awarded by competitive tendering, although in certain special cases, direct designation may be more advisable. A number of transmission companies may, then, be operating. Under competitive tendering, the acknowledged cost is the quotation submitted by the awardees, which during the operation of the installations may be subject to penalties or incentives depending on their actual availability. When the contract expires, operation and maintenance of the existing facility is tendered for an additional term.

With this regulatory scheme the transmission network owner can be termed as "*passive*", since it is not responsible for final network investment decisions, even if it is responsible for the expansion proposals. In this case, the remuneration should be based on the actual grid infrastructure and penalties and any incentives should be associated solely with the actual availability of each facility, but not with the performance of the entire network or the operation of the power system.

The system operator is the obvious choice for the institution to be in charge of preparing and proposing the network expansion plans, because of its intimate knowledge of the functioning of the power system and the transmission network in particular. However, even if the system operator does not own the grid, it has a clear incentive to further and accept the construction of a large number of grid

[17] The operator may or may not be integrated in the transmission company.

assets, since it is responsible for system security in one way or another.[18] Even though the final decision on new line construction is incumbent upon the regulator, this less technically minded institution may find it difficult to reject investments proposed by the operator, for fear that their lack might have an adverse impact on system reliability. The resulting grid may, then, tend to be larger than optimally required from the standpoint of economic efficiency.

If the system operator also owns the transmission grid, it might have an additional incentive to develop more lines and facilities than necessary. The more it invests, the larger the number of assets for which it may receive generous (if this is the case) and guaranteed remuneration.

Despite these potential problems, most authors dealing with this issue deem the traditional centralised grid expansion scheme to be the most effective way to obtain a sufficiently developed grid. See Joskow and Tirole [32], Newbery [45], Dismukes et al. [12] and Pérez-Arriaga et al. [58].

Traditionally regulated monopoly

This approach consists of awarding the transmission license to a single company, which is subject to some sort of incentive-based monopoly regulation.[19] This transmission company is also the system operator and typically assumes the following obligations: (a) to meet certain pre-established grid design and user quality of service standards and to enlarge grid facilities to be able to continue to meet standards; (b) to inform users of foreseeable congestion or "surplus capacity" at grid access nodes within a reasonable time horizon.

This "*active*" regulated transmission company is responsible for making investment decisions and its remuneration should be based on a "well-adapted" or efficient grid design and efficient operating cost criteria. In addition, economic penalties or incentives—incentive-based regulation schemes—associated with actual quality of supply levels—related to the transmission network and system operation—are recommended. In particular, RPI-X revenue or price caps can be applied (see Chaps. 4 and 5 for the basics of incentive-based regulation and how it is implemented).

Under the RPI-X revenue cap method, the annual regulated revenues may not be raised in the following period (typically 4 or 5 years) by more than the retail price index (RPI), minus an adjustment factor (X). The regulator must estimate the transmission company's total costs during the period and apply any efficiency adjustments that it deems suitable, along with inflation. The adjustment factor is the value that equates the present value of revenues to the present value of the estimated costs for the entire period (see Chaps. 4 and 5 for the definition and application of this factor). In the following price control period, the regulator

[18] The regulator is likewise indirectly responsible for system safety. Consequently, it should be also biased towards more system security than optimally needed. Account must be taken of the fact, however, that it is the regulator's obligation to secure system efficiency.

[19] This regulatory approach is similar to the scheme adopted in the UK with its *National Grid Company* and is the method most used to regulate distribution, as explained in Chap. 5.

studies the investments actually made by the licensee and its performance level when establishing company remuneration. Since investments in new transmission lines can be individually evaluated, the remuneration assessment for each period should explicitly address the investments planned, the depreciation on existing assets, the licensed company's financial status and any expected improvements in efficiency.

The drawback to this procedure is that it cannot guarantee optimal grid development, but a design that is subject only to compliance with minimum standards. Joskow [30] proposed that efficiency incentive design should hold this company responsible for re-dispatching costs over a certain ceiling or for the costs associated with system collapse. Such problems cannot always be attributed to the transmission owner or the system operator, however. Under such schemes, the company would be subject to very high risks that would not be proportional to its potential revenues.

Market player initiative with regulatory supervision

Market player initiative is a third standard approach[20] that leaves the initiative for grid reinforcement to its users, who can weigh their estimated charges derived from the investment costs involved against the expected benefits for them, including such factors as readier access, improved market prices, less congestion or lower line losses. Any valid coalition of users must be such that at least a minimum percentage of the regulated network charges derived from the proposed expansion must correspond to the members of the coalition. The regulator evaluates the public utility of the proposed reinforcement against pre-established criteria and accepts or rejects it. If the proposal is accepted, the regulator organises a tender for building and maintenance. The transmission company awarded the tender is remunerated according to the terms of its bid, while operation is left to the system operator. Transmission fees should be charged to grid users so that they reflect the benefit, or at least the use, that system agents derive from the grid.

A variation on this approach involves holding the coalition of users responsible for financing and building the investment, allowing the coalition to recover some costs via regulated transmission charges paid by other line users. In other implementations, the promoting coalition of users may receive the economic rights to the congestion rents associated with reinforcement (see Sect. 6.5.3 below).

This approach is more market oriented than the preceding two, although it involves rather complex administrative procedures and rests essentially on the existence of the right short- and long-term grid price signals as the mechanism for ensuring that players are incentivised to optimally reinforce the grid. One weak point of the method is that agents, on their own initiative, are unlikely to propose line construction where benefits are widely distributed or difficult to translate into economic terms, or if the margin of benefits over costs is not large [2, 9, 58].

[20] This is similar to one of the possible alternatives provided for in the Argentinean regulatory approach [39, 40]. It has been recently adopted by the New York ISO.

This is the case of lines designed to improve overall system reliability. This scheme should therefore be supplemented with a centralised back-up planning mechanism to ensure that such lines are also built.

Another alternative, geared to enabling grid users to participate in its development, is to organise long-term auctions in which agents bid for the right to use transmission capacity yet to be built. The academic community, has, however, identified major drawbacks to the exclusive use of long-term transmission capacity auctions to enlarge the grid [24, 42, 45].

Merchant lines

The transmission network regulations in many power systems allow the participation of merchant investors. Some systems have even tried to allow transmission to develop like a normal competitive business.[21] So-called "merchant lines", built under private initiative, sell their capacity to the highest bidder on the spot market or under long-term contracts.[22]

We have to distinguish two basic business models for transmission merchant lines. In the first one, a line obtains its revenues by arbitraging in the short term between the electricity prices at the end nodes of the line, i.e. by buying cheap on one end and selling more expensive at the other end. One general difficulty with this method is that only power transmission lines that do not eliminate or significantly reduce previously existing congestion are economically viable under this investment regime, since, once congestion disappears, congestion rents and therefore the business opportunity, vanish with it as the electricity prices tend to equalise [31, 56]. In addition, revenues from merchant type lines are subject to many risks, associated primarily with the construction of new regulated lines that may substantially reduce the congestion from which they benefit [6, 31]. Also, given the discrete nature of investment in new lines, building a merchant line whose capacity is less than efficient from the standpoint of the system as a whole could favour the appearance of market power. More than that, as explained in previous sections, building new lines may detract from the existing network transmission capacity. The outcome, higher congestion rents, would constitute a perverse incentive for such investors. Therefore, any merchant line must be subject to regulatory authorisation.

As we know, nodal pricing grossly under recovers the total cost of any well-developed transmission grid. Therefore, leaving network investment *entirely* to individual agents that try to make a profit from building transmission facilities and buying and selling energy at nodal prices would result in a hugely underdeveloped transmission grid or in bankrupt investors. On the other hand, it is always good for the system to leave open the possibility that some investor might find an attractive

[21] This was the case of the Standard Market Design proposed by FERC for the US in 2002.

[22] A very limited number of lines have been built under this regulatory approach to date: a few submarine cables in Europe, a couple of lines in Australia (which later returned to the regulated regime) and one line in the USA.

opportunity that was overlooked by the centralised planning process. But a frequent occurrence of merchant lines in a system should not be considered as good news, as merchant lines will be, in general, more expensive for the consumer than regulated lines, since the margin for profit is larger in the former.

The second business model for merchant transmission lines makes more sense in a less restricted number of situations [10]. In this case, the potential investor in a line agrees with a sufficiently large group of the potential beneficiaries of the line that they will cover the complete line costs. Once financing is secured, the promoter will sign long-term contracts with these line beneficiaries and will build the line. A major difficulty in this approach is the free-rider problem—i.e. those who benefit but expect other beneficiaries to shoulder the costs—and also the dispersion of beneficiaries, consumers in particular, that may render the search for contributors an insurmountable task. Only those investment projects with few well-defined beneficiaries and a large margin of benefits over costs have the potential to become merchant lines. But centralised planning will also likely identify them. Merchant investment appears to be more likely and also more useful when it focuses on singular projects that centralised planning has not identified or considers too risky.

Conclusion

In conclusion, although all these business models should be allowed to co-exist and each one of them may contribute to a comprehensive transmission expansion, it must be clear in any sound transmission policy that most lines should be built under regulated conditions, with the costs being allocated to the network users by regulated rates. Therefore, once it has been decided—based on a sound planning procedure—to build a new line, the main role of the regulator is to make sure that the line is built; therefore trying to reduce the risk of cost recovery of the investor as much as possible. If transmission planning has followed a well designed and transparent process, the risk of building non-beneficial lines is minimised. And the negative consequences of under investing in transmission are far greater.

6.3.3 Siting

Siting of transmission facilities has become a thorny problem in most developed countries, since transmission facilities in general cause inconvenience and do not provide any direct benefits for those who live in the environs. Siting is a less technical and more institutional issue, but it is interdependent with planning and cost allocation and adds another challenge to network expansion. Siting requires the proper consideration of the local concerns of those who will be affected by the presence of transmission lines, together with the objective of implementing a project that has been found beneficial for society. When several jurisdictions exist (local, province or state or autonomous region, supra-national or federal) it is

necessary to delimit responsibilities and to make sure that there is a clear decision-making procedure where all stakeholders are somehow represented.

It is expected that siting will become easier to address once satisfactory solutions are found for planning criteria, planning, cost allocation, investment and cost recovery. At least siting will be reduced to what really is, and no more.

6.4 Transmission Cost Allocation

6.4.1 Fundamentals

Electric power transmission is nearly universally regarded, also in liberalised systems, to be a natural monopoly. Therefore, transmission should be a regulated business. The economic regulation of network monopolies is discussed in Chap. 4. Under both traditional cost of service and incentive regulation, the annual revenues that the regulated transmission company is allowed are established by the regulator. These revenues are paid for by network users in the form of transmission grid tariffs or charges. This section focuses on methods for allocating the allowed revenues among network users in ways that further efficiency in the short term (encouraging agents to make optimal operation decisions), and in the long term (driving agents' decisions on the location of new generators and loads).

All transmission business costs must be recovered from the network users. They comprise the network investment costs (asset depreciation as well as a return on net fixed assets), operation and maintenance costs and other administrative and corporate costs related to transmission. Line losses and extra operation costs due to grid constraints are generation costs; therefore, these costs, as well as the system operator costs and those other costs related to the provision of Ancillary Services (most of these are generation costs) should be levied on system users through other charges.

The presence of the transmission grid, with its losses and constraints, leads to differences in energy prices from one node to another and sends site-related economic signals to grid users. Energy price differences result in some partial recovery of the total allowed revenues of the regulated transmission company when applied to power injections and withdrawals.[23] However, as Rubio and Pérez-Arriaga show [65], the net revenue resulting from the application of nodal prices amounts only to a small fraction of the total regulated cost of the grid. The fraction of regulated transmission revenues recovered from the application of energy prices is normally referred to as *Variable Network Revenues* (VNR), see Sect. 6.2.2 in this chapter.

[23] Revenues from the application of nodal prices comprise those obtained in the day-ahead and other short-term markets as well as those that happen well ahead of real time through the sale of Financial (or Physical) Transmission Rights over the capacity of likely to be congested corridors, or hedging differences in prices among different nodes, see Sect. 5.5.

Therefore, completing the recovery of the cost of the grid requires applying additional charges, normally called *Complementary Charges* (CC), which are equal to the fraction of the grid cost not recovered through energy prices. It is important that the total remuneration received by the transmission activity is independent from the actual value of VNR to avoid any perverse incentives to increase losses or congestions from the transmission owners. Complementary charges should also send economic signals to agents encouraging them to reduce the cost of expansion of the grid. This can be achieved if these charges incentivise agents to install new generation or load in those locations where reinforcements needed for the grid to cope with the resulting incremental flows are least costly. Additionally, complementary charges should be compatible with the application of efficient short-term economic signals, i.e. these charges should perturb nodal prices as little as possible, so as not to compromise the efficiency of system operation.

Transmission charges can be divided into connection charges and use of the system (UoS) charges. Connection charges are employed to allocate the cost of transmission facilities directly connecting large network users—large generators or industrial consumers and high-voltage distribution substations—to the existing transmission facilities. Each individual user being connected typically pays for its connection installations. Connection charges are therefore separated from UoS charges, which must cover the costs of the rest of the transmission facilities.[24]

The remainder of this section is devoted to the discussion of the design of UoS charges. Two phases in the design are clearly differentiated: (a) the allocation method employed to determine which fraction of the grid (i.e. "how much") should be paid by each network user and (b) the design of the most adequate format (i.e. "how") to apply these transmission charges. Both topics are discussed next.

6.4.2 Allocation of the Cost of the Main Grid to its Users[25]

6.4.2.1 The Principles

The allocation of the cost of a transmission network among its users must obey some basic principles that result from the combination of microeconomic theory, power systems engineering and sound regulatory practice [55]. After much trial and error in multiple power systems, some solid principles have been learned,

[24] The distinction is not always clear. Some connection links are very long and perhaps should be considered part of the common transmission infrastructure, especially if there is the possibility that other network users might benefit of connecting to this link in the future and the connection capacity is designed with this purpose.

[25] This section borrows concepts and some texts from the material used in the elaboration of the MIT Study "The future of the electric grid" [44], where one of the authors of this chapter was a participant.

although they are far from being widely applied [44, 49]. Only four high level principles suffice to guide the praxis of transmission pricing.

Principle 1: Allocate costs in proportion to benefits

"Cost causality", i.e. the responsibility of each network user in grid investment, should be the conceptual basis of any cost allocation methodology, even though it is difficult to implement. Cost causality is equivalent to "beneficiary pays" because transmission is built when the aggregated benefits of the additional investment to the network users exceeds the incurred costs. This implies that, in principle, both generators and consumers should pay because both (in general) benefit from the expansion of the transmission network.

A transmission project, which may consist of multiple facilities, is economically justified if its benefits—of all kinds—exceed its costs. By reducing or eliminating price differences, however, a transmission project could impose losses on generators in previously high-price areas or on load in previously low-price areas. In addition, some entities might suffer losses because of environmental harm. Regulators should approve projects with positive net benefits (i.e., net of any losses), even if they impose losses on some entities. Note that, as with any competitive system, there is no obligation to protect the losers from changes in the market values of their assets that result from new system conditions.

Dividing network costs among network users in proportion to their benefits strikes most people as equitable. If this is done for a project with benefits that exceed costs, all beneficiaries will end up being better off and are less likely to oppose moving forward with the project.[26] Conversely, if a project's costs exceed its benefits, it will be impossible to allocate costs in such a way as to make all entities better off. Thus, adopting the beneficiary pays principle helps with decisions about what should be built, as well as determining who should pay for what is built. Fairness is important, but support of consistent incentives for investments is the key reason for embracing this principle. When the benefits of a transmission project are very widely distributed, it might be reasonable to "socialize", i.e. to apply a flat charge to recover the project costs. Note that this does not mean abandoning the principle of beneficiary pays. Cost socialisation is in this case a good approximation to allocation to beneficiaries.

Should generators be charged transmission costs? Is it not true that "consumers end up paying all the costs", anyway? In wholesale electricity markets, both generation and load can benefit from new transmission capacity. Generators will be able to deliver their product at better prices, while consumers will experience

[26] In principle, if the project has positive net benefits, it is possible to compensate losers for their losses and make all affected entities better off. In practice, this is complicated and seldom if ever done. One could argue that compensation is not deserved for the loss of economic benefits (high prices to generators, low prices to loads) that exist only because of network congestion, but major environmental impacts may raise more serious issues in the future. Such impacts might be claimed, for instance, if a proposed line would cross a particular area but confer no benefits on its residents.

reduced energy costs and/or increased reliability. Cost allocation procedures should seek to apportion the costs of a line to generation and load proportional to aggregate economic benefits realised by the two groups. We can distinguish two extreme cases that illustrate the futility of any aprioristic or universal criterion to allocate transmission costs to generators or consumers. On one extreme, in highly competitive environments with no special opportunities for any generator to capture extra rents, all costs levied on generators will end up being passed on to load via wholesale electricity prices, either in the short term (if transmission charges are levied on a per MWh of produced energy then the generators in a competitive market will internalise the charges in their bids to the short-term market) or in the long term (if network charges are levied as an annual lump sum or on a per MW basis then potential new generators will delay investment until wholesale prices reach an adequate higher level to make the investment attractive). On the other extreme, some generators may enjoy unique location-specific (a very propitious hydro site or a cheap local gas field) or other advantages that restrict competition. Consequently, these advantageous generators will retain the benefits derived from transmission that is built to these locations. Not all generators operate in highly competitive environments, and moreover changing market conditions typically provide multiple opportunities to generators to enjoy short-term rents (and suffer short-term losses), so these generators can be charged transmission costs without any anticipated pass-through to consumers. Therefore, it cannot be argued that only consumers should be charged network costs, since "in the end consumers pay for all costs". On top of this, transmission charges are useful to send locational signals for siting of new generation facilities, so that total network investment costs can be minimised. This is particularly important for renewable generation, which frequently require costly transmission investments and has multiple possibilities for siting. A subtle regulatory issue is that any regulated subsidy to renewables must take into account the regulated transmission charge to be established, so that renewable generation is economically viable in the best locations under both a network viewpoint and an optimal utilisation of natural resources.

If regulation fails to allocate costs according to benefits, the consequences could be to discourage the realisation of a generation investment that otherwise would had been economically attractive and would had been beneficial for the system as a whole. This is clearly shown in the simple case of Fig. 6.13.

There is still an important practical issue worth mentioning. Lumpiness and economies of scale result in significant overinvestment (even if fully justified), particularly during the first years after the construction of a new transmission facility. Therefore, the entire cost of the asset should not be charged to the current beneficiaries, who may only need a relatively minor fraction of the entire capacity. Olmos and Pérez-Arriaga [49] point out that the loading level of each transmission line should be considered when determining the transmission charges applied to each network user. The fraction of the total cost of a line to be allocated to agents

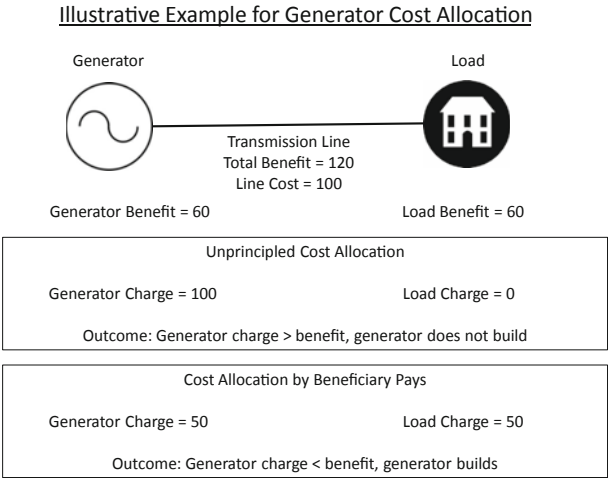


Fig. 6.13 Illustrative example of potential cost allocation outcomes to generators [44]

according to their responsibility in the construction of the line should, for instance, be limited to the ratio of the loading rate of the line to that of other similar lines in the system. The remainder of the cost of this line should probably be socialised, since current users of the grid cannot be deemed responsible for the construction of the fraction of the capacity of a line that is expected not to be used until long time in the future (for lines that are underutilised in the present).

Principle 2: Transmission charges should not depend on commercial transactions

Transmission charges should depend on the location of the users in the network and on the temporal patterns of power injection (for generators) and withdrawal (for loads). However, transmission charges should not depend on commercial transactions that occur between users (that is, who trades with whom). Transmission charges should be levied on those who benefit from the existence of any given transmission facility, regardless of any trading relationships.

This principle was already introduced in Sect. 6.1.3 and, because of its (admittedly) counterintuitive content, will be proved here with a mental experiment. Imagine a power system with M generators and N loads in a competitive wholesale market environment. They are free to trade with one another. Assume that all of them are given at a particular instant of time complete information about the bids in the next hour—i.e. each generator and load knows the exact quantities and prices that everyone else is willing to buy and sell at. These market agents are then allowed some interval of time to engage in a transaction. Given that all of them have complete information about what is being offered, it is clear that all the best supply offers will be taken by the demand and only the more expensive generators that were not necessary to meet all the demand will not be scheduled to produce. The set of accepted generators and loads will result in a certain pattern of

flows in the network.[27] If the experiment is repeated 1 million times, the resulting bilateral trades will be all different, since each time the individual generators or loads may try to buy or sell from or to a different partner. But in the end, the same loads will be supplied and the same set of least expensive generators will be taken; therefore resulting in exactly the same pattern of network flows each time. We can also argue that, with an appropriate nodal pricing scheme, the nodal prices and therefore the economic benefits for each network user would be identical in these million cases. Therefore, there is no reason to apply network charges that discriminate according to any existing commercial transactions. Instead, transmission charges should depend on the location of the network users within the system topology and on the temporal patterns of power injection and withdrawal [55].

This second principle means that a generator located in a region A that trades with a load serving entity in a region B, such that physical network connection exists between A and B, should pay the same transmission charge as if, instead, it were contracted to supply a neighbouring load sited within its own A region. And conversely for the load in region B, which could purchase the electricity from a generator in regions B or A. This principle follows from what is called "the single system paradigm": if there is open network access and no barriers to inter-regional trade, the decentralised interaction between the regions and their agents should approximate the ideal outcome of an inter-regional efficient generation dispatch, regardless of who trades with whom, inside or outside each region. The independence of the transmission charges from the commercial transactions directly follows. The application of this principle should not be affected by the existence of any contracts voluntarily signed by any agents, since they should modify neither the physical real-time efficient dispatch of generation nor the demand. This second principle is not tantamount to socialisation of network costs; as indicated before, transmission charges should depend on the location and the timing of network utilisation.

Failure to make transmission charges independent from commercial transactions can result in "pancaking" or piling up of transmission charges, where network users are required to pay accumulating fees for every region through which their power is deemed by contract to pass between the buyer and seller, regardless of actual power flows. Pancaking makes transmission charges depend on the number of administrative borders between buyer and seller. Such pricing tends to stifle trade and to prevent buyers from accessing low cost sellers. The resulting perverse incentives could lead to inefficient transmission investments and would significantly complicate operations in networks. Pancaking has been recognised in both the US and EU as undesirable; this recognition led to FERC Order No. 888 (Open Access) in the United States and to a system in the EU that provides a standardised mechanism for accessing and paying for the transmission system

[27] The matching of bids could be done in such a way that it would guarantee a priori that no network constraints would be violated. The effect of losses on the competitiveness of the different generators could be also accounted for in a bid matching process that would ideally emulate an optimal power flow (see Annex C of this chapter).

within the EU [19]. This principle is generally adhered to in US practice within ISOs, but not in inter-ISO transactions.

Principle 3: Transmission charges should be established ex ante
Transmission network charges for new network users should be determined ex ante and not updated, or at least not for a reasonably long time. This is the only way to send the predictable—although not necessarily uniform—economic locational signals that investors need to choose the most convenient sites with a low financial risk. This is of particular interest for wind and solar generators, which usually have many potential installation sites.

The locational impact of transmission charges is mostly meant to incentivise potential new generators to site at convenient places from the transmission network viewpoint, i.e. where the presence of the new generator will reduce (or, at least, not increase) the need for network reinforcements. Transmission charges may also have some impact on the retirement decision for old plants with scant operation profit margins. No significant impact is expected on the siting decisions of consumers, since transmission is a minor component of the total electricity payment, which normally is not a major ingredient of the consumers' budget. And, this is the important point we want to make here, no locational impact is possible for new generation investments once they are into the construction period or in operation.

What is proposed here is that, when a new generator requests connection to a certain point of the grid, the system operator should provide the transmission charges to be levied to this generator for the next 10 years (or a similar figure). For typical monetary discount rates, and given the uncertainty surrounding most of the major factors that affect the profitability of the power plant, 10 years should be enough for the new generator to decide whether to invest or not.

The trajectories of transmission charges on potential new generators applying for connection at the beginning of a year T, must not be changed during the following 10 years *for these generators*. However, in the presence of additional information during year T, the trajectory of network charges that will be announced at the beginning of year T + 1 and applied to any new entrants during this year will be a different one, in general.

Principle 4: The format of the transmission charges matters
In the design of transmission charges one must clearly differentiate the determination of *how much* each network user has to pay from the specific format of implementation of the charge, i.e. as a volumetric charge (€/MWh), as a capacity charge (€/MW), as a lump sum (€) or as a combination of them. The choice of format for the transmission charges has implications on the short- and long-term behaviour of the agents of the market. For instance, a volumetric transmission charge (€/MWh) to generators becomes an additional component of their variable production costs; therefore distorting their efficient competitive position in the wholesale market and increasing the short-term marginal price of electricity. On the other hand, a capacity charge (€/MW) to generators will have to be considered

as an additional fixed cost for investors in new generation facilities. This topic is discussed in detail in the next Sect. 6.4.3.

6.4.2.2 The Implementation

How could these principles be implemented in an actual power system? How could one assign the responsibility of individual generators and loads in the construction of transmission facilities? Or, in other words, how could one discover the beneficiaries of any given transmission investment? The answer to these questions must necessarily be found in the investment process, since what justifies the investment in a new network facility is the fact that its cost is less than the aggregated benefits for the network users. The "beneficiary pays" approach to transmission pricing was adopted for the first time in 1992 in the development of the regulatory compact associated with the liberalisation and restructuring of the Argentinean power sector.[28] At the same time, the UK adopted a transmission pricing method – Investment Cost Related Pricing, ICRP—that was based on the impact that every generator and load would had on some streamlined transmission expansion model. ICRP is still in use today, although it is also subject to revision [46]. Both conceptual approaches are dual views of the same underlying principle: transmission expansion is required because of the growth of connected generation and demand and, at the same time, it is justified because the aggregated benefits for that generation and demand (as well as the benefits of existing agents) exceed the costs of transmission investment. The same approach was also contemplated in the early 1990s to address the allocation of the cost of a transmission line—the SIEPAC project[29]—that should join six Central American countries bringing major benefits to the region. A rigorous formulation of the connection between transmission expansion planning and cost allocation to beneficiaries can be found in [28].

In practice, the application of these sound concepts has proven to be more difficult than expected. Cost allocation to beneficiaries can be approached as a subproduct of the network expansion planning process: as planning requires the comparison of network costs and the aggregated benefits for the system agents for a long period of time, the planning process must contain the information regarding the economic beneficiaries of the considered transmission project. This may be complex, but it is a promising and economically sound approach.

---

[28] One of the authors of this book, Ignacio Pérez-Arriaga, proposed this approach to the Argentinean authorities and was a component of the team that developed the corresponding legislation.

[29] More information in the website: http://www.crie.org.gt/index.php/principal. Some of the authors of this book (Tomás Gómez, Ignacio Pérez-Arriaga, Andrés Ramos and Michel Rivier), as well as other researchers of the Institute for Research in Technology at Comillas University, were involved in multiple aspects of this project, including the design of the cost allocation procedure of the transmission system.

It is necessary to establish a distinction between new transmission facilities and those already in operation for some time. The method, with its complexities, can be applied to new or recent investments. Additional and basically insurmountable difficulties appear when we look for the beneficiaries of a network asset that entered in operation many years ago. In this case, the obvious "without the line" counterfactual, to be compared with the present situation "with the line", may not provide any valuable information for cost allocation purposes. Two typical situations have been found in practice. In the first one, eliminating the considered line means chaos in the power system, resulting in some permanent unmet demand. Obviously, this would never happen in an actual power system and the "without the line" counterfactual is nonsensical, as other measures would have been taken, in absence of the line, to supply the entire demand. In the second situation the removal of a line has basically no economic impact on the network users when the power system has enough built-in redundancy, and only sophisticated reliability analysis could detect any differences. Most cases lie in between these two extremes.[30]

As a consequence, strict cost causality or cost allocation to beneficiaries has been tried and actually employed in only a few power systems. In most cases an "engineering way of thinking" has prevailed: those who "use" the transmission facilities should pay for them. One can think of "use" as a proxy for "benefit" or "responsibility in the investment", but there is not a clear nexus between them that can be proven. If the method employed to measure "network use" appears to be reasonable, one might expect a strong correlation between the outcome of both approaches.

But, what is a reasonable method to measure network utilisation? Unfortunately, computing the electrical utilisation of lines by agents is not a simple task either, since there is no indisputable method to do it. Several approaches to determine the use of the network by agents have been proposed and applied, with results that vary significantly from one another for reasons that will be explained later. It is important to keep in mind that the final objective is not computing the use of the network by each agent, but determining the responsibility of this agent in the construction of the line.

What remains of this section contains a brief description of some of the best-known methods for allocating transmission grid costs. The fact that they are described here does not mean that they are recommended necessarily. In fact, most of them are not recommended at all.

[30] The situation is not better with the ICRP method as applied in the UK. The evaluation of the impact on a well-adapted transmission network of adding or removing generators and loads is plagued with difficulties due to the lumpiness of the network investments, the strong economies of scale and the redundancy in the network design due to reliability criteria. See [46] for details.

A. Methods without locational components

Current approaches to transmission cost allocation in most countries do not include any locational components. This is not desirable in general, particularly with the current trend towards deployment of wind and solar generation in a multiplicity of locations and with the integration of power systems into larger regional or multinational grids.

Postage stamp

This very popular method consists of covering the total transmission costs by applying a uniform rate to all network users based on some simple measure of transmission utilisation. In practice, this results in a uniform charge per MW connected or per MWh injected into or withdrawn from the network to recover the allowed regulated revenues. This is the method used in most European countries, by many US electric utilities and many countries in the world. The charge is frequently applied only to consumers. Sometimes different charges are computed for consumers and generators, based on a priori breakdown of the total transmission cost into two—more or less arbitrary—fractions.

The name "postage stamp" refers to the fact that the transmission charge is wholly unrelated to the place where power is injected or retrieved. These charges entail no geographic discrimination that can convey suitable locational signals to steer individualised decision making towards the best interests of the system as a whole. This simple method may only be recommended when grid characteristics require no further sophistication, i.e. densely meshed grids with low demand and generation growth and requiring no major reinforcements.

Ramsey pricing

Some electricity systems have well-developed grids that would not benefit greatly from locational signals and do not need additional grid investments that could potentially benefit some users much more than others. In such systems, providing that short-term loss and congestion signals are conveyed in one way or another, the primary aim of grid cost allocation should be to interfere as little as possible with the market and investment decisions that players would make if there were nothing to allocate. Under these circumstances, second best or Ramsey criteria may provide a valid justification to some cost allocation decisions. Ramsey pricing aims to allocate most of the grid costs to users that are least elastic to the resulting charge. Under conditions of perfect competition (which rarely occur in real markets), generator elasticity to additional charges is very high, as explained before, and any network charges to generators would ultimately fall on the consumers. Given, then, a sufficiently competitive market, transmission network costs should be allocated primarily to consumers, charging the least elastic demand (typically domestic consumers, in more developed countries, at least) more than the most elastic demand (typically large industrial consumers).

Some countries apply Ramsey pricing criteria, more or less explicitly, in the design of their slightly more complex postage stamp transmission network tariffs.

B. Network utilisation methods

As indicated previously, there is no unquestioned way of quantifying unambiguously the utilisation of a transmission network by its users. As the results obtained may depend largely on the method adopted, the choice of one or another is not immaterial. The technical literature contains a plethora of approaches that have been proposed or actually applied. Here, a selection has been made that contains some of the most popular ones: average participations, contract path, MW-km or MW-mile and marginal participations. Just a few of them seem to be able to reasonably approximate the principle one of cost allocation, while it is easy to find case examples indicating flagrant flaws in most of the others. Detailed critical evaluations can be found in Pérez-Arriaga et al. [57].

Contract path

This is the most rudimentary of the network utilisation methods, and one that has been widely used in the past. It clearly violates principle number 2. In the contract path method, the cost of a given transmission grid service is calculated from the path that the energy has to follow from the point of injection to the point of withdrawal (contract path). The route taken by electric power between these two points is determined by mutual agreement among the parties. In other words, buyer, seller and transmission company agree to the most "logical" path that energy should follow over the grid for the intents and purposes of establishing grid charges. Such charges are then determined as a fraction of the cost of the lines where the transaction "flows".

This method was developed in the context of the bilateral trade—known as "wheeling"—that preceded the creation of organised wholesale markets, an arrangement where two typically vertically integrated utilities agreed upon a transaction that crossed a third company's grid.

When applied in a multiple system context this method also consists of "determining" the group of transmission systems that every transaction must cross to reach its destination. This usually leads to payment of a toll for each system crossed. The result is known as "pancaking", i.e. the toll paid is the result of summing or piling up the tolls stipulated by each of the crossed systems. That the result depends on how the boundaries between systems are defined is an indication that this approach, which obviously discourages energy trading, is essentially flawed.

MW-mile

The MW-mile (obviously, also MW-km) is one of the earliest methods that attempted to provide a more precise measure of grid usage, taking not only the MW transmitted but the length of each line "used", into consideration, for instance. The underlying principle is that it should not cost the same to transmit 10 MW over 100 km as it does over 10: grid use is not the same and this should be factored into the transmission charge. Again, this is violation of the second principle of cost allocation.

The method first defines a baseline case with a load flow regarded as representative of system operation and including all the transactions to be analysed.

The resulting transmission flow in MW in each line is multiplied by the length of the line in km, and all the resulting products are summed. This gives the total MW-km associated with the baseline case. One of the transactions is then eliminated and the new load flow is computed. The total MW-km is then found for the new situation. The difference between the two sums is the amount of MW-km attributable to the transaction eliminated from the second calculation. The sum to be paid for the transaction is found by multiplying grid costs in the baseline case times the fraction, in MW-km, that the transaction represents of the total, likewise in the baseline case. This is repeated for all transactions.

This method has been and continues to be widely used. However, it is based on a mistaken assumption, since transmission charges should not be made to depend upon commercial transactions.

Marginal participations

The marginal participation (MP) method distributes line use, and consequently cost, on the basis of the marginal effect that each consumer and generator has on line flows. The effect on the grid is obtained by calculating the variation in all flows when a user's consumption or production is increased by 1 MW. This is repeated for each network user and for each one of the representative scenarios that are considered. The variation in the flow obtained for each line, player and scenario considered serves as a basis for calculating a value that provides a measure of electricity system use. This value is the sum of the products of the variation in flow in each scenario times the power consumed or generated by the user in question, times the duration in hours of the scenario. The sum of the effects found for a given player on all the grid facilities is divided by the sum of the impact of all users on those facilities to find the proportion of the grid cost to be paid by the player.

All this seems very reasonable, but there is an underlying assumption that, in the opinion of the authors of this chapter, renders the method useless. In order to calculate the marginal effect caused by each user, a slack bus or balance node that responds to the increases in generation or demand must be defined, since the balance of generation and demand in the grid has to be maintained at all times. The choice of the location of the slack bus in the system conditions the absolute results obtained for each player, although not the relative figures, which remain constant regardless of the slack bus chosen (see property 5 of nodal prices in Annex B of this chapter) [72]. However, the basic underlying assumption of this method, namely, that *all incremental changes must be referred to a single slack node*, *whichever it is,* is not justified. Why a single node and not seven? Why should all changes should be referred to the same node? Some proponents of the method claim that the slack bus problem disappears if a "distributed slack bus" is defined, whereby the 1 MW injected by a generator (or withdrawn by a load) sinks (or comes from) all the demands in the system, in proportion to the respective demands. The claim cannot be accepted, precisely because of the above mentioned property 5 of nodal prices: If choosing any single node as slack does not make sense, choosing any combination of nodes does not change the relative charges to each node, but only the absolute values.

Therefore, all the allocation factors obtained with any choice of slack node or combination of slack nodes are fatally flawed, since the main underlying assumption is flawed. Thus, it is very questionable that these figures should be used as a measure of the use of the line. The arbitrariness in the choice of the slack node would become even less acceptable when computing network charges in a multiple power system context, i.e. in regional or multinational systems, such as the European Union Electricity Market, the Central American Electricity Market or any of the US Interconnections.

Variations of this method are used in the Argentinean and Chilean electricity systems, where it is known as the areas of influence ("*areas de influencia*") scheme.[31] A similar method, known as cost reflective network pricing (CRNP) is in place in Australia. A more sophisticated version of this method is used in the Single Electricity Market of Ireland, whereby the incremental flows created by a network user only count towards its network charges when these flows coincide in direction with the existing flows in the baseline or reference case.

Average participations

The average participation (AP) method is based on the actual pattern of grid flows. Application of a simple heuristic rule allows one to "trace downstream" each flow withdrawn from the grid to determine the fraction of the flow of each line that can be attributed to each generator and demand at any given instant of time. The heuristic rule that is used by the algorithm is very simple: At any branching point in the network (a node) a fraction of any incoming or withdrawing flow in a transmission line divides exactly in proportion to the values of the existing total flows. This rule makes much intuitive sense, but it cannot be proved since, as indicated in Sect. 6.1.3, electricity does not propagate as water in a pipe and it cannot be traced or ascribed to any specific line.

The chief advantages of this method are its simplicity and clarity of use, and the absence of the problems involved in marginal methods, since no slack bus is involved. In the extensive experience with the AP method of the authors of this chapter, no case examples that question its soundness have been found.

This method has been applied in New Zealand's Trans Power,[32] the Polish grid company and, more recently and with additional features, in the Central American electricity market.

---

[31] As an explanation of the choice of this method in Chile or Argentina, it may be argued that the slack bus has been placed in the very dominant load centre (Santiago and Buenos Aires, respectively). In this way, most of the demand does not pay (in Argentina demand does not pay transmission charges at all, anyway) and the charges to generators grow with the distance to the main load centre, as one should reasonably expect.

[32] A first reference to this method was found by the authors of this paper in a document of Trans Power by the professor of the University of Canterbury, Grant Read: Pricing and Operation of Transmission Services: Long Run Aspects In Turner, A. (ed.) Principles for Pricing Electricity Transmission, Trans Power August 1989. More recent references are Bialek [4] and Kirschen [36].

AP is not the only network utilisation method that appears to yield reasonable results. A very different but sensible approach is the Aumann–Shapley approach in Junqueira et al. [33]. Both methods have given close numerical results when they have been applied to the same case examples.

The third principle of transmission pricing indicates that something is wrong with static cost allocation methods, where the order of installation of lines, generators and loads is ignored. This will be discussed in more detail in Sect. 6.4.3. Here it will be mentioned that the algorithms of the AP or the Aumann–Shapley methods can be modified to take account of dynamic considerations in the evolution of the power system. How? AP and Aumann–Shapley aim at determining the "average" use of the grid by each generator or load as if all facilities had always been in place. However, the responsibility of agents in network reinforcements is directly related to the incremental flows produced by the decisions of these agents to install new generators or loads in specific places. Hence, usage-based cost allocation factors produced by methods like AP or Aumann–Shapley should be modified to take account of the different possible patterns of change of the flows in the system caused by the installation of each generator or load and the time when this generators or load and the lines in the system were built. The application of these principles to the process of computation of transmission charges is discussed in detail in Olmos and Pérez-Arriaga [49].

C. Economically based methods

The methods that are based on the first principle of transmission pricing: "beneficiaries pay" and its symmetrical "responsibility for investment" have been already discussed. Only a brief account will be provided here. Even if its application is difficult in some cases, they should be the conceptual reference or guide for any other cost allocation scheme.

Beneficiaries pay

In the new regulatory provisions for competitive electric power markets, investment in a new transmission line is justified when the value of the aggregated net benefits for all the players, producers and consumers, is greater than the present cost of the line. Therefore, conceptually speaking, the transmission charging procedure would consist of allocating the additional grid cost in proportion with the benefits that the grid affords to each of its users.

The beneficiary pays method entails assessing the financial benefits that each grid user obtains from the existence of each individual transmission facility and allocating the cost of that facility among the players in proportion to the benefit obtained.

Benefit is defined to be "the financial impact for a grid user associated with the existence of a grid facility or suite of facilities". That is to say, benefit is not understood here in absolute terms, but as the difference between two situations. In practice, the financial benefit accruing to each player is assessed by comparing its benefit with and without the facility in question.

As noted above, when investment in a facility is justified, the accumulated benefits exceed its cost. Therefore, users do not pay additional charges greater than the savings deriving from the line, if the charges are applied in proportion to the obtained benefits. One of the primary virtues of this method, then, is that it is based on a fully justified economic principle, guaranteeing that the economic efficiency of long-term user decisions is not distorted.

The difficulties that exist with the application of this method have been described earlier. This is particularly the case with lines that have existed for some time. The computation of the benefits depends, in any case, on multiple assumptions and information, such as the fixed and variable costs of the generating units, which is not normally available in competitive market contexts and have to be estimated.

Despite the above difficulties, the "beneficiary pays" concept inspired the regulatory approaches adopted in Argentina (1992) and California (1998). Presently, it is considered the guiding principle that should inspire any grid charging schemes, either implemented or under consideration.

Responsibility for investment

This method seeks to allocate transmission costs on the basis of responsibility for investment, which is consistent with the economic principles set out above. It involves evaluating the additional grid investment costs induced by each user (known as deep costs[33]), in addition to the strict connection costs (known as shallow costs). In other words, transmission charges are calculated in proportion to the extra investment cost that some algorithm estimates that is occasioned by each user. The method is in use in the UK (IRCP, investment cost related pricing) and Colombia. Its implementation difficulties have been described and it is presently subject to review by OFGEM.

6.4.3 Designing Use of System Network Charges

This is the second phase in the determination of transmission charges. The fourth principle of transmission pricing reminds us that designing transmission charges involves not only developing the methodology for computing the responsibility of agents in the cost of the transmission grid, but also providing adequate answers to many implementation issues. We now focus on the most relevant aspects of the implementation of locational transmission grid charges that are not directly related to the cost allocation algorithm applied. These include the definition of the operation scenarios to be considered for tariff computation; the structure of the

[33] The terms "shallow" or "deep" applied to grid costs or charges are used to differentiate the grid investments required to physically connect a new agent to the system (i.e. a line or transformer exclusively for that agent) from grid reinforcements of existing grid required to evacuate the power produced by the new agent.

charges and their updating procedure; or how to deal with grandfathering issues in the implementation process. A detailed discussion of these topics can be found in Olmos and Pérez-Arriaga [49].

The definition of a set of scenarios that is representative of the diversity of situations that may exist in the future once the considered generators or loads have entered into operation, and the weight to be assigned to each scenario, is of essence in the computation of transmission tariffs. The relative weight given to each scenario in the allocation of the cost of a line or another transmission facility should be in accordance with the reasons justifying its construction, i.e. with the planning process. For instance, in those simpler cases where the justification of the network investment is based just on the reduction of losses and congestion costs, it is suggested that the total cost of the line should be apportioned into two parts: one representative of the weight that the reduction of transmission losses had on the decision to build the line and another one representative of the weight of the decrease in congestion costs. Then, once the set of representative scenarios has been established, the relative weight given to each scenario in the process of allocation of the cost of the fraction of the line deemed to be built to reduce losses should be proportional to the system losses in this scenario. Similarly, the relative weight given to each scenario in the process of allocation of the cost of the fraction of the line attributable to the reduction of congestion costs should be proportional to the level of congestion costs.

Perhaps, the most critical issue in the implementation of network tariffs is the definition of their format: as a €/MW, €/MWh or just as an annual lump sum charge, as this charge may have distortionary impacts on the efficient behaviour of the power system agents. Assuming that efficient short-term locational signals at wholesale level are sent via nodal energy prices (locational marginal prices, LMP in the US terminology) or some proxy (such as the application of loss or congestion factors), any additional €/MWh charge will only have a distortionary effect. For instance, if transmission tariffs are applied in the form of energy charges (€/MWh), i.e. a charge that depends on the amount of energy produced or consumed by the corresponding agent, in a wholesale short-term market the network users will internalise these charges in their energy bids to the Power Exchange or in their bilateral contracts; therefore causing a distortion in the original market behaviour of these agents and the outcome of the wholesale market. This is of particular relevance for generators, as they typically determine the market price with their bids in most power systems. However, this is also relevant for consumers that have internal generation facilities and net metering (see Chap. 8), as these internal generators will enjoy an unfair competitive advantage over the remaining generators.

An annual capacity charge (€/MW) also runs into problems. The application of a unique €/MW transmission charge per unit of installed capacity to all generators connected to the same node, regardless of their production profiles is clearly unfair (the same occurs with a €/MW charge per unit of contracted capacity to demands connected to any given node that have widely different utilisation factors). The key point to be realised is that the procedure that has been described here to allocate

transmission costs (based on scenarios and individual benefits or responsibilities or utilisations of generators and demands) allows the computation of individualised annual charges for each network user. Therefore, once the annual amount to be paid by each network user is known, it could be charged as a single lump sum, or conveniently broken down into monthly instalments or, perhaps for aesthetic reasons, as a €/MW charge conveniently tailored to each generator or demand so that it results in the previously computed annual amount.

Finally, we examine the application of the third principle of transmission pricing, which is almost universally ignored. Olmos and Pérez-Arriaga [49], as well as MIT [44], propose that the transmission tariff (or trajectory of annual tariffs) to be applied to each generator or load must be computed once and for all (or for a significant number of years) before its installation, i.e. at the time the potential network user requests connection access and has to make the decision about whether to proceed or not with the new investment. Only in this way does the locational signal provide useful feedback. Annual revisions of the value of transmission charges create uncertainty about future charges for the investors and render the locational signal useless.

Strict application of the transmission pricing principles may be relaxed for generators and loads that were connected some years ago (five to ten, for instance), since the importance of the siting decision has disappeared and the application of the first principle becomes more questionable, as explained before. The differences in the charges when transitioning from one regime to the next (or when evolving from a prior tariff regime to a more orthodox one) may be socialised (preferably to demand).

6.4.4 The International Practice

The international practice of transmission cost allocation at the national or system operator level is very diverse. Transmission tariffs in most countries do not contain any locational signal. They disregard the need to allocate line costs efficiently (see for instance ETSO [18], [59], Lusztig et al. [41]). Regulators have frequently settled for simple transmission charges that socialise the cost of the network to its users. However, in our view, as time passes and new types of generation compete for access, sending clear locational signals—which include transmission tariffs—will become more relevant.

The most common scheme is the plain “postage stamp” method, whereby every load pays a flat charge per kWh of consumed energy at any time, or per contracted kW of capacity. In some instances generators also pay on a per kW or per kWh basis. As previously discussed, this method distorts wholesale market bids. A few systems have introduced some sort of locational transmission charges, and more systems are now considering doing the same because of the anticipated large

penetration of wind and solar plants that could unnecessarily stress the transmission grid in the absence of any locational signal. In the EU, the term "locational signals" is commonly used in regulatory documents as a desideratum. However, no progress has been made in this regard at the European level[34] with exception to the UK, Ireland and, up to a certain point Sweden, who have all implemented it at the country level. The principle "beneficiary pays" is commonly accepted in official documents in the US, see [20] for instance, although its practical implementation is so far very rudimentary, to say the least. In recent official documents of the EU this principle has also been adopted, although still in general terms.

In the US, no serious attempts have been made so far to extend intra-regional (RTO) cost allocation methods to inter-regional level. On the contrary, in the EU an inter-transmission system operators (TSO) compensation mechanism (ITC) has been in place since 2002 with the following characteristics [48]: Countries—represented typically by one TSO, sometimes more than one—compensate one another for the utilisation of their networks using some metric that is based on network usage. The net balance of compensations and charges for each country—either positive or negative—is added to its total network cost from which the transmission tariffs are computed. Every country is free to design its internal network tariffs. Payment of the national transmission tariff gives every agent the right to access the entire EU transmission network without any additional charge. Although some computational aspects of this method could be much improved, this overall hierarchical approach has been a major contributor to facilitate electricity trade in the EU and, despite its simplicity, has a solid conceptual basis. Note that this method implicitly and automatically allocates the cost of any new transmission investment in the EU territory.

In the EU, the principle of allocation to beneficiaries has been adopted, although still in general terms.

6.5 Access to the Transmission Grid

6.5.1 Fundamentals

One of the objectives of the regulation of the transmission network is to guarantee non-discriminatory access to all grid users. This calls for transparent and objective rules for authorising grid connections and allocating limited transmission capacity in areas where the existing infrastructure happens to be insufficient, either transitorily or on a more permanent basis.

[34] The ITC mechanism that is described below cannot truly be considered to provide locational signals.

In systems where transmission and competitive generation were unbundled in the wake of industry reform, all players authorised to participate on the wholesale market are implicitly entitled to access the transmission grid. Obviously, however, grid capacity imposes a physical limit to access and the potential conflicts that may arise can be regulated in a number of ways, which, as will be seen below, are closely related to how responsibility for planning new investment is allocated. We can distinguish three types of situations, which require a separate treatment: (a) requests to connect to the transmission network; (b) network constraints of a local nature, just affecting one or a few nodes; (c) network constraints of a global network-wide nature.

Requests of connection to the grid

One approach to authorisation for grid connections is to limit new access rights to the existence of surplus transmission capacity, an ambiguous concept at best that depends on operating conditions, among others. Where there is no surplus capacity at the requested network connection point, consumers at least should be offered alternative connection points and measures should be adopted, so that service is provided as soon as possible. However, a different approach appears to make more sense for generators: to accept all requests and establish local market mechanisms to solve congestion problems. This procedure would preclude granting access priority to more senior or former connections and allow more efficient generators to take the place of their less efficient counterparts, just as it happens on the market as a whole.[35]

The costs of connection are recovered by dedicated connection charges. Connection costs vary with distance to the closest network of adequate voltage and capacity and the characteristics of the connection itself. There are different approaches to charge for connection costs, depending on the level of considered contribution to the cost of the dedicated facilities (which could be shared by other users) and the need for deeper network reinforcement. These are the three basic possibilities:

- No charges, then all connection costs are socialised.
- Shallow network charges, then the connection charges are meant to cover the cost of the dedicated facilities and possibly the cost of reinforcements in the local area; but the costs of any other necessary reinforcements are socialised.

[35] A diversity of approaches are followed in different countries. In some cases priority in dispatch in case of conflict is given to generators with "firm" connection, i.e. those who arrived first, before the volume of connected generation capacity started to exceed the network limits in that part of the power system. In some of these systems, when any of these generators with "firm" connection has to be constrained off, for whatever network-related reason, the generator is paid an economic compensation. In general this should not be considered a sound regulation, as it goes against the basic principle of penalising or rewarding those generators who site in inadequate or favourable locations, from a grid perspective. In some special cases it can be argued that unexpected delays in the construction of transmission lines transitorily justifies such economic compensations.

- Deep network charges, then the connection charges cover the cost of the dedicated facilities and of the necessary network reinforcements.

Difficulties arise when the reinforcements due to a new connection also benefit the existing grid users; or potentially future users, and therefore the system operator may decide to oversize the connection facilities. When the distance between the consumer or generator and the network connection point is large, the connection might have to be considered a part of the main grid instead of a connection.

Solution of local network constraints

This situation happens when, for instance, the outcome of a day-ahead market does not include any generator in a given zone of the network. Without any generator connected there, the voltage may not be maintained within the established limits for a secure operation. Therefore, at least one local generator must be constrained-on.

An ad hoc auction involving only the available local generators could be used to select the generator(s) that will solve the problem. The difficulty appears when only one generator can fix the problem, or when all the generators that can fix the problem belong to the same company. Then, if there is no possible competition there can be no market. The only reasonable solution in this case is to reach an agreement between the regulator and the involved company, so that the generator receives a regulated remuneration that satisfactorily covers its fixed and variable costs. A good number of countries maintain the ordinary market rules even in this total absence of competition, and pay the generator the price of its bid in the day-ahead market. This is obviously an invitation to the generator to abuse its market power and a headache for the regulator to chase continuously for inadequate behaviour of the generator. Note also that, if the generator operates most of the time under constrained-on conditions and is paid its bid, it will not be able to recover its fixed costs unless the bid goes beyond its pure variable costs. All this clearly indicates the need for a fully regulated solution to this type of problem.

Solution of generalised network constraints

In most cases, network constraints impact a large region of the network and any kind of local approach does not make sense. The correct treatment in this case is to run an optimal power flow model as described in Annex C and apply nodal prices to generators and loads, at their respective nodes, at regular intervals (some of the US system operators compute nodal prices for all nodes of the transmission network every 5 min,[36] others do it hourly). As explained before, nodal prices (also called locational marginal prices and spot prices) completely internalise all network effects—losses and constraints,—as well as the location and economic and technical characteristics of generators. Nodal prices implicitly manage network constraints in the most efficient way.

[36] see for instance http://www.pjm.com/markets-and-operations/energy.aspx.

However, as indicated before, nodal prices have been regarded to be unnecessarily sophisticated in systems where congestion is seldom a problem, and different simplifications have been proposed: (a) ad hoc zonal pricing, i.e. introduce price differentiation only when a congestion appears; (b) single pricing, i.e. ignore transmission congestion when the electricity market is cleared and apply ad hoc supply and clearance mechanisms, e.g. re-dispatch generation units (see below), to deal with any network restriction that may appear and charge the additional incurred cost to the responsible agents; some systems provide financial compensation for those less expensive generators that have to be constrained-off because of network limitations, with the additional costs of re-dispatching units and economic compensations being charged to consumers.

The above procedures may be supplemented with agreements for firm transmission rights or financial transmission rights[37] that enable market players to hedge congestion-related financial risk, Sect. 6.5.3 below. The risks involved include the loss of revenues if a generator is eliminated from dispatching, nodal or zonal market price volatility, or the material inability to complete a transaction. The provisions of these transmission contracts vary: physical versus financial rights; line- or corridor-based (flowgates) versus node to node. Such contracts must be awarded in ways that neither create nor intensify the potential for the exercise of market power, avoiding the award of physical rights to most of the capacity at an interconnection to a single player, for instance.

6.5.2 Management of Generalised Network Constraints

The following sections discuss how the different approaches to energy prices, with or without a locational component, combine with different types of contracts for the agents to hedge against the uncertainty in prices, when exacerbated by the presence of the network.

Implicit auctions: nodal or zonal pricing

As explained in Sect. 6.2, nodal pricing sends short-term economic signals that price the energy injected into or withdrawn from the grid at its marginal value. These individualised prices for each transmission grid node automatically internalise the economic effect of ohmic losses and grid constraints, including congestion. They send out appropriate signals for:

- correctly allocating limited grid capacity among players
- determining the economic value of grid constraints

[37] See Hogan [26] and Pérez-Arriaga [51] for an introduction to the subject, and an excellent manual "Financial Transmission Rights" on the New Zealander transmission company at www.transpower.co.nz as well as the Harvard Energy Policy Group website at http://ksgwww.harward.edu/hepg/ for background on the recent debate about the various formulae.

- determining constraint costs, who should pay and how much, who should be paid and how much.

They are applicable when the wholesale market is cleared taking into account the effect of the transmission grid, i.e. instead of a single market price for electric power, a separate price is obtained for each node.

Because the allocation of grid capacity is wholly settled on the energy market, this method is known as an *implicit auction* (for grid capacity, since the grid is implicitly taken into account as if a perfectly efficient auction of the grid capacity would have taken place). Gilbert [23] showed that implicit auctions maximise the use of transmission capacity.

An approximate version of nodal pricing is to apply zonal prices instead of nodal prices. Once the most frequent points of congestion are identified, the grid nodes affected by internodal congestion are grouped into areas or zones. This distinguishes energy prices by zone in lieu of by nodes, a procedure that simplifies market clearance and enhances transparency and simplicity. Clearing market prices with supply and demand curves is comparatively easy in systems where constraints tend to clearly delimit two or more zones. And it becomes virtually impossible if the boundaries between zones are blurred.

The process may be described as follows. After defining two zones in the system between which constraints may potentially occur, the market operator initially clears purchase and sales offers for the system as a whole. The next step is to calculate the flow over the lines connecting the two zones and determine whether the market has been viably cleared. Where it has not, the market operator distinguishes between supply and demand from the two zones, raising the strike price in the importing zone (i.e. accepting higher selling prices and withdrawing demand offers lower than the new price) and lowering the strike price in the exporting zone (withdrawing sales offers higher than the new price and accepting cheaper purchase offers) until exports from the export zone and imports in the import zone concur with the maximum available transfer capacity between the two.

Re-dispatching mechanisms and loss factors

Grid effects may be disregarded when clearing supply and demand in an electricity market, for reasons of transparency, simplicity or because grid costs have a relatively small impact on the market. However, security and efficiency require that network constraint violations and loss reduction be taken into account at a certain point of the process.

Re-dispatch of generation

In the absence of nodal or zonal pricing for handling grid constraints when clearing the market, an explicit method must be devised to allocate limited grid capacity and determine who is to pay and who to be paid for the use of this scarce resource.

Two families of methods have been proposed in this regard: on the one hand, methods based on re-dispatching generation units, a decision made by the system

operator, as explained in this section; and on the other, methods based on market mechanisms such as auctions, which are analysed in the following sections. The former manage transmission constraints ex-post (after the energy auction is cleared) while the latter manage them ex-ante (before the energy auction is cleared).

When grid constraints (congestion or voltage problems) are detected, the system operator defines the action that should be taken to deal with situations that would lead to unfeasibility. Typically, it determines which players must withdraw from the system and which are to be included to:

- accommodate the constraint [16, 60, 71]
- maintain the balance between generation and demand [1, 22].

The system operator draws on its extensive knowledge of the system in this process and in light of the various possible alternatives attempts to apply the following criteria:

- minimisation of constraint costs
- minimisation of the operating changes required.

Then, the system operator makes decisions based on specific bids geared to accommodate grid constraints.[38]

This method differs from nodal or zonal pricing (which can be viewed as the orthodox benchmark method) in that the only generators whose financial results are affected are the ones involved in re-dispatching. The other generators receive no signal whatsoever. Energy no longer dispatched by the system operator (removed to solve the network constraint) may be paid at the respective agent's bid price (if a specific bid related to the constraint solving mechanism is in place), at the opportunity price (energy market price less the price of the agent's bid) such as in the English system, or not at all, such as in the Spanish system.[39] When additional energy is requested, it is normally paid at the respective agent's bid price. Therefore, this method does not result in efficient prices applied to all network users. Particular care must be taken to prevent the potential exercise of market power by units aware that they are indispensable to alleviate routine system constraints. Since the additional costs involved are usually charged to demand regardless of the solution adopted, the system operator seeks the least expensive solution for consumers.

A simplified version of this method is available for simple network configurations where several well-defined zones with interconnection constraints between

[38] In some cases, such as Spain, the agents' own bids during the preceding energy market session were used. Nonetheless, factors relating to start-up cost internalisation, for instance, which takes place very differently depending on whether start-up is required to participate for several hours on the energy market, or to accommodate a constraint within a period of a very few hours, suggest that the two types of bids should be distinguishable.

[39] Units that are re-dispatched to rebalance the system (but not to overcome the constraint) are nonetheless paid the opportunity price.

them are readily identifiable. This scheme, known as "countertrading", is used in Norway and Sweden, within the Nordel regional market. In this case, once a constraint is detected after market clearance, the system operator purchases energy in the importing zone and sells it in the exporting zone to encourage higher production in the former and lower production in the latter. The cost of the operation is transferred to the grid tariffs.

Loss factors
The effect of transmission losses can be approximately taken into account in wholesale electricity markets that clear with a single (flat over the entire geography) energy price through the application of loss factors. Therefore, there is no need to forgo the short-term loss signals that contribute to the economically efficient system operation. The losses attributable to each player, either computed as a marginal (preferably) or average value, can be applied in the form of corrective factors to determine the prices to be paid or earned by this player or, rather preferably, the net amount of energy produced or consumed by the former. For instance, a generator that injects 100 MWh at a node where the loss factor is 0.96 will be paid the corresponding nodal price, but only for 96 MWh, or for 104 MWh if the loss factor happens to be 1.04 at this node. This should lead players to internalise the losses they are responsible for in their offers; therefore modifying their competitive position in the market clearing process.

Explicit transmission capacity auctions
Most market agents do not want to depend on uncertain and highly variable short-term electricity prices for their transactions. Of course, the situation gets worse in general when the transmission network is factored in. The natural way of hedging against this risk is to sign medium or long-term contracts. When the contracting agents happen to be on the opposite sides of a frequent network congestion, or if an agent is connected to an area of the network that suffers from frequent congestions, with the typical associated price changes, the agent(s) may want to purchase in advance the use of a fraction of the network capacity that might be congested (other options will be discussed later), so that they become immune to the differences in prices across the congestion. This purchase can be facilitated if the regulator, or the system operator on his behalf, or any private entity creates a trading platform for this type of product, i.e. auctions for direct (*explicit*) allocation of transmission capacity, separately from the market clearing of electric energy. The alternate approach—implicit auctions—consists of waiting for the short-term market and allocate energy and transmission capacity altogether in a market clearing process that jointly considers generation and transmission, as in the optimal power flow of the Annex C of this chapter.

Frequent line congestion is an indication that the capacity of the line in question is a scarce resource that must be allocated among their several potential users. One natural way to do so is to organise a line capacity auction, so that the scarce transmission capacity is awarded to the users willing to pay the most for it. The auction of the available capacity guarantees that both the allocation among the

agents and the (marginal) price for the right to use the line are efficient (assuming ideal inter-agent competition).

The practical implementation of this method always involves ex-ante auctions of the lines that the system operator foresees may become congested, i.e. auctions held prior to energy trading itself.[40] This type of auction has in fact been used primarily in regional contexts where several energy markets co-exist in independently dispatched areas. If an agent wishes to participate on the energy market in another (outside its own) system or to conclude a bilateral agreement to purchase energy from or sell energy to an agent in another system, it must first acquire the right to use the interconnection for the corresponding amount of capacity in the transmission capacity auction. Otherwise, the agent may not be allowed to participate in the market in question. The mechanism for the simplest case, i.e. two neighbouring systems, each with its own energy market, connected by an interconnecting line, would be as follows.[41] First, a capacity auction is held for the interconnecting line (or corridor) between the two systems, in which all the agents in both systems interested in accessing the neighbouring market participate. The awardees in this auction, in exchange for paying the price resulting from the auction (marginal price), acquire the right to participate in the neighbouring energy system. This mechanism enables both system operators to ensure that the flows between the two systems are technically feasible and interconnection capacity allocation is quasi efficient (why it is not wholly efficient, even with ideal energy markets, is explained in a later discussion). As illustrated below, the price resulting from the interconnection capacity auction is related to the price differences in the two energy markets (estimated a priori, because trading actually takes place after the conclusion of the interconnection capacity auction).

The following simplified example illustrates how these auctions work. Assume two systems, A and B, joined by an interconnection line such as shown in Fig. 6.14. Assume that in the absence of the line (Fig. 6.14a) the prices of energy in the two systems, given their characteristics, are respectively €40 and €50/MWh. The opportunity for exchange between the two systems is obvious. Generating units that remain idle in system A, because their variable cost exceeds €40/MWh but not €50/MWh see an opportunity to sell their energy on the neighbouring market. The system A generating units with operating costs of under €40/MWh would also be happy to sell their output in system B, where the price is higher. System B demand is also keen on buying system A energy at a lower price. Furthermore, part of the demand in system B that may not be consuming at €50/MWh, would do so at a lower price. Obviously, where possible thanks to the existence of an interconnecting line, electricity will flow from system A to system B until the prices in the two balance out. Assume, as in Fig. 6.14b, that the prices

[40] An ex-post line capacity auction (i.e. organised after energy trading) could be envisaged once the lines with overloads are identified, but given the many practical problems involved in real situations, to date they have been mere academic exercises.

[41] This is a method applied primarily in regional markets. See also Chap. 10.

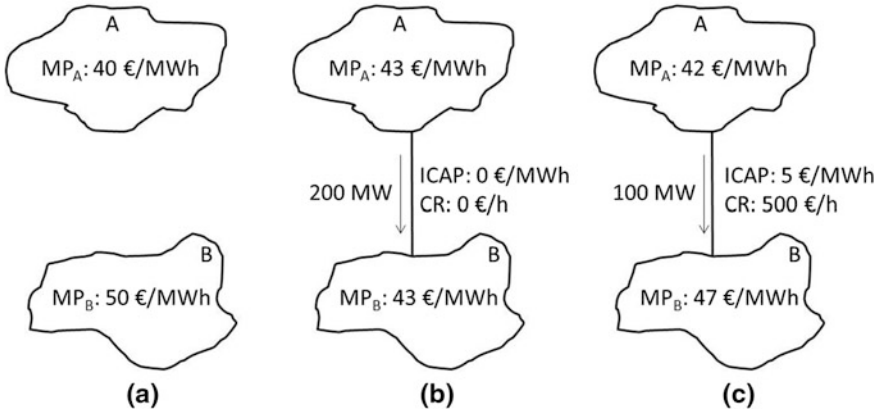


Fig. 6.14 Market prices, line flows, explicit auction prices and congestion rents for the separate (explicit) allocation of energy in two systems, A and B, and interconnection capacity between them

even out at a flow of 200 MW and a price of €43/MWh. At that time, some of the system A units feed the demand in system A and others export energy across the line to feed the demand in system B. As discussed earlier, because of the physical laws that govern the flows in an electricity system (see Sect. 6.1) there is no way of knowing which system A unit generates the energy that stays in system A and which the energy that is exported across the line to system B. This is not an issue, however; since the price is identical in the two systems, distinguishing which of the system A units is in fact using the line is irrelevant. In fact, both units (the one regarded to be exporting energy and the one that is not) receive the same market revenue per unit of energy produced.

Now assume, however, as in Fig. 6.14c, that the interconnection capacity between the two systems is 100 MW, and consequently the full 200 MW that would equal the price in the two systems cannot be transferred. Assume also that the resulting prices would be €42 and €47/MWh, respectively. In other words, a grid constraint (line congestion) appears between the two systems. Under these conditions, all the system A producers would be keen on generating the 100 MW which, exported across the line, are sold and used in system B, where the price is higher than in system A. Since, physically, the energy is exported by the system as a whole rather than any of the units specifically, and more units want to use the line than actually can, the line capacity (100 MW in this case) is auctioned among all the agents. The system A generators would be willing to bid to use the line, but never for an amount higher than the difference of price between the two systems. The marginal price for the auction would consequently be €5/MWh (€47-€42/MWh). The units that win the auction may participate in the system B energy market and charge €47/MWh, but have to pay €5/MWh in the grid capacity auction, for a net revenue of €42/MWh, while the units that do not win the auction participate in the system A energy market, likewise with a net revenue of €42/MWh.

This result is perfectly efficient, because it concurs with the results that would be obtained if the two systems constituted a single market with nodal prices. All the generating units connected to the "system A" node are paid the price at that node, €42/MWh, and all the units connected to the "system B" node, the price at that node, €47/MWh. Since exactly the same units would be dispatched, line flow would also be identical. Furthermore, the revenue resulting from the line capacity auction, €500/h (=€5/MWh × 100 MW), concurs exactly with the sums denominated earlier as line congestion rents, i.e. the difference between buying energy at one or the other end of the line (€47/MWh × 100 MW − €42/MWh × 100 MW = €500/h). In case b, since the price of energy is expected to be equal in the two systems, the result of the interconnection capacity auction would be zero, an indication that in that case such capacity would not constitute a scarce resource. The auction revenues concur with the congestion rents, which in this case are nil.

As this example shows, ex-ante capacity auctions can be held prior to energy trading in each system to determine the access rights to neighbouring markets across the available interconnection capacity. Moreover, this method proves to be a good approximation to the most efficient method, i.e. implicit auctions (nodal or area prices). It is, therefore, a much more orthodox approach than the re-dispatching described in the preceding section.

However, the method is not entirely flaw free, for a number of reasons [5]. First, the agents must submit their bids for the grid capacity auction before knowing the result of the energy prices in each system, which are established in a subsequent trading session. Consequently, they must work with estimated energy prices. If the estimates are inaccurate, the results are less than optimal.

Second, when several interconnected markets are involved, rather than two neighbouring markets such as in the example, given the distribution of flow along alternative paths, exports from one system to another may entail the simultaneous use of several interconnections, even if the two concerned systems are neighbours. Agents would have to participate in all the line auctions needed for inter-system transactions and be awardees in them all. Unless joint, coordinated auctions are held among all the systems (which is not often the case), the results will not be optimal.

Third, the flow in an interconnecting line may be the result of netting two opposite flows. In the example in Fig. 6.14, a system B generator may wish to sell 30 MW to system A (based on its estimates of the price of energy in the two systems). If that generator then participates in the auction and is awarded use of the line, 130 MW could be awarded in the opposite direction (even though the maximum line capacity is 100 MW). Under those circumstances, the final dispatching would be the same. Awarding an auction in that manner is risky, however, because the system B generator may not be able to sell its energy in system A because its price is too high. In that case, the 130 MW accepted in the opposite direction would not be feasible. If, by contrast, only 100 MW of grid capacity were accepted in the system A to system B direction, despite the 30 MW accepted in the opposite direction, the final net flow across the line might be just

70 MW, much lower than the actual capacity. The result would be inefficient use of the line.

Fourth, when circumstances are not as simple as in the example, the sensitivity of the flow in each potentially constrained line must be determined with respect to the energy involved in each transaction.[42] Before submitting their bids for the use of an interconnection line, agents should know the impact of their output on the line flow. In the preceding example this sensitivity factor is obviously one, since one MW exported from system A to system B always induces the same flow across the line, regardless of which system A unit produces it. But in a more complex situation (with meshed interconnected systems), this sensitivity factor must be determined by the system operator. Its calculation prior to dispatching is another source of potential inefficiency. Moreover, a number of interconnections may be required to reach some systems (bearing in mind, in particular, that loop flows arise). Agents should therefore submit coordinated bids to all the interconnection auctions they may potentially be using for their transactions. This constitutes yet another source of potential inefficiency, since the mutual interdependence among flows in different parts of the grid simply cannot be avoided. The result might be unfeasible flow patterns or less than full use of capacity on congested lines due to market agents' inability to obtain sufficient capacity on all the lines they need.[43]

Account must also be taken of the fact that capacity auctions may favour the exercise of market power [23]. The awardee of interconnection capacity between systems or areas may control the appearance on its own market of new outside competitors. An agent might find offering line capacity at below cost to be to its benefit, if it can thereby block the interconnection, even without making use of it, and offset its loss by exercising greater market power. This issue must be addressed by the regulator when designing and regulating such auctions.

A number of measures may be taken to palliate this problem. First, total interconnection capacity should not be auctioned over the long term. Since grid capacity auctions are held prior to energy trading, they can be organised as far in advance as desired. As discussed in the following section, capacity auctions can also be organised as the sale of transmission rights, a product used by agents to hedge their grid-associated risks (such as the appearance of constraints). In this respect, agents would have an incentive to have grid capacity assigned sufficiently in advance to establish their operating strategies.

Auctions may even be held fairly far in advance and cover long time spans. When that happens, the full interconnection capacity must not be auctioned, because that would enable one agent to block access long enough and far enough in advance to discourage possible competitors. The standard procedure is to divide

[42] An alternative approach that avoids some of the difficulties of this flow-gate approach is the use of point-to-point contracts in the auctions, as explained in Sect. 6.5.3 of this chapter.

[43] Some of the articles in [7] showed, however, that theoretically this system should converge, under ideal conditions, towards efficient equilibrium. Other studies on the use of flow-gate capacity auctions followed by energy-only markets can be found in Oren and Ross [50] and Tabors and Caramanis [70].

line capacity, so that one part is auctioned in the long term (1 year or over), another in the medium term (one or several months) a third in the short term (1 week or several days) and a final tranche immediately before very short-term energy trading (day-ahead or intraday).

Second, the capacity awarded in these auctions must be associated with "*use it or lose it*" or "*use it or sell it*" clauses. These clauses require the awardee to state at a given time (nomination time) if it is going to use the grid capacity awarded or otherwise. If not, the right reverts to the system, which makes it available to the rest of the agents (via another auction, for instance). The former awardee would either forfeit the amount paid ("use it or lose it") or be indemnified by the amount obtained in the auction ("use it or sell it").

While the market power stemming from this type of grid capacity might be mitigated to some extent by these measures, it is not eliminated altogether. In some systems, even more drastic solutions have been implemented, by prohibiting the predominant agents from participating in capacity auctions, for instance.

Implicit auctions may be difficult to implement, especially in regional (multinational) systems where energy dispatching in each control area or country has traditionally been in the hands of a separate entity (the local SO or TSO). When capacity is allocated separately, the SO in each area can independently clear the local energy market, at least if the capacity auction is held before energy is dispatched. This is not the case in implicit auctions. A coordinated explicit auction followed by independent energy auctions within each area or country may strike a reasonable balance between efficiency and practicability. Actually, explicit auctions for transmission capacity have traditionally been held for frequently congested lines or flowgates [29].

6.5.3 Transmission Rights

Short-term energy market prices may be highly volatile due to the special characteristics of the "electricity" product (in particular, the fact that it cannot be stored in any market-relevant amounts). The price of energy (ignoring other more complex issues such as the grid effect) rises and falls from 1 h to the next depending on countless factors, such as the generation available at the time, demand, fuel prices, hydroelectric availability or renewable resources. Agents naturally seek to protect themselves from the risks entailed in such volatility by trying to stabilise energy market revenues and expenses. Some of these markets have reached maturity, as in Nordel or some US markets, for instance.

These products may vary widely. While the scope of this chapter does not cover details on these products,[44] given the simplicity of the contract for differences and its usefulness in the definition of the transmission rights, it is discussed here.

[44] See Chap. 7 for a more detailed account of contracts in electricity markets.

Contracts for differences are bilateral financial agreements between an energy seller and a buyer. The agreement establishes a reference price and a volume of energy. Both parties participate in the short-term market to buy and sell energy at the market price, but by virtue of their agreement, the buyer pays the seller the difference between the market and agreed energy prices whenever the former is lower than the latter, and vice versa if it is higher. This enables each to stabilise its revenues and expenses, respectively. Both agents assume a supplementary risk with respect to the price of energy, which makes them natural counterparties in contracts for differences. Even so, other types of agents may also participate, speculating on the price of energy.

The transmission grid adds a new type of risk to energy prices and unit output capacity. As shown earlier, the implementation of efficient congestion management mechanisms leads to energy prices with some manner of spatial differentiation. As a result, these prices acquire further volatility, associated with the existence of grid congestion or system constraints.[45] Transmission contracts or transmission rights (TRs)[46] are mechanisms geared to hedging the risk associated with such volatility. The use of transmission capacity rights was initially proposed by Hogan [26].[47]

Imagine an electricity system in which the price varies depending on geography. A producer and a consumer located on different grid nodes would be unlikely to conclude a contract for differences such as described above to hedge the risk associated with short-term market volatility if they are aware that their nodes may be impacted by a grid constraint. Assume that the nodal price at the node where the producer is connected is €35/MWh and is scantly affected by the appearance of system constraints. Assume that the nodal price at the node to which the consumer is connected, however, is €40/MWh if no constraint arises and €45/MWh otherwise. If the price difference is stable (at €5/MWh, for instance), a contract for differences could be concluded using the two prices (€35 and €40/MWh) as reference prices. In that case, if the prices become €37 and €42/MWh, respectively, the generator transfers €2 to the consumer. The difference between the two prices is a grid revenue (related to the grid variables or congestion rents discussed earlier). Nonetheless, due to the constraint, the price at the consumer's node may rise to €45/MWh, while at the generator's node, the price remains flat at €35/MWh. The generator cannot pass on to the consumer the extra €5/MWh, because its price has not risen by that amount. Under such circumstances, those two are no longer

[45] Contrary to claims by some authors, the appearance of congestion in electricity systems must be a common occurrence (see [69]).

[46] Actually, this product is known by very different names in the literature. Some describe its characteristics, i.e., financial versus. physical, while others are used for the same product, i.e. financial transmission rights and congestion revenue rights. ETSO [17] proposes transmission risk hedging products (TRHPs) as a generic name for any market-based solution available for hedging transmission risk in congestion management.

[47] In a draft for an article written shortly before his death in 1988, Professor Fred Schweppe of MIT developed the fundamentals underlying firm grid rights.

natural counterparties in connection with grid risk and cannot hedge one another against price volatility induced by grid constraints.

When this happens, however (i.e., when the difference between the two prices grows and consequently the buyer–seller pair must pay more as a whole), it is the variable grid revenue that increases. The grid itself becomes the natural counterparty to the buyer–seller pair with respect to grid constraint-induced risk. Factoring in the likelihood of a constraint, the mean expected price to the consumer might be €41/MWh, for instance. The producer–buyer pair could hedge constraint-related risk (i.e. charge and pay the same whether or not the constraint ultimately materialises) by reaching the following agreement with the grid. The pair buys the rights to be charged for the actual variable revenue between the two nodes at a closed price, which would be the mean expected congestion rents for that transaction (€6/MWh). No matter what happens in terms of the constraint, then, the consumer's payment (€41/MWh) and the producer's revenue (€35/MWh) are flat. The difference, also logically a constant, is what is paid to the grid. If the consumer node price is €40/MWh because no constraint arises, the generator receives €35/MWh on the market, the consumer pays €40/MWh on the market and together they pay the €6/MWh grid agreement price, but receive the revenue generated, which comes to €5/MWh. Consequently, the grid agreement generates a €1/MWh extra cost which, added to the €40/MWh paid by the consumer on the market, amounts to the target price, €41/MWh. Conversely, if the consumer's node price is €45/MWh because a constraint does arise, the generator receives the €35/MWh market price and together they pay the fixed €6/MWh grid agreement charge, but receive the variable revenues generated, which amount to €10/MWh. Therefore, the grid agreement generates a €4/MWh profit which, subtracted from the €45/MWh paid by the consumer on the market, comes to the €41/MWh target price.

By virtue of this agreement, the grid, instead of having volatile revenues (ranging from €5 to €10/MWh) has a stable revenue of €6/MWh, the expected mean. Nonetheless, this affords the grid no advantage whatsoever, since its total remuneration as a regulated business is established by other mechanisms that cannot depend on variable grid revenue.[48] Since the grid is the natural counterparty, such agreements are normally concluded with the system operator, the entity in charge of managing grid variable revenues or congestion rents. As with contracts for differences, however, since this type of agreement is no more than a bet on the price difference between two system nodes, any other agent could participate as a speculator. The system operator cannot commit to a higher volume of variable revenues than the grid can generate, for otherwise it would be taking a risk. This principle is generally known as the revenues adequacy criterion and can be readily understood when applied to the simple case of two nodes connected by a single line. If the line capacity is 100 MW, the system operator should not conclude agreements for more than that amount, otherwise the variable grid revenues (line congestion rents) would not suffice to pay for the commitments undertaken.

[48] Dependence on variable grid revenue is acceptable when merchant lines are involved.

While for the sake of simplicity, the contract has been described here as an agreement between the producer–buyer pair and the grid (system operator), normally the producer reaches an agreement with a consumer (or retailer) to deliver a certain amount of energy at the latter's node. It subsequently concludes a contract with the grid (system operator) or a speculator to hedge the risk associated with possible price differences between its connection and delivery nodes resulting from grid constraints.

Such contracts, particularly when handled by the system operator, are usually acquired on organised markets, where standardised transmission rights are auctioned. As in the capacity auctions mentioned in the preceding section, the timeframes and volumes may vary and their design may be relevant to market power-related issues. The transmission rights contract described above is a purely financial agreement and the rights are in fact known as financial transmission rights. Since the rights refer to two system nodes (or areas), they could be defined as *point-to-point transmission rights*. Rather than the right to transmit energy, what is acquired is the right to collect variable grid revenues at the energy prices prevailing in two system nodes (or two areas). For that reason they are also called congestion revenue rights in some systems.

These are not the sole approaches possible, however. Physical TRs may be traded rather than financial TRs, and flow-gate TRs rather than point-to-point TRs. And, like any other financial product, they may be acquired as obligations or options. The differences between these schemes are discussed briefly below. An excellent review can be found in Batlle and Gómez-Elvira [3].

Given that Financial Transmission Rights entitle owners to receive the difference between the energy prices at the nodes that these rights refer to, the aggregate value for market agents of all the simultaneously feasible transmission rights (defined as obligations) that can be issued in the system would equal the expected overall net revenues from the application of energy prices. Due to the fact that, as explained in Sect. 6.2.2, these revenues tend to be much smaller than the total transmission network cost in a typically well-developed transmission grid, it is highly unlikely that the financing of investments in the transmission grid through the issuance of FTRs would result in an appropriate development of the grid. Most of the required reinforcements could not be financed through this scheme.

Physical versus financial rights

As noted above, financial transmission rights allow their holder to collect the respective variable network revenues or congestion rents. Physical transmission rights entitle their holder to use the network transfer capacity between two points or areas in the case of point-to-point rights or the capacity of a line or corridor in the case of flow-gate rights. Therefore, holders of point-to-point physical rights can inject electricity at one node and retrieve it at the other [7, 66] at no extra cost. Owners of point-to-point financial rights could simply purchase power at one node and sell it at the other, with congestion rents providing a perfect hedge against the respective price difference. Therefore, in ideal markets, physical rights would have no advantage over financial rights.

Physical rights also partially hedge the volume risk associated with transmission capacity, however. Their holder has priority physical access to capacity and can therefore guarantee that the commodity will be delivered unless the lines to be used are not in operation, since this capacity is actually reserved (until some specified deadline prior to system operation) for the rights holder. TRs are therefore associated with firm energy supply contracts, which provide long-term security of supply to the end node market, and must be physical. They are used in the Central America Market, for instance (see [14]).

Physical rights may be closely associated with the explicit network capacity auction described in the preceding section. They are generally used to allocate interconnection capacities between different systems, such as in many European interconnections, while financial rights are used within a single market in which prices differ depending on location (nodal prices or market coupling between different systems, see Chap. 10), such as in the PJM system. Physical rights are always associated with use-it-or-lose-it or use-it-or-sell-it clauses to mitigate the potential exercise of market power.

In any event, an academic debate has taken place about the real differences between and the pros and cons of physical and financial transmission rights.[49] Under ideal market circumstances (perfectly competitive and efficient market structure), the two approaches would be equally efficient and useful. Market power concerns and market structure deficiencies (i.e. lack of harmonisation and coordination where several systems are involved, lack of market liquidity, existence of transaction and operating costs for the bidding process, etc.) may lead to slight differences between them, however. An excellent discussion of this issue can be found in Batlle and Gómez-Elvira [3].

Flow-gate versus point-to-point (or area-to-area) rights

Point-to-point transmission rights are related to the capacity required for a transaction between two or more nodes or areas. Flow-gate transmission rights are related to the capacity of a specific line or corridor which may become congested (a flow gate).

Point-to-point rights allow their holder to conduct a transaction between two nodes or areas or capture the congestion rents associated with a hypothetical transaction between these two nodes. Congestion rents in this case are equal to the transaction volume times the difference in energy prices between the injection and

[49] Chao and Peck [8], Tabors and Caramanis [70], Oren and Ross [50] stress that transmission rights need to be physical in the short term to conduct physical transactions. According to these authors, however, "use-it-or-sell-it" clauses must be applied to make capacity hoarding more difficult. Kench [35] states that while both physical and financial rights can provide for efficient congestion management, physical rights are better tools for ensuring that market agents contract power at efficient prices. Hogan [26] recommends the use of financial rights, because if the respective congestion rents are earned, for the holder of financial capacity rights it is immaterial whether the physical transaction is concluded or the respective payment is received. Hogan further claims that unlike physical rights, financial rights do not affect physical short-term energy dispatching.

withdrawal points. Flow-gate rights allow their holder to use a specific line or corridor or collect the respective congestion rents.

A substantial number of academic studies have been published both in favour and against the use of each of these types of transmission rights [8, 27, 66, 70]. The point-to-point approach would appear to be more beneficial, since it enables the holder of the right to conclude a transaction (or earn the respective congestion rents) between the respective sending and receiving nodes (or zones), regardless of grid configuration, active constraints or other contingencies. This transaction may also involve several flow gates (the loop-flow characteristic of energy transmission enlarges the number of flow-gates potentially involved in a transaction). Coordinated action is therefore required of agents simultaneously acquiring several flow-gate rights. Moreover, since flows depend on other agents' behaviour and contingency scenarios, the former have no guarantee that they have acquired the appropriate flow-gate rights [26, 67]. Point-to-point rights require this coordination to be implemented in the auction clearing process, which becomes much more complex and centralised. Consequently, flow-gate rights also have some advantages over point-to-point rights[50] ([50, 70], among others). As discussed in [7], secondary flow-gate rights trading may be bilateral, with no need for the intermediation of a centralised institution.

Point-to-point rights are better adapted to centralised trading (single or coordinated markets with geographic energy price differentiation), such as on the PJM market, whereas flow-gate rights are the better option in decentralised environments (with interconnection capacities between several systems), such as in many European interconnections. Hence, the election of the type of rights to implement should be made contingent upon the trading model chosen [43].

Stated more simply, the identification of areas comprising a series of nodes that are similarly affected by constraints would contribute to enhancing the coordination and liquidity of auctions involving either point-to-point or flow-gate rights. Such areas are broadly known as single price areas (SPAs). The definition of SPAs is central to the design and implementation of grid congestion management mechanisms. If large enough SPAs cannot be defined, then implicit nodal auctions, i.e. nodal pricing, should be deployed in short-term energy dispatching. Defining SPAs is no simple task, however. For a more detailed discussion, see [34, 38, 47], among others.

Options versus obligations

A transmission rights option allows its holder to use its (physical) transmission capacity or receive the respective (financial) congestion rents. Holding an option is not tantamount, however, to being required to exercise it. By contrast, obligations constitute commitments to use the (physical) transmission capacity to which the

[50] Many authors have expressed the concern that auctions for point-to-point rights may be subject to liquidity problems. Holding an auction to buy and sell point-to-point capacity rights entails no loss of liquidity, since these algorithms are implicitly exchanging flow-gate rights.

obligation refers, or receive the dispatching value of this capacity, which may be negative and thus involve a (financial) payment. Options allow agents greater flexibility when adjusting their transmission rights to their commercial positions and are consequently more valuable to rights holders than obligations.

As discussed above in connection with explicit grid capacity auctions (see preceding section), the two approaches may impact grid use differently; in obligations, holders may capitalise on the netting effect of transactions in the opposite direction, while in options this is not possible. Actually, the TSO must auction options for the use of line or corridor transmission capacity separately for each direction. Selling options for the (physical or financial) use of line capacity in one direction does not raise the amount of transmission capacity that can be auctioned in the other direction, because inasmuch as they are options, their net effect cannot be foreseen. If physical rights options were netted, the resulting power flows might not be feasible [27, 66]. Similarly, if (financial) congestion rights were netted, the dispatching outcome might not suffice to pay all transmission rights holders (the revenues adequacy criterion might not be met). Therefore, options reduce the total amount of transmission capacity that can be potentially sold in the long term and may result in a lower use of capacity.

In practical terms, the feasibility of implementing obligations has been proven [27], whereas options have yet to be physically implemented on any significant scale. Nonetheless, options have been recommended by some authors [8, 43].

Options allow agents greater flexibility when adjusting their transmission rights to their commercial positions and are consequently more valuable to rights holders than obligations. The effect of the use of options on market liquidity is mixed. All other things being equal, the more flexible products are, the greater the demand. This may increase liquidity. At the same time, the fact that the transmission capacity sold in both directions under options cannot be netted may reduce the amount of capacity sold and therefore the liquidity of the capacity market.

6.6 Conclusions

This chapter discusses the main technical and economic characteristics that make the transmission grid a natural monopoly and therefore subject to regulation.

The key features of transmission regulation have been identified and analysed. Investment in new transmission facilities can be organised under various approaches that are not mutually exclusive: (i) centralised planning under cost of service regulation, (ii) a single transmission company whose investment decisions are incentivised by RPI-X regulation schemes, (iii) a coalition of grid users that drive new investments and (iv) merchant lines.

Access to limited transmission capacity is another key issue for transmission regulation. Access priority affects the position of players in different electricity markets, and therefore their potential costs and benefits derived from energy

trading. Several methods for allocating the existing limited interconnection capacity among interested or affected market players have been reviewed. These methods range from re-dispatching generating units by the system operator to market-based methods where limited capacity is auctioned among interested market players.

The importance of locational short-term and long-term economic signals associated with the transmission grid and how they should be used in transmission regulation has been stressed here. Short-term signals given by locational energy pricing such as nodal or zonal prices provide incentives for optimal and efficient system operation and for allocating limited interconnection capacity. Their expected value also provides a useful signal for future investors. Long-term signals, given by locational transmission charges, are needed to share the allowed revenues from regulated transmission installations among grid users, while encouraging new supply- and demand-side actors to locate efficiently. Finally, several methods are discussed for the design of efficient locational transmission charges, so that they maintain their locational impact, but avoid any distortionary effects on the behaviour of the market agents.

Annex A
Data for the Case Example

This appendix provides the technical data of the system depicted in Fig. 6.7, which was used to illustrate the fact that nodal prices internalise the effect of losses and grid constraints on the marginal cost of supplying electricity in each node. Table A.1 provides the variable cost of all generation units in the system, expressed in €/MWh. Table A.2 provides the figures for the line parameters.

Table A.1 Cost of generation units expressed in €/MWh

Bus	Generation								Load	
	Unit 1		Unit 2		Unit 3		Unit 4		Demand	ENS
	Capacity MW	Cost €/MWh	Capacity MW	Cost €/MWh	Capacity MW	Cost €/MWh	Capacity MW	Cost €/MWh	MW	Cost €/MWh
1	300	30	75	65	125	70	100	75	1	1,500
2	100	59	50	67	50	74	–	–	240	1,500
3	160	30	100	61	50	76	50	80	40	1,500
4	–	–	–	–	–	–	–	–	160	1,500
5	–	–	–	–	–	–	–	–	240	1,500
6	200	30	–	–	–	–	–	–	80	1,500
7	–	–	–	–	–	–	–	–	100	1,500
8	100	30	–	–	–	–	–	–	15	1,500
9	–	–	–	–	–	–	–	–	100	1,500

Table A.2 Values of the parameters of lines in the system

Line		Reactance	Resistance	Capacity limit
From	To	p.u.	p.u.	MW
1	2	0.058	0.016	500
1	4	0.04	0.01	500
2	3	0.034	0.008	500
2	4	0.08	0.02	500
2	5	0.04	0.01	500
2	6	0.08	0.02	500
3	5	0.02	0.004	500
3	8	0.08	0.02	500
4	6	0.12	0.03	500
5	6	0.04	0.01	500
5	8	0.08	0.02	500
6	7	0.12	0.03	500
6	9	0.04	0.01	500
7	9	0.04	0.01	500

Annex B
Further Properties of Nodal Prices

Further properties of nodal prices are presented in this appendix. It is important to understand these properties when designing or evaluating certain methods of allocation of costs or economic responsibilities in relation with the transmission network [57].

Property 4: Nodal price breakdown into energy-, ohmic loss- and grid constraint-related components

Nodal prices differ from one node to another due to ohmic losses and the effect of grid constraints. Given the importance of suitable economic signals for regulatory design, a number of attempts have been made to break down the expression for and value of the nodal price into essentially three components. The first is associated with the marginal cost of energy in the system, irrespective of location issues. This first component is a single value for the entire system that corresponds to, for instance, the marginal generator's operating cost. The second is associated with the additional effect of ohmic losses on the price of energy and depends on the specific location of the node. The third is associated with the additional effect of grid constraints on the price of energy and, likewise, depends on the specific node location.

The advantages of this breakdown are obvious. If the weight of each of the effects in the nodal price could be distinguished and delimited, each effect could be handled separately, and efficient, individual economic signals could be designed. Rivier and Pérez-Arriaga [62] analysed this possible breakdown in detail. Expressions such as B.1 were already formulated by Schweppe et al. [68], in which

the nodal price associated with node k, ρ_k, is expressed as the sum of γ, which is independent of node k, and η_k, which depends on node k in particular. As expression B.1 shows, η_k can in turn be broken down into another two terms, one related to losses (LF_k, the node k loss factor, defined as the increase in total system loss induced by an increase in demand at node k) and the other to the effect of grid constraints ($\mathrm{NC}_{j,k}$, the constraint factor, defined as the increase in the parameter (such as flow or voltage) affected by the jth grid constraint when demand rises at node k). Parameter μ_j is the extra cost to the system of reducing the operating limit imposed by the jth constraint (such as line capacity or the maximum or minimum voltage limit) by one unit.

$$\rho_k = \gamma + \eta_k = \gamma + \gamma \cdot \mathrm{LF}_k + \sum_j \mu_j \cdot \mathrm{NC}_{j,k} \tag{B.1}$$

LF_k and $\mathrm{NC}_{j,k}$ are therefore sensitivity factors, measuring respectively changes on losses and on any constrained parameter of the system, for an increase in demand at node k.

Rivier and Pérez-Arriaga [62] showed, however, that while such a breakdown is possible, it is fairly arbitrary. As the discussion in the following section shows, the value of γ, LF_k and $\mathrm{NC}_{j,k}$ depends on the arbitrary choice of a system slack node. Consequently, Eq. B.2 is more fitting than 7.5, where s is the index of the node chosen as the slack node.

$$\begin{aligned} \rho_k &= \gamma_s + \eta_{k,s} = \gamma_s + \gamma_s \cdot \mathrm{LF}_{k,s} + \sum_j \mu_j \cdot \mathrm{NC}_{j,k,s} \\ &= \rho_s \cdot (1 + \mathrm{LF}_{k,s}) + \sum_j \mu_j \cdot \mathrm{NC}_{j,k,s} \end{aligned} \tag{B.2}$$

where

- γ_s is the nodal price at slack node s (if a marginal generation unit is connected to node s, it also concurs with the marginal operating cost of that unit, ρ_s).
- $\mathrm{LF}_{k,s}$ is the node k loss factor (sensitivity factor), defined as the increase in total system losses attributable to the changes in flows when an increase in demand at node k is met by generation induced at the slack node s.
- $\mathrm{NC}_{j,k,s}$ is the node k constraint factor (sensitivity factor), defined as the increase in the parameter (such as flow or voltage) affected by the jth grid constraint, attributable to the increase in generation induced at slack node s to meet an increase in demand at node k.

If these expressions were defined with a different slack node s, the values of γ_s, $\mathrm{LF}_{k,s}$ and $\mathrm{NC}_{j,k,s}$ (and consequently $\eta_{k,s}$) would change, while the nodal price, ρ_k, would remain constant.

Therefore, the nodal price cannot be rationally broken down in univocal terms into the cost of energy, the extra cost associated with ohmic losses and the extra cost associated with grid restrictions. One alternative that might appear to be

rational (albeit also somewhat arbitrary) would consist of choosing the node to which the system marginal generation unit is connected at any given time as the slack node. Under that approach the general term γ_s would indicate the cost of generating one additional MW of energy with the cheapest unit available, which might be defined as the pure energy component of the nodal price. The additional terms would indicate, specifically for node k, the extra cost due to ohmic losses and grid constraints. Note that, as the earlier examples showed, the identity of the marginal unit depends on the ohmic losses and grid constraints in place at any given time. Furthermore, when ohmic losses and grid constraints are brought into play, more than one marginal unit may exist in the system, each connected to a different node.

Rivier and Pérez Arriaga [62] showed that expression B.2 can be readily generalised when the slack node chosen is distributed across the entire system. A *slack node*[51] $s*$ can therefore be chosen as a linear combination of several nodes in system i, each with a relative weight t_i.[52]

In that case, the equations would be expressed as in B.3.

$$\rho_k = \gamma_{s^*} + \eta_{k,s^*} = \gamma_{s^*} + \gamma_{s^*} \cdot \mathrm{LF}_{k,s^*} + \sum_j \mu_j \cdot \mathrm{NC}_{j,k,s^*} \tag{B.3}$$

$$\gamma_{s^*} = \sum_i \rho_i \cdot t_i; \text{ and } \sum_i t_i = 1 \tag{B.4}$$

where

- LF_{k,s^*} is the loss factor for node k (sensitivity factor), defined as the increase in total system losses that would induce the increase, distributed in accordance with all the generation (t_i) in all slack nodes i, required to meet a given increase in demand at node k.
- NC_{j,k,s^*} is the node k constraint factor (sensitivity factor), defined as the increase in the parameter (such as flow or voltage) affected by the jth grid constraint that would induce the increase, distributed according to all the generation (t_i) at all slack nodes i, needed to meet a given increase in demand at node k.

In this case, one possibility would be to define the slack node as a distributed combination of the nodes to which a marginal generator is connected. Even here, however, the nodal price would continue to be rather arbitrary, for two reasons. First, ohmic losses and grid constraints determine which units are marginal and where they are sited; and second the t_i coefficients must be constant, whereas the

[51] When modelling transmission networks, the difference (the slack) between the total active power input and total active power output plus the computed total losses is balanced by power injected or with drawn at an arbitrarily selected *slack node*.

[52] Some authors have proposed the use of a distributed slack node comprising all nodes with demand, whereby the response of every node is proportional to its amount of actual demand. This scheme is used, for instance, in the computation of transmission charges in the Irish Single Electricity Market.

distribution of these units would vary depending on the position of node k, the slack node whose demand is being increased in the nodal price calculation.

By way of conclusion, breaking down the nodal price may constitute an entertaining exercise with potential practical applications, but sight must not be lost of the conditioning factors and actual meaning of the outcome.

Finally, Rivier and Pérez-Arriaga [62] also showed that expression B.2, which relates the prices at nodes k and node s, can be extended to a more general expression that relates the nodal prices of any two nodes, k_1 and k_2, as per B.5.

$$\rho_{k_1} = \rho_{k_2} \cdot (1 + \mathrm{LF}_{k_1,k_2}) + \sum_j \mu_j \cdot \mathrm{NC}_{j,k_1,k_2} \tag{B.5}$$

where

- ρ_{k_1} and ρ_{k_2} respectively represent the nodal prices at nodes k_1 and k_2.
- LF_{k_1,k_2} is the node k_1 loss factor with respect to node k_2, defined as the increase in total system losses induced by the increase in generation needed at node k_2 to meet an increase in demand at node k_1.
- NC_{j,k_1,k_2} is the node k_1 constraint factor with respect to node k_2, defined as the increase in the parameter (such as flow or voltage) affected by the jth grid constraint, induced by the increase in generation needed at node k_2 to meet an increase in demand at node k_1.

Rivier and Pérez-Arriaga [62] also discuss other properties of nodal price of lesser significance.

Property 5: Impact of the choice of reference node on the value of the sensitivity factors.

The sensitivity factor $\mathrm{NC}_{j,k}$ that is associated with line flows frequently appears in regulatory applications that are generally related to the allocation of network costs (see Sect. 6.5). That factor, which measures the variation in line flow (associated with constraint j) when demand at node k rises by one unit, represents the incremental measure of node k's use of line l (where line l is the line affected by constraint j). The scientific literature refers to this sensitivity as PTDF (for power transfer distribution factor). Hereafter $\mathrm{PTDF}_{k,l}$ will mean the variation in the flow in line l when one additional unit of power is injected at node k.[53]

[53] This definition is used because it is in keeping with the normal use made of this factor in the scientific literature. Up to now, factor NC has been defined as sensitivity with respect to an increase in demand at node k. Hence, those two are the same but with the sign reversed. Note that PTDFs are also often defined in the scientific literature as the variation in flow in the event of a one-unit increase in a given transaction between any two nodes in the system.

As noted above, since the value of such sensitivities depends on the slack node used in their calculation, the notation would more correctly be $\mathrm{PTDF}_{k,l,s}$. This actually indicates the variation in the flow in line l when one extra unit of power is injected in the system at node k, at the same time as node s removes the power needed to satisfactorily balance the system (in the absence of losses, 1 MW is injected at node k and 1 MW is removed at node s, generating a variation in the flow in line l of $\mathrm{PTDF}_{k,l,s}$ MW).[54]

Given the use made of these factors in some applications, especially with regard to the allocation of network costs, their possible variability depending on the slack node chosen for the calculation is a matter that merits discussion. Where the effect of losses is ignored in these factors, thereby "linearising" the expressions that relate flows to the injection and withdrawal of power at the nodes, the superposition principle allows the relationship between such factors calculated for two different slack nodes, s_1 and s_2, to take the following form:

$$\mathrm{PTDF}_{k,l,s_2} = \mathrm{PTDF}_{k,l,s_1} + \mathrm{PTDF}_{s_1,l,s_2} \tag{B.6}$$

Note that, since $\mathrm{PTDF}_{s_1,l,s_2}$ does not depend on node k, under these circumstances the slack node may be deduced to generate a uniform shift in all the sensitivities associated with a given line for all the nodes in the system. Consequently, the difference in the absolute (but not the relative) value of the sensitivities for different system nodes remains unaltered, despite the change in slack node.

This property may be useful in certain applications when what is sought is not an absolute value, but instead, a relative value of line use sensitivities when power is injected at system nodes.

The explicit inclusion of losses calls for changes in these expressions. When changes take place at the slack node, the loss factors are related as follows (see [62]):

$$\left(1 + \mathrm{LF}_{k,s_2}\right) = \left(1 + \mathrm{LF}_{k,s_1}\right) \cdot \left(1 + \mathrm{LF}_{s_1,s_2}\right) \tag{B.7}$$

and result in the following expression:

$$\mathrm{PTDF}_{k,l,s_2} = \mathrm{PTDF}_{k,l,s_1} + \mathrm{PTDF}_{s_1,l,s_2} \cdot \left(1 + \mathrm{LF}_{k,s_1}\right) \tag{B.8}$$

In this case, the absolute difference between sensitivities when the slack node changes is not strictly invariable, due to the impact of the loss factor associated with node k. Since that factor is normally very close to zero; however, the conclusions may be maintained, depending on the application for which the calculations are intended.

[54] Where losses are taken into consideration, the amount of power to be removed by the slack node would not be 1 MW, but a somewhat smaller or larger amount, depending on how system losses would be impacted by the exchange.

Annex C
Computation of Nodal Prices

The power system model that will be used here uses linearised and simplified versions of the equations governing grid flows that nonetheless reflect the physical fundamentals of flow distribution under Kirchhoff's laws. This simplified grid model, which is often cited in the literature (see, for instance, [73]), is known as the DC model because while it simulates the behaviour of an AC system, the equations resulting from the simplifications are somewhat reminiscent of DC system circuits.

The model is described by the set of Eq. C.1. For the sake of simplicity, ohmic losses in each line have been represented as a function of the flow over this line and assigned to the end nodes of the line, thus being equivalent to an extra demand in each of the two nodes (half of the losses would be assigned to each node). In addition, in order to make the formulation simpler, the only considered grid constraints are the maximum capacities of lines. The model aims primarily to show how nodal prices can be obtained in both dispatching models and market clearing schemes. An exhaustive discussion of more complex formulas and expressions that would portray transmission grid behaviour more reliably falls outside the scope of this book.

$$
\begin{array}{lll}
\max \sum_i \{B_i(d_i) - C_i(g_i)\} & & \\
s.t. & & \\
d_i - g_i + \sum_m \left\{\tau_{im} \cdot \phi_{im} - L_{i,m}(\phi_{im}, R_{im})\right\} = 0; & \forall i & \pi_i \\
\tau_{im} \frac{\theta_i - \theta_m}{x_{im}} = \phi_{im} & \forall i, m & \xi_{im} \\
\phi_{im} \leq \bar{\phi}_{im}; & \forall i, m & \mu_{im} \\
\theta_{ref} = 0 & & \\
g_i \leq \bar{g}_i; & \forall i & \beta_i \\
d_i \leq \bar{d}_i; & \forall i & \alpha_i
\end{array}
\tag{C.1}
$$

where i is an index representing the set of nodes; m is an alias of i; $B_i(d_i)$ is the utility of consumers at node i as a function of their total demand di[55]; $C_i(g_i)$ is the total operation cost incurred by agents at node i when producing $g_i$units of power[56]; τ_{im} is a binary variable whose value is 1 when nodes i and m are connected by a line and 0 otherwise; ϕ_{im} is the flow over the line between nodes i and m in direction i to m; $L_{i,m}(\phi_{im}, R_{im})$ is the fraction (half) of transmission losses in the line between nodes i and m that has been assigned to node i and has therefore been represented as an extra load in this node: losses over a line depend on the flow and resistance of the line; θ_i is the phase angle at node i; x_{im} is line reactance

[55] Alternatively, under competitive market conditions, $B_i(d_i)$ can be understood as the aggregated selling bid function of all the demand d_i that is located at node i.

[56] Alternatively, under competitive market conditions, $C_i(g_i)$ can be understood as the aggregated selling bid function of all the generation g_i that is located at node i.

(typically, the dominant element in the impedance of a transmission line) between nodes i and m; $\overline{\phi_{im}}$ is the maximum flow allowed over the line from i to m in that direction; θ_{ref} is the reference phase angle; $\bar{g}_i$ is the maximum power output at node i and $\bar{d}_i$ is the maximum demand for energy (power) at that node. In addition, π_i, ξ_{im}, μ_{im}, β_i and α_i are the dual variables for the respective constraints, which, together with the primal variables, are obtained when solving the optimisation problem.

The nodal price at node k, ρ_k, can be shown [63] in this simple case to be the dual variable in the respective kth nodal balance equation, π_k.[57]

$$\rho_k = \pi_k \tag{C.2}$$

The system economic dispatch can be modelled using an alternative formulation, see the system of Eq. C.3, which represents exactly the same problem. Here, the transmission grid-induced ohmic losses can be included much more easily.

$$\begin{array}{lll} \max \sum_i \{B_i(d_i) - C_i(g_i)\} & & \\ s.t. & & \\ \max \sum_i \{B_i(d_i) - C_i(g_i)\}; & & \\ \sum_i (d_i - g_i) + L(d,g) = 0; & & \gamma \\ \phi_l = \sum_i PTDF_{i,l} \cdot (d_i - g_i); & \forall l & \xi_l \\ \phi_l \leq \bar{\phi}_l; & \forall l & \mu_l \\ g_i \leq \bar{g}_i; & \forall i & \beta_i \\ d_i \leq \bar{d}_i; & \forall i & \alpha_i \end{array} \tag{C.3}$$

where most symbols have been already explained. In addition, $L(d,g)$ represents the total transmission losses in the system as a function of power injections and withdrawals; $\text{PTDF}_{i,l}$, is the power transfer distribution factor for the flow over line l with respect to the power injection at node i (i.e. the sensitivity of the flow across this line to the power injected at this node); ϕ_l is the flow across line l and $\bar{\phi}_l$ the maximum amount of power allowed over line l in the direction of the actual flow in the scenario considered. Finally, when used as an index, l refers to the set of all the lines in the system.

With this formula, the expression for the nodal price (whose value does not depend on the choice of the slack node) is obtained as a combination of several dual variables present in the problem:

$$\rho_k = \gamma + \eta_k = \gamma + \gamma \cdot \frac{\partial L}{\partial d_k} - \sum_l \left(\mu_l \cdot \text{PTDF}_{k,l}\right) \tag{C.4}$$

where $\frac{\partial L}{\partial d_k}$ is simply the loss factor corresponding to node k, LF_k, mentioned in preceding items. Hence this is the same expression as in C.3.

[57] Strictly, the nodal price expression is $r_k = \pi_k + a_k$, although α_k is non-zero only at nodes where demand goes entirely unserved.

The formula for the optimisation problem proposed in C.3 depends on the choice of the slack node, s. Choosing a different slack node changes both the expression for the total losses, L, and the value of the PTDF parameters. The results (generation and demand dispatching, value of the objective function and nodal prices) are obviously not impacted, although the value of the dual variable γ is. Equation C.4 is consequently more suitably re-written as C.5, which is the same expression used in B.2.

$$\begin{aligned}\rho_k &= \gamma_s + \eta_{k,s} = \gamma_s + \gamma_s \cdot \mathrm{LF}_{k,s} - \sum\nolimits_l \left(\mu_l \cdot \mathrm{PTDF}_{k,l,s}\right) \\ &= \rho_s k \cdot \left(1 + \mathrm{LF}_{k,s}\right) - \sum\nolimits_l \left(\mu_l k \cdot \mathrm{PTDF}_{k,l,s}\right)\end{aligned} \tag{C.5}$$

The natural choice for the reference node s is the marginal generator in the system at any given time. Unfortunately, in general, there are several active network constraints at a given time in any power system, resulting in the simultaneous existence of several marginal generators, therefore rendering the choice of the reference node s an arbitrary decision.

6.10 Annex D Proof

It follows the simple proof of the statement in Sect. 6.3.1 on the equivalence of the outcome of transmission planning under traditional and competitive regulations.

Within the traditional approach, in general, the transmission network must be jointly optimised with the generation investments. Here, the objective is to maximise the consumer's welfare (utility function minus costs):

$$\max \{\mathrm{U(D)} - \mathrm{CFG} - \mathrm{CVG} - \mathrm{CT}\} \tag{D.1}$$

where U(D) is the utility function of consuming a demand D, CFG are the generation fixed costs, CVG are the generation variable costs and CT are the transmission costs (which can basically be considered as fixed costs).

When demand is assumed to be given and generation planning is also prescribed from the outset, transmission planning becomes the typical minimisation of generation operation costs via network reinforcement:

$$\min \{\mathrm{CVG} + \mathrm{CT}\} \tag{D.2}$$

Within the competitive approach the entity in charge of transmission planning (the independent system operator (ISO), typically, under regulatory supervision) must apply the following optimisation criterion in order to identify the network reinforcements that must be proposed to the regulatory entities for authorisation:

$$\max \{\text{Net benefit of consumers} + \text{Net benefit of generators}\} \tag{D.3}$$

where the total cost of any justified investments is implicit in these net benefits, as network charges to consumers and generators. In general, it is a good guideline in the design of the rules for competitive markets that the ideal outcome coincides with the one that the traditional approach would produce under the same circumstances. This is exactly what has been accomplished here, as it will be shown next.

In a competitive wholesale market, the following expression is always true.

$$\mathrm{PD} - \mathrm{IG} - \mathrm{IVT} - \mathrm{RNC} = 0 \tag{D.4}$$

where PD is the total payment (at wholesale level) of consumers, IG is the total income of generators (net of any network payments), IVT is the global variable income of the transmission network (based on application of nodal prices to both consumers and generators) and RNC is the residual network charge of the transmission network (i.e. the part of the total network cost CT that is not recovered by IVT).

The preceding expression allows one to replace the objective function of the maximisation problem in the traditional approach by this one that is entirely equivalent:

$$\{\mathrm{U(D)} - \mathrm{PD}\} + \{\mathrm{IG} - \mathrm{CVG} - \mathrm{CFG}\} + \{\mathrm{IVT} + \mathrm{RNC} - \mathrm{CT}\} \tag{D.5}$$

which shows that the maximisation problem in the traditional approach can be replaced by the following equivalent problem in the context of the competitive approach:

$$\max \{\text{Net benefit of consumers} + \text{Net benefit of generators}\} \tag{D.6}$$

since the transmission network is regulated so that $\mathrm{CT} = \mathrm{IVT} + \mathrm{RNC}$. Note that, embedded in the net benefits of consumers and generators are the complete payments for any justified investment in transmission facilities.

References

1. Alomoush MI, Shahidehpour SM (2000) Contingency-constrained congestion management with a minimum number of adjustments in preferred schedules. Electr Power Energy Syst 22:276–290
2. Anderson KP, McCarthy A (1999) Transmission pricing and expansion methodology: lessons from Argentina. Utilities Policy 8(4):199–211
3. Batlle C, Gómez-Elvira R (2011) Forward cross-border transmission capacity allocation: physical versus Financial Transmission Rights. IIT Working Paper, May 2011. www.iit.upcomillas.es/batlle/publications
4. Bialek J (1996) Tracing the flow of electricity. IEE Proc Gener Transm Distrib 143(4): 313–320
5. Boucher J, Smeers Y (2001) Towards a common European electricity market-paths in the right direction… still far from an effective design. Harvard Electricity Policy Group, Web page: http://www.ksg.harvard.edu/hepg/Standard_Mkt_dsgn/Smeers_Interconnections1_4jni_3.do1.pdf

6. Brunekreeft G (2003) Market-based investment in electricity transmission networks: controllable flow. CMI electricity Project paper. Applied Economic Departament de Economía Aplicada. Cambridge University. Web page: http://www.econ.cam.ac.uk/electricity/publications/wp/index.htm
7. Chao HP, Peck S (1996) A market mechanism for electric power transmission. J Regul Econ 10:25–29
8. Chao HP, Peck S (2000) Flow-based transmission rights and congestion management. Electr J 13(8):38–59
9. Chisari OO, Dal-Bó P, Romero CA (2001) High-voltage electricity network expansions in Argentina: decision mechanisms and willingness-to-pay revelation. Energy Econ 23:696–715
10. Coxe R, Leonardo M (2010) Survey of non-traditional transmission development. Paper presented at the annual general meeting of the IEEE power and energy society, Minneapolis, Paper 978-1, July 2010
11. De Dios R, Sanz S, Alonso JF, Soto F (2009) Long-term grid expansion: Spanish Plan 2030. In: CIGRE conference http://www.cigre.org
12. Dismukes DE, Cope RF III, Mesyanzhinov D (1998) Capacity and economies of scale in electric power transmission. Utilities Policy 7(3):155–162
13. EIPC (2010) Eastern interconnection planning collaborative. http://www.eipconline.com/
14. EOR, Ente Operador Regional (2005) Reglamento del Mercado Eléctrico Regional (The Regional Electricity Market Procedures, in Spanish). Libro III del RMER, De la Transmisión
15. ENTSO-E, European Network of Transmission System Operators for Electricity (2010 and 2012) Ten-year network development plan (TYNDP) 2010–2020, Chapter 8. https://www.entsoe.eu/system-development/tyndp/tyndp-2010/. Similarly for TYNDP-2012
16. ETSO, European Transmission System Operators (1999) Evaluation of congestion management methods for cross-border transmission. http://www.etso-net.org/, p 22
17. ETSO, European Transmission System Operators (2006) Transmission risk hedging products. solutions for the market and consequences for the TSOs. ETSO Background Paper, 20 April 2006. http://www.entsoe.eu
18. ETSO, European Transmission System Operators (2008) Overview of transmission tariffs in Europe: Synthesis 2007. ETSO Tariffs task force. http://www.etso-net.org/upload/documents/11.a.%20Final_Synthesis_2007_18-06-08.pdfS, p 27
19. European Commission, COM(2011) 658 final (2011) Proposal for a regulation of the European parliament and of the council on guidelines for trans-european energy infrastructure
20. FERC (2010) Transmission Planning and Cost Allocation by Transmission Owning and Operating Public Utilities, Docket No. RM10-23-000, June 17, 2010. Washington, DC. Retrieved from http://elibrary.ferc.gov/idmws/common/opennat.asp?fileID=12372947
21. FERC, Federal Energy Regulatory Commission (2011) Order 1000. Transmission planning and cost allocation
22. Galiana FD, Ilic M (1998) A mathematical framework for the analysis and management of power transactions under open access. IEEE Trans Power Syst 13(2):681–687
23. Gilbert R, Neuhoff K, Nwebery D (2004) Allocating transmission to mitigate market power in electricity networks. Rand J Econ 35(4):691–709
24. Helm D (2003) Auctions and energy networks. Utilities Policy 11(1):21–25
25. Heyeck M (2007) The next interstate system: 765-kV transmission. Electr Light & Power 85(1):32
26. Hogan WW (1992) Contract networks for electric power transmission. J Regul Econ 4:211–242
27. Hogan WW (2002) Financial transmission right formulations, Cambridge. http://www.ksg.harvard.edu/hepg, Center for Business and Government, John F. Kennedy School of Government, Harvard University
28. Hogan WW (2011) Transmission benefits and cost allocation. White paper. www.hks.harvard.edu/hepg/Papers/2011/Hogan_Trans_Cost_053111.pdf

29. IAEW, Institute of Power Systems and Power Economics and CONSENTEC (2001) Analysis of electricity network capacities and identification of congestion. Aachen, Consulting fur Energiewirtschaft und -technik. Report for the European Commission, Directorate-General Energy and Transport
30. Joskow PL (2005) Making transmission owners accountable. Panel on investments in transmission. In: The economics of electricity markets conference, Toulouse
31. Joskow PL (2006) Patterns of transmission investment. In: Lévêque F (ed) Competitive electricity markets and sustainability. Edward Elgar Publishing Limited, Cheltenham, pp 131–187
32. Joskow P, Tirole J (2002) Transmission investment: alternative institutional frameworks. Panel presentation at the wholesale markets for electricity conference, Toulouse, Francia, 22–23 Nov 2002
33. Junqueira M, daCosta LC, Barroso LA, Oliveira GC, Thome LM, Pereira MV (2007) An Aumann–Shapley approach to allocate transmission services cost among network users in electricity markets. IEEE Transactions on Power Systems 22(4):1532–1546
34. Kavicky JA, Shahidehpour SM (1997) Determination of generator siting and contract options based on interutility tie line flow impacts. IEEE Trans Power Syst 12(4):1649–1653
35. Kench BT (2004) Let's get physical! or financial? A study of electricity transmission rights. J Regul Econ 25(2):186–214
36. Kirschen D, Allan R (1997) Contributions of individual generators to loads and flows. IEEE Transactions on Power Systems 12(1):52–60
37. Kumar A, Srivastava SC, Singh SN (2004) A zonal congestion management approach using real and reactive power rescheduling. IEEE Trans Power Syst 19(1):554–562
38. Latorre G et al (2003) Classification of publications and models on transmission expansion planning. IEEE Trans Power Syst 18(2):938–946
39. Littlechild SC, Skerk CJ (2004a) Regulation of transmission expansion in Argentina part I: state ownership, reform and the fourth line. CMI electricity Project paper. Applied Economic Departament de Economía Aplicada, Cambridge University. http://www.econ.cam.ac.uk/electricity/publications/w
40. Littlechild SC, Skerk CJ (2004b) Regulation of transmission expansion in Argentina part II: developments since the fourth line. CMI electricity Project paper. Applied Economic Departament de Economía Aplicada, Cambridge University. http://www.econ.cam.ac.uk/electricity/publications/wp/
41. Lusztig C, Feldberg P, Orans R, Olsonet A (2006) A survey of transmission tariffs in NorthAmerica. Energy 31 (6–7):1017–1039
42. McDaniel T (2003) Auctioning access to networks: evidence and expectations. Utilities Policy 11(1):33–38
43. Méndez R, Rudnick H (2004) Congestion management and transmission rights in centralized electric markets. IEEE Trans Power Syst 19(2):889–896
44. MIT, Massachusetts Institute of Technology (2011) The future of the electric grid. http://web.mit.edu/mitei/research/studies/the-electric-grid-2011.shtml
45. Newbery DM (2003) Network capacity auctions: promise and problems. Utilities Policy 11(1):26–32
46. OFGEM, Office of the Gas and Electricity Markets, (2012) Project TransmiT. http://www.ofgem.gov.uk/Networks/Trans/PT/Pages/ProjectTransmiT.aspx
47. Olmos L (2006) Regulatory design of the transmission activity in regional electricity markets. Ph D dissertation, Universidad Pontificia Comillas
48. Olmos L, Pérez-Arriaga IJ (2007) Comparison of several inter-TSO compensation methods in the context of the internal electricity market of the European Union, Energy Policy. vol. 35, no. 4, pp. 2379–2389, April 2007
49. Olmos L, Pérez-Arriaga IJ (2009) A comprehensive approach for computation and implementation of efficient electricity transmission network charges. Energy Policy 37(12):5285–5295

50. Oren SS, Ross AM (2002) Economic congestion relief across multiple regions requires tradable physical flow-gate rights. IEEE Trans Power Syst 17(1):159–165
51. Pérez-Arriaga IJ, Rubio-Odériz F, Puerta Gutiérrez JF, Arcéluz Ogando J, Marín J (1994) Marginal pricing of transmission services: An analysis of cost recovery, IEEE Transactions on Power Systems. vol. 10, no. 1, pp. 65–72, February 1995
52. Pérez-Arriaga IJ (2002) Cross-border tarification in the internal electricity market of the European Union. In: Proceedings of the power systems computation conference (PSCC), Seville
53. Pérez-Arriaga IJ, Olmos L (2003) Network cost allocation in the internal electricity market of the EU: two main approaches for Inter-TSO payments calculation. Working paper, Universidad Pontificia Comillas
54. Pérez-Arriaga IJ, Olmos L (2006) Compatibility of investment signals in distribution, transmission and generation. In: Lévêque F (ed) Competitive electricity markets and sustainability. Edward Elgar Publishing Limited, Cheltenham, pp 230–288
55. Pérez-Arriaga IJ, Smeers Y (2003) Guidelines on tariff setting. In: Lévêque F (ed) Chapter 7 in the book 'Transport pricing of electricity networks'. Kluwer Academic Publishers, Boston
56. Pérez-Arriaga IJ, Rubio FJ, Puerta JF, Arceluz J, Marín J (1995) Marginal pricing of transmission services: an analysis of cost recovery. IEEE Trans Power Syst 10(1):546–553
57. Pérez-Arriaga IJ, Olmos-Camacho L, Rubio-Odériz FJ (2002) Report on cost components of cross border exchanges of electricity. Prepared for the Directorate General for Energy and Transport/European Commission, Madrid. Available at ec.europa.eu
58. Pérez-Arriaga IJ, Gómez T, Olmos L, Rivier M (2011) Transmission and distribution networks for a sustainable electricity supply. In: Galarraga I, GonzáLlez-Eguino M, Markandya A (eds) Chapter 7 in the book 'Handbook of sustainable energy'. Edward Elgar, Cheltenham
59. PJM Interconnection (2010) A survey of transmission cost allocation issues, methods and practices, Valley Forge
60. Rau NS (2000) Transmission loss and congestion cost allocation: an approach based on responsibility. IEEE Trans Power Syst 15(4):1401–1409
61. RealiseGrid project (2010) Working Package D3.3.1. Possible criteria to assess technical-economic and strategic benefits of specific transmission projects. http://realisegrid.rse-web.it/default.asp
62. Rivier M, Pérez-Arriaga IJ (1993) Computation and decomposition of spot prices for transmission pricing. In: Proceedings of the power systems computation conference (PSCC), Avignon
63. Rivier M, Pérez-Arriaga IJ, Luengo G (1990) JUANAC: a model for computation of spot prices in interconnected power systems. In: Proceedings of the 10th PSCC conference, Graz, 19–24 Aug 1990
64. Rubio-Odériz F (1999) Metodología de asignación de costes de la red de transporte en un contexto de regulación abierta a la competencia (Methods for transmission cost allocation under a regulatory context of open competition, in Spanish). Doctoral thesis Universidad Pontificia Comillas, Madrid (Spain)
65. Rubio FJ, Pérez-Arriaga IJ (2000) Marginal pricing of transmission services: a comparative analysis of network cost allocation methods. IEEE Transactions on Power Systems 15(1):448–454
66. Ruff LE (2000) Flowgates vs. FTRs and options vs. Obligations. http://www.ksg.harvard.edu/hepg
67. Ruff LE (2001) Flowgates, contingency-constrained dispatch and transmission rights. Electr J 14(1):34–55
68. Schweppe FC, Caramanis M, Tabors RD, Bohn RE (1988) Spot pricing of electricity. Kluver Academic Publishers, Boston
69. Stoft S (1999) Financial transmission rights meet cournot: how TCCs curb market power. Energy J 20(1):1–23

70. Tabors R, Caramanis M (2000) Real flow, a preliminary proposal for a flow-based congestion management system, Cambridge, MA. http://www.ksg.harvard.edu/hepg/flowgate/Real%20Flow.pdf
71. Tao S, Gross G (2002) A congestion management allocation mechanism for multiple transaction networks. IEEE Trans Power Syst 17(3):826–833
72. Vazquez C, Olmos L, Perez-Arriaga IJ (2002) On the selection of the slack bus in mechanisms for transmission network cost allocation that are based on network utilization. In: Proceedings of the power systems computation conference (PSCC), Seville
73. Wood AJ, Wollenberg BF (1996) Power generation, operation, and control, 2nd edn. Wiley, New York

Chapter 7
Electricity Generation and Wholesale Markets

Carlos Batlle

> *The market is not an invention of capitalism. It has existed for centuries. It is an invention of civilization.*—Mikhail Gorbachev

The generalized worldwide process of restructuring and liberalization of the electric power sector has primarily concerned the generation activity. Although liberalization and restructuring of the power industry has not been universally adopted, it is undoubtedly the prevalent framework in most countries.

In *The Wealth of Nations*, [31] contended that a free market economy is more productive and more beneficial to society. Smith noted that individuals, in the pursuit of their individual self-interests, interact on the market place guided by an "invisible hand" that inadvertently leads them to reach in the end socially optimum results. In a way, the market price acts as this "invisible hand" that drives the activity and ensures the efficient allocation of resources.

It is well known that, wherever possible, competition is beneficial because it places pressure on individuals to act more efficiently. In the context of electricity systems, this competition is not only expected to make suppliers to reduce costs but also help to naturally send sound economic signals to consumers. That is, consumers are made aware of the costs incurred to meet their demands.

However, the introduction of a competitive framework in electricity systems is not as straightforward as in other economic activities. The particular characteristics of the underlying commodity and the large diversity of typologies in electricity systems worldwide have led to the implementation of an enormous variety of

The author wants to thank Pablo Rodilla, Michel Rivier, and Ignacio Pérez-Arriaga for their support in the development of this chapter.

C. Batlle (✉)
Institute for Research in Technology, Comillas Pontifical University, Sta. Cruz de Marcenado 26, 28015 Madrid, Spain
e-mail: Carlos.Batlle@iit.upcomillas.es

C. Batlle
MIT Energy Initiative, MIT, Cambridge, MA, USA

C. Batlle
Florence School of Regulation, European University Institute, Florence, Italy

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_7,

alternative wholesale market designs. The objective of this chapter is to introduce and shed some light on this complex topic.

Whatever the reason that motivated the regulatory change, the success or failure of the new regulatory framework should be judged on the grounds of the overall efficiency achieved for the electric power industry.

7.1 The Necessary Steps Toward Competitive Electricity Markets

The key steps in electricity industry liberalization and restructuring include[1]:

(1) Privatization to enhance performance and reduce the ability of the state to use these companies as a mean to achieve costly political agendas.
(2) Unbundling: separation of the electricity businesses that can be conducted competitively (generation and retail) from the natural monopolies (transmission and distribution), which must be regulated.
(3) Horizontal restructuring to ensure competition (otherwise market power may put in danger the whole scheme, see Sect. 7.5).
(4) Designation of an Independent System Operator (ISO). This ISO would be responsible to maintain network stability and should ensure open entry to the wholesale market and full access to the transmission network.
(5) Establishment of a wholesale market where generators compete to supply electricity on an hourly, daily, weekly, monthly, and annual basis, or longer. This wholesale market also has to suitably integrate market-based mechanisms aimed to acquire operational reserves services (see Sect. 7.4).
(6) Unbundling of retail tariffs and rules to enable access to the distribution networks in order to promote competition at retail level. Open access to the retail market, so that all consumers can choose their supplier of electricity.[2]

This chapter focuses on step two and, in particular, step five.[3] The study of the wholesale market is essential to understand the workings of the electricity industry in a regulatory environment open to competition.

These steps are necessary but not sufficient to ensure an efficient electricity market. In certain circumstances markets may be unable to deliver optimum resource allocation, and this must be taken into account in electricity market design. Under such circumstances, normally termed market failures, regulatory intervention is required to maximize social welfare. Intervention mechanisms must be carefully analyzed in the wholesale market design phase, with a view to

[1] See also the textbook conditions proposed by Joskow [18].

[2] This is a necessary condition for the full liberalization, but it is not required for the implementation of a market mechanism at generation level.

[3] The last step, the liberalization of the retail business, is analyzed in Chap. 9.

distorting market operation as little as possible by regulation and verifying that the intended objectives are achieved.

7.1.1 The Multiple Dimensions of the Generation Business

The design of a stable regulatory framework for the efficient and reliable delivery of electric power at present and in the future is one of the major concerns of electricity market regulation policies. The type of regulation chosen for electricity production must wed sound economic criteria with the technological aspects of electricity generation, in a format that is compatible with the industrial structure and the legal context prevailing in each country. It must also cover all the considered timescales, from investment decisions made several years before plants come on stream to decisions made in real time to select the unit that must respond to an imminent change in demand.

Among many other things, the regulatory framework identifies the decision maker in each case. This is one of the key factors that differentiates the regulatory designs in place in different electric power systems.

In the present discussion, the generation activity is not addressed as a whole, but decomposed into stages in which the discriminating variable is the distance to real time. This leads to a division of three major phases in the electric power generation business:

- The first timescale involves decisions concerning the installation of additional generation capacity to replace obsolete facilities and cover demand growth. Depending on the type of regulatory framework, decisions in this timescale may be centralized in a government agency, incumbent upon vertically integrated utilities under the supervision of the regulatory authority, or left to the initiative of private investors.
- The second timescale revolves around the scheduling of existing generation capacity (a short- to medium-term issue, i.e., generator maintenance management, fuel supply contracts, hydro reservoir management, start-up schedules, and close to real-time economic dispatch of generation units). Here again, decisions may be made centrally or left to the initiative of producers and consumers, either in some organized fashion or bilaterally.
- The third timescale involves generators' short-term responses oriented to keep the generation-load balance (a short-term issue, i.e., the use by the System Operator of the various types of so-called operating reserves). In this phase of the generation business, maintaining security is the overriding concern and System Operators everywhere are in charge of centrally managing the system, which must necessarily also include the grid. In most competitive markets, the boundary between this phase and the preceding one is known as "gate closure". Thus, the timescale involved by this third phase includes all decisions between gate closure and real time, when electric power actually flows to the end user.

Regulation must strike a balance among three traditional objectives of electricity supply: economic efficiency, reliability of supply, and environmental impact. Each of these objectives can also logically be decomposed into the three timescales. All are interrelated with one another and across timescales.

A few general comments are in order here. In some countries the power industry has been restructured and liberalized fairly abruptly. This was typically the case in government-owned electricity utilities, such as in the UK or Argentina. By contrast, where utilities were primarily investor-owned, such as in the USA or Spain, the process was more gradual. Globally viewed, the traditional regulation paradigm has been transformed very slowly, step by step, starting from the longer timescale (free entry for small independent generation facilities) and evolving toward real time and ancillary service markets.

While the vast majority of the technical literature on electricity systems describes regulatory frameworks for markets fully open to competition in a rather matter of fact fashion, many of the electricity market designs in place around the world are somewhere in between a traditional centralized and a fully competitive model. As any number of experiences show, deregulation has not only been incomplete but in certain cases regressive, for some recently implemented regulatory mechanisms indisputably constitute a partial return to the centralized paradigm.

7.2 Expansion of Generation Capacity

As generation investment costs typically account for the largest share of the total cost of electricity, this is the area where the potential benefit from regulatory reform is greatest. Moreover, a country's economy is impacted very differently depending on whether investment in generation is made with public resources or by private investors. Privatization of formerly state-owned generation assets often results in much needed income for the national treasury. This is why regulatory change frequently goes hand-in-hand with change in the ownership and control of electric utilities, via privatization or corporatization[4] and the entry of private investors. However, changes in ownership must be distinguished from regulatory change per se.

7.2.1 A Gradual Process: From Independent Power Production to Unrestricted Entry and Access

The gradual exposure of traditional vertically integrated utilities to competition was initiated in response to concerns about energy security and independence from

[4] Corporatization refers to the transformation of state assets or agencies into state-owned corporations in order to introduce corporate management techniques to their administration. Corporatization is sometimes a precursor to partial or full privatization (see more in Wikipedia).

foreign energy sources as a result of the 1970s oil crisis. This and other concerns led the United States and other developed countries to enact rules that favored the development of renewable energy and cogeneration. The most significant was the 1978 Public Utilities Regulatory Policies Act (PURPA) in the USA. PURPA enabled a certain kind of non-utility electric power producers (so-called "qualifying facilities", essentially small renewable generators and cogeneration) to become alternative electricity suppliers by requiring electric utilities to buy power from these facilities at the "avoided cost" rate, i.e., the cost the electric utility would incur were it to generate or purchase from another source.

The precise definition of "avoided cost" was left to the discretion of each individual state. The resulting spectacular development of qualifying facilities in some states led to the need to establish mechanisms for the competitive selection of more efficient producers and establish conditions to govern their funding. The competition mechanism, which began to be used for this purpose in Maine in 1984, soon spread to other states.

Although the volume of generation from qualifying facilities was very small compared to generation from vertically integrated utilities, these new participants introduced a completely new scenario in the electricity industry. What had been a closed shop for many years, opened its doors to independent power production, albeit with very specific characteristics.

Other factors favored the advent of independent power production in the USA and elsewhere, which was not restricted to renewables or cogeneration only. Electricity prices under traditional regulation had consistently declined for many years, aided by technological developments and the fact that economies of scale had not yet been fully exploited. That also changed in the late 1970s and early 1980s, leading to more stringent regulatory reviews of utility costs. Delays in the regulatory process ("regulatory lags") and the occasional recognition by regulatory authorities of only part of companies' generation investment costs ("prudent and justified expenditures"), which were examined retroactively, raised vertically integrated utilities' risk, encouraging them to seek alternatives to cover the investment needed in generation.

When new investment was needed, these utilities began to buy the production required from third parties under a variety of instruments, subject to the approval of the regulatory authorities. The costs involved were then passed on to consumers with no risk to the incumbent utility. The conditions governing these arrangements were not established by PURPA or imposed by regulators, but were the result of inter-party negotiations and laid down in what were called "power purchase agreements" (PPA). Trading between the independent power producers (IPP) and any other parties except the vertically integrated utility were banned or severely limited.

Such independent power production became standard practice in the USA. Ad hoc methods to select suppliers were replaced by extremely detailed "competitive bidding" rules to determine the lowest cost generation portfolio. Finally, the 1992

Electricity Act allowed independent power producers[5] to trade freely in the power system and sell wholesale power anywhere to any vertically integrated utility or distribution company. This entailed open access to the transmission network. Selling to end consumers or "retail competition" was not allowed, however. This area of the market was liberalized years later at the individual state level. Other countries, such as Chile, UK, Norway, and Argentina got off to a later start but carried liberalization through to completion much more swiftly.

Similar transformations in developing countries were driven by a number of factors. After years of huge investment projects and subsidized rates, which were often insufficient to recover costs, state-owned electricity companies in many countries lacked the resources to continue investing. As a result, they resorted to independent producers, private companies keen on entering the generation business. Plant construction and operation was often tendered. The agreement stemming from the award was usually either a build-operate-transfer (BOT) or a build-operate-own (BOO) arrangement. Under the first, the plant had to be transferred to the incumbent company after a given term, while under the second the investor could retain ownership of the facility. Several electric power systems are presently organized around these arrangements today, Mexico being a paradigmatic example [19].

The massive entry of independent power producers in the electricity industry in the 1990s and the early twenty-first century was favored by an environment of declining interest rates, controlled inflation, liberalization of capital movements, and development of financial markets. The present prevalence of private investors in the electricity industry, hitherto mostly controlled by State companies in many countries, has brought fundamental change to the perception of risk and investment priorities.

7.2.1.1 Power Purchase Agreement Contracts[6]

A power purchase agreement (PPA) affords potential investors and financial institutions a sound legal guarantee. The lack of adequate legal safeguards increases risk, resulting in higher capital costs.

In traditionally regulated systems, generators' remuneration is governed by a set of laws. Generally known as a "regulatory contract", this arrangement is based on the Government's implicit commitment not to change the rules if the change is detrimental to business. A regulatory guarantee is by definition skewed, however, since regulators have their own objectives that may not necessarily include protection for producers. In the present context of frequent regulatory change the world over, international financial institutions do not generally consider this so-called "regulatory contract" a sufficient guarantee to grant funding under the most favorable conditions. The regulatory contract is based on a relationship of "trust"

[5] They were called "exempt wholesale generators" in the Act.

[6] This section is based on [8].

between the regulator and business agents, which does not typically exist with new entrants (especially if they are foreign companies).

But contracts are not even a full guarantee that problems will not arise. Political interference can nullify legal guarantees, albeit often with great difficulty. For instance, when short-term energy prices decline steadily for whatever reason, regulatory authorities and governments in particular are often tempted to renegotiate any existing long-term contracts, initially regarded as advantageous, but overly expensive for consumers in the new scenario.

PPAs are often only the most visible part of a much more complex network of agreements among manufacturers, equipment suppliers, maintenance contractors, fuel suppliers, financial institutions, insurance companies, consumers, and a long list of entities involved in the project. Ideally, each type of risk should be allocated to the parties best able to control it. The main objective of a PPA is inter-party risk sharing. The risks involved in generating electric power include variations in fuel prices, costs of labor and materials, unexpected plant failures, the uncertainty as to whether a contract counterparty will be found, and regulatory amendments.

At the very least, a PPA should specify a price per MW and a price per MWh. The former may consist of an annual charge to remunerate the generation plant for its nominal capacity plus its annual fixed operation and maintenance costs and is usually indexed to its availability record. The price per MWh is typically the cost of fuel times a plant-specific efficiency factor, plus a variable operation and maintenance item. The cost of fuel may be indexed to an international market reference price and the other costs to the consumer price index (CPI) or similar. The contract must include the definition of *force majeure*, i.e., situations beyond the control of the parties to excuse compliance with certain stipulations.

Because of their very nature, PPAs are usually very long-term contracts (at least ten or fifteen years), typically closely linked to the physical aspects of the project and inflexible in structure. They can be viewed as an extension of the traditional regulatory scheme, suitable for cases where the objective is not a drastic transformation of the existing regulatory framework, but only to enable private investors to enter the business to meet the need for new capacity.

7.2.2 Liberalization of Generation Investment: Opening the Door to Wholesale Markets

Chile made major progress toward the liberalization of generation expansion in the early 1980s. Prior to the Chilean initiative, the only experiences were the qualifying facilities in the USA and elsewhere, the incipient entry of independent generators as the result of auctions or negotiations with the incumbent company and the regulator, and PPAs. In 1981 a group of Chilean economists, inspired by the Chicago School of Economics, proposed and implemented radical reform of the Chilean power industry, with a view to introducing economic rationality and attracting foreign investment.

In essence, the reform established freedom of installation of generation facilities and a remuneration scheme based on market prices. The specifics of the Chilean regulation include features such as some manner of government intervention where private investment fails to meet needs and a mechanism for determining market prices outside actual market conditions at any given time. But for the first time in many years, it established an electric system in which private investors could freely install new plants and sell the electricity generated at market prices.

The first principle underlying this revolutionary reform, later adopted by many other countries, was deregulation of generation investment. In these systems, the free entry principle ensures that any investor is allowed to install new generation capacity, subject only to the customary legal obligations in connection with land use or environmental impact. Beyond that, all that is required is a licence or authorization that the authorities are obliged to grant indiscriminately to ensure fair competition among potential agents. Under such arrangements, the central planner is replaced by an unspecified number of decentralized planners. One possible objection that may be raised is that the optimal mix of generation technologies is more difficult to establish under this approach. Experience in other industries has shown, however, that efficient markets in which participants are sent appropriate economic signals typically reach suitable solutions, with less risk of committing the glaring errors often made by central planners. At least the possibility and the seriousness of error making is widely distributed.

In a competitive environment, companies decide to invest only when it makes economic sense, therefore eliminating vested political, industrial, or private interests that may be present in centralized planning. Here also, Adam Smith's "invisible hand" aligns the interests of the agents making decentralized decisions with the interests of overall economic efficiency, as discussed below.

The second basic principle of Chilean reform was to replace cost-of-service remuneration with market prices for the generated electricity.[7] This obviously raised the question of whether this regulatory scheme would be able to attract the necessary investment. For this scheme to work, the expected revenues from electricity sales and (to a lesser extent) ancillary services (see Sect. 7.4) would have to cover investment and operating costs, and provide for a reasonable rate of return on the capital invested.

This liberalized scheme for generation expansion planning was later adopted in a substantial number of electric power systems around the world. England and Wales, New Zealand, Norway, Argentina, and Colombia were among the first to follow suit. Many others took the same route in the late 1990s, including Spain, California, and New South Wales. Free entry to the generation business and market-based pricing were the two features common to all these systems.[8]

[7] In the Chilean scheme generators also participated in a capacity market, whose results were similar to the capacity payments later adopted in a number of Latin American countries.

[8] See Batlle et al. [5] for a description of the expansion of electricity systems in Latin America.

Marginal pricing and investment cost recovery

The marginal cost of production is a concept that plays a fundamental role in the analysis of competitive markets. As explained in Chap. 2, the use of marginal costs to compute market prices has its justification in microeconomic theory. Specifically, theory has it that on a perfectly competitive market, prices should equal marginal production costs.

In principle, ideally marginal energy pricing is the most suitable remuneration mechanism [32]. As microeconomic theory has shown, the short-term marginal energy price, defined as the production cost of responding to a unit change in energy demand, is the appropriate signal for attracting new investors. If the margin between installed available capacity and peak demand narrows, the price rises, therefore ideally providing the incentive for the entry of new investors.

However, the translation from costs to prices is not immediate, since the cost structure of a generator is far more complicated than just the addition of fixed investment costs plus variable costs proportional to production: plant operation is subject to a large number of technical constraints, limited energy plants (hydro reservoirs, and also storage facilities, as for instance batteries or pumping units) are difficult to manage, certain types of generation are highly unpredictable in the short term, such as run-of-the-river (non-impounded) and wind generation, or the output from CHP plants and finally, demand-side modeling is still only scantly developed.

This does not mean that marginal pricing is not the right choice in the case of electricity generation business. Pérez-Arriaga and Meseguer [27] demonstrate, by investigating the optimal economic signals that generators and consumers must receive in a competitive market under diverse circumstances, that even in the presence of a variety of planning and operation constraints for the generators, the approach is consistent with the goal of a correct regulatory policy: the maximization of global net social benefit. See Chap. 2 for a brief demonstration that, besides leading the system to the maximization of the operation efficiency, marginal prices allow investors to recover their investment costs (obviously only for those cases in which the investment is economically rational).

7.3 Generation Management and Scheduling

7.3.1 From Central Dispatch to Wholesale Markets

This section reviews the schemes in place for efficient generation unit dispatching, i.e., efficient demand coverage. Each approach to generation investment is characterized by a specific scheme for generator management and remuneration. These schemes range from the centralized optimization of generation dispatch under traditional regulation to arrangements provided for in the latest regulatory systems, in which no market price is calculated, and plant operation is the result of bilateral agreements based on supply and demand.

In this section, "generation management and scheduling" covers the following activities and decision making:

- long-term decisions (years) on plant transformation and overhauling, long-term fuel purchases and power sale contracts, plant maintenance programming, multi-annual reservoir management and nuclear fuel cycle management;
- medium-term (months to days) decisions concerning fuel, annual reservoir and pumping management, futures contracts for fuel and power.

Short-term (same-day) decisions to connect steam or hydroelectric generating units, overnight shut down management, hourly generator scheduling and operating reserves; decision making for shorter timescales, incumbent upon the System Operator, are considered in Sect. 7.5.

Regardless of whether this suite of decisions is made by a vertically integrated utility under traditional regulation or decentralized market agents, the reasonable underlying assumption is that, when shorter term decisions have to be made, the longer term decisions are already in place. For instance, when weekly decisions are made whether to connect or not some steam plants based on expected system behavior, the information on generating capacity, maintenance or the status of long-term fuel contracts is a given and not subject to change.

7.3.1.1 The Traditional Unit Commitment

For the sake of simplicity, this section focuses only on the so-called unit commitment problem, i.e., the determination of which units should be in operation and which should remain disconnected at any given time, to ensure that demand is met at the lowest possible cost, subject to any existing constraints. Given the highly detailed nature of such decisions, unit commitment is only applied within a timescale of about one day to one week before real time, depending on system characteristics.

More specifically, the unit commitment problem consists of supplying the estimated demand profile in a one day or longer horizon at the lowest cost, given each power plant's technical characteristics and cost functions. What has to be determined is which generators should be simultaneously connected to the system at any given time, i.e., when each should start up and shut down, and the distribution of total production among the units connected to the system in each time interval, usually hours or half hours.

When working so close to real time, system details are very important, and account must be taken of aspects such as steam plant generating unit start-up and shut down timing and costs, the hydrological restrictions in place in river basins, stations in tandem arrangement, demand chronology profiles, and the generating capacity to be held in reserve to respond immediately to fortuitous equipment failure. Other considerations weighing in these decisions include long-term hydroelectric management, system reserve capacity requirements, and network constraints that may render an economically efficient solution unfeasible.

Roughly speaking, assuming that demand is inelastic and that the marginal utility of demand is constant and hereafter referred to as the cost of the non-served energy, C_{nse}, [€/MWh], the problem to be solved can be expressed as the minimization of the cost of generation plus the total cost associated with the non-served energy nse, [MWh]:

$$\min(C_G + C_{\text{nse}} \cdot \text{nse})$$

where C_G, [€/MWh], is generation cost (obviously depending on the amount of energy generated).

For instance, an extremely simplified version of this minimization problem may be expressed as:

$$\min \sum_{i,t} (C_{i,t}^{\text{start-up}} + C_i^{\text{fuel}} \cdot g_{i,t}) + \sum_t C_t^{\text{nse}} \cdot \text{nse}_t$$

$$\text{subject to} \quad \sum_i g_{i,t} + \text{nse}_t = \text{d}_t$$

$$\text{Operational constraints} \begin{cases} g_{i,t} \geq \underline{g}_i \\ g_{i,t} \leq \bar{g}_i \\ \sum_t g_{i,t} = \bar{e}_i \\ \text{etc.} \end{cases} \quad \forall t, i$$

where for every unit i and time interval t, $C_{i,t}^{\text{start-up}}$, [€], denotes the start-up cost and C_i^{fuel}, [€/MWh], the fuel cost of each generating unit producing a given amount of energy, $g_{i,t}$ [MWh]; d_t is the demand, $\underline{g}_i$ and $\bar{g}_i$ are the minimum and maximum output of the unit and $\bar{e}_i$ represents the maximum energy available for the whole time horizon (e.g., the day) for the case of an energy limited plant (e.g., a hydro unit).

Under traditional regulation, centralized optimization is based on the cost structures of each generating unit, which are either supervised by the regulator or included in the PPAs. Where generating plants are remunerated under cost-of-service arrangements, each plant receives its annual fixed costs plus the variable costs deriving from the time it is actually operating. Plants with PPAs are paid according to the terms of their agreements and their production records.

7.3.1.2 The First Step: Wholesale "Markets" Based on Audited Costs

The electric power system reform implemented in Chile changed the former paradigm. Investment decisions are no longer incumbent upon the regulator and generators are remunerated on the grounds of system marginal cost. Fixed remuneration no longer exists.

The establishment of fair, transparent, and effective access to the transmission network and generating unit scheduling for all players was one of the primary hurdles to introducing competition in this activity. The two new entities created in the Latin American context at the start of the deregulation process for this purpose are defined below.[9]

- The System Operator (SO), which must be independent of generators and marketers, is responsible for managing the transmission network, essential to ensure effective competition, and the security functions, which in general also involve the generation plants.[10]
- The Market Operator (MO) is responsible for operation decisions that are based only on the economic data provided by generating units (the audited generation costs). In all the Latin American countries that adopted the Chilean model (all the countries that deregulated the industry, with the possible exception of Argentina, whose design is slightly different), this new institution, often called a market agent committee, consists of market agents' representatives (normally generators, retailers, distributors,[11] large consumers, the SO, and the regulator). It is responsible for operating the system and obtaining the unit commitment on the basis of the generators' declared costs/bids and calculating the system marginal cost for each time block (not necessarily hours, in some cases days or even weeks, as for instance in the Brazilian market, whose MO calculates weekly marginal prices).

This scheduling and operating model does not mean, however, that generating units are free to bid any "opportunity cost" whatsoever at which they are willing to generate electricity. The MO calculates the optimal schedule in accordance with the criteria described in the preceding section, but introducing two major changes: the cost structure of the conventional thermal plants assumed in the minimization problem is not the same as the structure laid down in long-term contracts; rather, they are what are termed as "audited costs". Moreover, hydro plants declare their inflows and reservoir levels and it is up to the MO to decide how the plants are to be operated, on the grounds of medium- or long-term optimization criteria (the same criteria that were in place prior to system reform).

As stated, in addition to scheduling, the MO calculates the time-blocks system marginal costs used as the (marginal) prices that remunerate all the generating

[9] In the schemes more open to competition that are presented later, the roles of the SO and MO differ from what is presented here.

[10] For instance, the European Commission [13] defines the tasks of the Independent System Operator stating that 'is responsible for granting and managing third-party access, including the collection of access charges, congestion charges, and payments under the inter-TSO compensation mechanism (…) is also responsible for operating, maintaining, and developing the transmission system. (…) has full responsibility for ensuring the long-term ability of the system to meet reasonable demand through investment planning (…) is responsible for planning, including obtaining the necessary authorizations and for the construction and commissioning of new infrastructure'.

[11] In their role of retailers of the consumers under regulated tariffs.

units in the system. Roughly speaking, these marginal prices are expected to be close to the cost of fuel for the marginal units (the most expensive units) committed in each time block. However, when binary (such as start-up costs), timing (e.g., the optimal operation of limited energy plants, such as hydro plants) or network constraints are taken into consideration, calculating this price is always complex and subject to considerable controversy.

This pricing mechanism, based on minimizing operation on the grounds of units' audited costs constitutes a major drawback. Since generating units cannot bid their opportunity costs (which roughly speaking would be the "avoided cost", i.e., the cost of the next more expensive unit), peak-time units would only be dispatched at a higher price than their marginal costs (to enable them to recover their investment costs) in the event of scarcity. In such a scenario, the system marginal price should be the aforementioned cost of non-served energy. But this is not actually the case. The actual value used is always capped by the regulator at a much lower level than would be required to enable the peak-time units to recover their investment costs.

To remedy this shortcoming, the Chilean model added a supplementary mechanism to the system marginal calculation, the so-called "capacity market" (which involves a demand-side obligation to hedge expected peak consumption for two years in advance under an agreement with a generating unit), in an attempt to provide additional remuneration. In Argentina, a capacity payment was devised instead of a capacity market (see Chap. 13 for a brief description of the design initially implemented in Argentina).

As stated, this scheme, based on the audited costs of generating units (and hydro inflows and reservoir levels), is currently implemented in most Latin American countries that reformed their electricity industries. Colombia is possibly the only one where scheduling is based on daily bids (and where generators submit a single bid for the entire day).

The full deregulation of the generation business, however, called for an even bolder move. This was introduced with the compulsory pool model discussed in the following section, in which scheduling follows the same rules, but generators are free to establish their own opportunity costs and therefore to bid the price at which they are willing to be committed.

7.3.1.3 The Second Step: From the Cost-Based Prices to the Bid-Based Prices

Pioneers

Ten years after the Chilean reform, in the early 1990s (1991–92), England and Wales and Norway took the system one step farther. These two markets differed in a number of ways, initially with respect to the nature of the underlying electric power systems: while the E&W generating units are almost all steam facilities (which means that unit commitment is subject to many technical constraints), the Norwegian system is based almost entirely on hydro production.

Out of the many aspects of these innovative designs that might be highlighted, the most prominent features, common to both, are considered here. To begin with, the system marginal price was based not on costs but on bids freely submitted by the generating units (freely does not, of course, imply lack of regulator supervision).

England and Wales

The market in England and Wales [36] was initially a compulsory day-ahead market (electricity could only be sold on the pool) whose clearing mechanism was the same optimization algorithm that was used prior to system liberalization to dispatch generating units for the following day, at half-hour intervals.[12] Under the former arrangements, the model inputs were the data furnished by the generating units on their technical constraints (such as technical minima and ramps) and production costs (such a start-up costs and fuel costs expressed as a piece-wise linear function). The calculation was based on a classic unit commitment and the system marginal price was computed for each half hour. Network constraints, however, were ignored by the model and subsequently treated in a simplified manner.

Price calculation in the England and Wales Electricity Pool[13]

In the England and Wales Electricity Pool all generators were paid the same price every half hour, which was computed *ex ante*. As the unit commitment model that was used was deterministic, an additional term was necessary to account for the possibility that available generation could not meet demand at each half hour of the next day. Therefore:

$$\text{Marginal price} = \text{SMP} \times (1 - \text{LOLP}) + \text{VOLL} \times \text{LOLP}$$

where SMP is the system marginal price computed with the deterministic GOAL model; VOLL is the value of lost load for the consumers, i.e., the value established by the regulator for the system marginal price when there is non-served energy; LOLP (separately estimated with another model) is the probability that available generation cannot meet all the demand during the considered half hour and 1-LOLP is, obviously, the probability that all demand is met. Thus, the equation above computes the expected value of the marginal price for the system for the considered half hour when calculated one day in advance.

SMP is computed as "the highest genset price". Once the unit commitment has been determined for the next day, for each generator a marginal price is computed at each half hour of the day that allows the generator to recover its operating cost, where any start-up costs and no-load costs (i.e. all nonlinearities in the generator's cost function) are included. For instance, the start-up costs

[12] It was a heuristic model named GOAL, later replaced by another one that was based on Lagrangian relaxation.

[13] This description was written by Prof. Pérez-Arriaga.

are distributed during all the non off-peak demand hours. Then SMP at each half hour equals the highest of all the generator prices for that half hour. In this way it is made sure that all generators recover their operating costs.

The computation of LOLP happened to be clearly biased and systematically resulted in high values that did not correspond to reality. During the years that this scheme of remuneration was used, there never was a case of non-served energy due to lack of available generation. However, generators obtained substantial income from the VOLL × LOLP term. Moreover, some companies apparently manipulated this mechanism, by declaring unavailable units at critical times and causing the algorithm to result in artificially high values of LOLP [21]. Many people have classified the VOLL × LOLP term as a capacity mechanism. Conceptually it is only one necessary element in the correct computation of the ex ante marginal price. However, when misused, as it was the case in the UK, it became a systematic extra remuneration for generators, which was linked to situations of system stress, and its de facto became a capacity instrument.

On top of SMP, consumers had to pay an uplift charge, which covered the generation costs incurred because of transmission constraints (which were ignored by GOAL) and the payments of some of the ancillary services. As it has been seen, the England and Wales pool, although a competitive market at the core, contained a fair amount of regulation.

On March 2001, the Electricity Pool model of England and Wales was replaced by a fully bilateral model, the New Electricity Trading Arrangements (NETA), which in 2005 became BETTA (British Electricity Trading Transmission Arrangements), with the integration of Scotland. The electricity trading arrangements in England and Wales are now managed by ELEXON, the Balancing and Settlement Code Company.

As for example described by Green [15] 'the guiding principle of NETA's (and hence BETTA's) design was that electricity should be treated as much like a "normal" commodity as possible, while still recognizing the physical characteristics of electricity. This means that there is a balancing mechanism run by the system operator, National Grid, to ensure that demand and generation are kept in balance and transmission constraints are respected, but no other market was centrally organized. Instead, most electricity is traded bilaterally (or internally, for integrated firms), with some trading on electronic exchanges to aid transparency'.

Norway

The market established in Norway in 1993 (Statnett Marked AS[41]) differed in a number of meaningful ways. First, it was implemented as a voluntary and purely

[14] Currently NordPool (www.nordpool.com), owned by the national grid companies Fingrid, Energinet.dk, Statnett, Svenska Kraftnät, provides a marketplace for trading both physical and financial contracts in Finland, Sweden, Denmark, and Norway.

financial market on which generating units can trade their energy in different timescales, up to the day-ahead spot market, on which financial positions are cleared. As it is later explained in more detail, bids are not fully simple (mainly quantity/price bids only, unrelated to any technical constraint, since, as later described, a number of semi-complex conditions have been progressively allowed).

These two approaches, the E&W and the Norway wholesale markets, represent not only the pioneering experiences but also probably the two opposite poles regarding the extent to which the resulting wholesale spot market design resembles the traditional centralized paradigm.

These spot (or day-ahead) markets based on an organized daily auction (plus in some cases a free bilateral market in which market agents agree to match their production/consumption programs) play a central role in the liberalized context for they determine the expected generation schedule for the day after.

Electricity wholesale market sequence

The electricity wholesale market is composed of all the commercial transactions of buying and selling of energy and also other services related to the supply of electricity (the so-called ancillary services, which are essential for this to occur in adequate conditions of security and quality, see Sect. 7.4). These transactions are organized around a sequence of successive markets where first market agents (supply and demand) trade energy, and then, the SO acquires from these agents (mainly from the supply) the above-mentioned ancillary services products related to the supply of electricity in periods closer to real time.

The overall trading timetable covers a number of timescales: months or years before a trade is to be implemented; "gate closure"; real time when the transaction takes place; and post-transaction settlement.

The generation and load parties must notify the SO of their expected physical schedules at real time by market gate closure (one day, one hour, or possibly less before real time). One of the many ways of splitting this sequence of markets and transactions is into the following categories:

- long-term markets,
- day-ahead markets (DAM), and
- intraday plus balancing markets (in the EU) or real-time (in the US) markets.

Additionally, the SO acquires operating reserves (for example, secondary reserves or 10-min spinning reserves, see description on ancillary services later on) in different timescales, sometimes in the long term (e.g., with two years in advance) or once the energy market closes.

When the market structure is competitive and opened enough (what naturally leads to significant levels of volatility and liquidity), a financial long-term market arises. The primary purpose of these long-term markets is to provide hedging mechanisms for producer and consumer entities. But they can also be used by arbitrageurs and speculators, who critically contribute to market liquidity.

These long-term markets function prior to the day-ahead auction. The morning of the day before (D-1) those agents that have not previously committed their supply in a bilateral contract, submit their offers and bids to the market operator, who clears the auction and gets first preliminary schedule results for the day after.

When the MO does not run an extensive nodal auction (i.e., when the transmission constraints in all detail are not considered when clearing the auction, see description below), the SO checks that the schedule resulting from the declared bilateral contracts plus the DAM is feasible. If there is any transmission constraint the SO solves the constraint at the least possible cost and gets the final feasible schedule.

Once the DAM feasible schedule is known, additional mechanisms have to be implemented to allow market agents themselves or the SO fix any deviations from this program that might occur, either because the schedule resulting from the DAM is not feasible for a generating plant[15] (only in the case that simple bids are considered, see explanation in the next section) or because for some reason the plant cannot perform as expected. This rebalance can be done right after the DAM schedule is finished, but others might need to be done later in the day D-1 or a few hours before real time. For instance, the weather forecast might happen to be inaccurate, so wind producers might be interested in selling more or less than what it was estimated at the time the DAM closed on D-1.

In power system operation, once the market is closed, the SO must ensure that supply matches demand in the real time. This task calls for ceasing any further market transaction (at some point in time at which the SO considers there is no time enough for agents to react efficiently) and leaving all the power system control to the SO. Thus, the regulator together with the SO needs to determine the point in time at which this economic trading should be over; a time called "gate closure". Until gate closure the market agents are allowed to balance their positions and correct their deviations without any type of intervention of the SO. After gate closure, the final production schedule is determined for all participants and only the SO can act to adjust any deviation.

- In some cases, (in the EU) subsequent trading is allowed within the day (centralized in the so-called intraday markets organized by the power exchange) and the gate closure is set very few hours ahead of real time. At gate closure, market agents are supposed to have submitted their balancing bids (upwards and downwards) for the so-called balancing market run by the SO. This auction determines the least-cost resources for the SO to fix the potential imbalances.
- In other cases (namely in the US markets), once the day-ahead market clears in the afternoon of the day before, generating units have to submit their bids for both the so-called real-time energy market (similar to the balancing bids in the

[15] These deviations may be due to several reasons. For instance, a thermal unit could have bid expecting to produce in four consecutive hours the day after, but in the final DAM schedule it has been committed in eight hours. If the unit just counts on fuel for the expected four hours, the generator will require a way to readjust the schedule, in order to purchase the committed supply in the four hours in excess.

EU model) and the regulation market (similar to the reserves market in the EU model, providing AGC services, see description of Ancillary Services below in the chapter). These are the tools for the SO to correct the imbalances and to maintain the system stability. Instead of clearing the market agents bids (expressed as just selling and buying quantity-price pairs), the SO runs an optimization tool considering all the technical constraints of the different units (the so-called Security Constrained Economic Dispatch) and calculates prices for each five-minute interval. Gate closure is in some cases set the day before (thus, unless exceptionally justified, generating units cannot trade or modify their bids after the market closes) as at the time of this writing it is the case in PJM or seventy-five minutes before real time, as in NYISO.

Finally, in both models, in most systems the SO acquires in the long-term other ancillary services, namely very short-termed reserves that might be needed to respond to very specific contingencies. These complementary reserves are described later in more detail (Fig. 7.1).

The market implemented in East Australia works in a slightly different way. As taken from the introductory report from the Australian Energy Market Operator, AEMO [2], daily bids are submitted before 12:30 pm on the day before supply is required, and are reflected in pre-dispatch forecasts. Generators may submit rebids up until approximately five minutes prior to dispatch. In doing so, they can change the volume of electricity from what it was in the original offer, but they cannot change the offer price. AEMO issues dispatch instructions to generators at five-

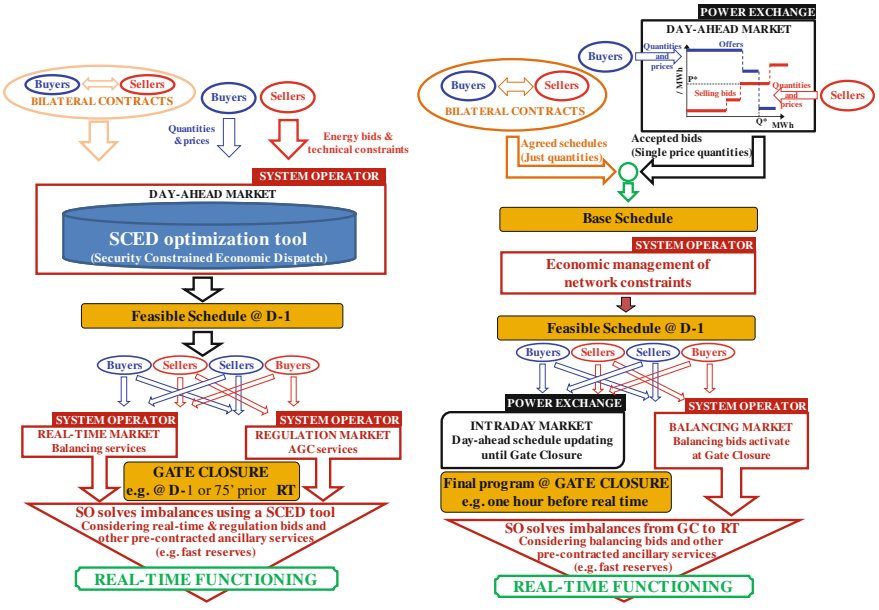


Fig. 7.1 Market sequence in the US (*left*) and in the EU (*right*) models

minute intervals throughout each day based on the offers that generators have submitted in the bidding process. In this way, there are 288 dispatch intervals every day.

Next we review in larger detail the main design characteristics of the components of these energy markets.

7.3.2 Long-Term Markets: Over-the-Counter and Futures Markets

Prior to the market gate closure, generating units, suppliers, and qualified consumers can always trade freely their future supply needs. Long-term (in most cases one-year and typically not more than two years) contracts—in all their diverse formats—are the dominant form of transaction in wholesale electricity markets. Most market participants do not want to be subject to the uncertainty of the short-term electricity prices in liberalized markets, and long-term contracts provide hedging against this risk. In addition, speculators without any interest in physically buying or selling electricity may wish to bet against the uncertainty of the short-term electricity prices, and participate in organized markets selling all sorts of financial products with the short-term price of electricity as the underlying reference price.

Long-term contracts can adopt two basic formats: either physical or purely financial contracts. Physical contracts entail physical and cash delivery on expiry. The delivery point is the high voltage grid in general or some prescribed node in it—in some cases some important node or "hub" that is chosen as reference; in other cases the node where the buyer is located. Note that the selling party in a physical bilateral contract does not necessarily have to produce electricity (although this is the most frequent case); it suffices with purchasing electricity from other participants or in a short-term market and making sure it can reach the agreed delivery point. In some power systems it is possible to purchase a transmission right to transport the scheduled volume of power from the generation node to the agreed delivery point, when bottlenecks between the two nodes are expected, so as to reduce the risk that the agreed quantity may not be supplied by the contracting generator. Ideally, (as this depends on the specific regulation of the power system or systems involved) physical contracts provide the guarantee that the power will be delivered at the consumption point if the generator is producing and (if this is the case) the contracted physical transmission link is available, regardless of any situation of scarcity that may happen in the concerned power system(s). In the short-term the involved system operators must approve the previously declared schedule of the transaction, in order to prevent any violation of operating constraints. The system operators may put in place some mechanism for the settlement and management of any real-time imbalances between the declared and the actual transaction schedule.

On the other hand, a purely financial contract only entails cash delivery on expiry, when the differences between the specified reference index and the contractual price are settled. For instance, in the contract termed "contract for differences" the buyer pays the contractual price to the seller for the contracted volume, and the seller pays the reference index to the buyer for the same volume. The hourly or half-hourly spot prices of the day-ahead market at the power exchange are normally selected as the reference index. Market liquidity is important to provide a reliable index. These contracts are not related to the physical dispatch of the plants or to any actual consumption; in fact these financial contracts can be signed by any legal entity without any physical relation to the power sector. See below a case example that shows the implications of this type of contract.

Trading may be conducted essentially under one of two schemes: bilateral over-the-counter (OTC) contracts, and organized futures markets.

Over-the-Counter markets

In the OTC market model, each pair of counterparties reaches an agreement and concludes their trades independently. Generators and suppliers negotiate their contract terms and electric power is physically transmitted: these *forward contracts* are physical contracts, they influence actual dispatching, and they take place outside any organized market.

Brokers, organizations that bring buyers and sellers together (centralizing purchase and sale offers), also operate on these markets, but do not centralize default risk.[16]

There is no "single price", but organizations such as Dow Jones and Platt's attempt to compile information on the closing prices in OTC transactions and publish indices that players can use to establish the price of their operations.

One of the advantages of bilateral contracts is that they accommodate customized formats to match counterparty requirements, therefore reducing the basis risk.[17]

However, this advantage also entails drawbacks, because the more highly customized the contract, the higher the cost of possible assignment to a third party. To facilitate such transactions, the European Federation of Energy Traders (EFET) has setup a standard master agreement for the delivery and acceptance of electricity (the General Agreement) [9] in an attempt "to improve the conditions of energy trading in Europe and… promote the development of a sustainable and liquid European wholesale market".

---

[16] Default or credit risk is the risk that a counterparty will be unable to meet its obligations, i.e., the risk that the counterparty will default on its contract over the life of the obligation.

[17] Basis risk in finance is the risk associated with imperfect hedging using futures. It may arise due to the difference between the price of the asset that is to be hedged and the asset underlying the derivative, or to a mismatch between the futures expiration date and the price of the actual selling date of the asset (www.wikipedia.com).

Table 7.1 PJM Western Hub Peak Calendar-Month Real-Time LMP Swap Future

Contract Unit	80 Megawatt hours (MWh) (5 MW per peak hour)
Price quotation	The contract quantity shall be 80 Megawatt Hours (MWH) and is based on 5 megawatts for peak daily hours. Transaction sizes for trading in any delivery month shall be restricted to whole number multiples of the number of peak days in the contract month
Minimum fluctuation	$0.05
Floating price	The Floating Price for each contract month will be equal to the arithmetic average of the PJM Western Hub Real-Time LMP for peak hours provided by PJM Interconnection, LLC (PJM) for the contract month
Termination of trading	Trading shall cease on the last business day of the contract month
Peak days	"Peak day" shall mean a Monday through Friday, excluding North American Electric Reliability Corporation holidays
Peak hours	From Hour Ending (HE) 0800 Eastern Prevailing Time (EPT) through HE 2300 EPT

http://www.cmegroup.com/trading/energy/

Futures markets

Derivatives (*futures* and *options*) contracts are traded on a commodity exchange where the delivery date, location, quality, and quantity have been standardized. A derivative is a standardized contract where all terms associated with the transaction have been defined in advance, leaving price as the only remaining point of negotiation [34]. Standardized contracts allow participants to benefit from market liquidity and transparency, trading anonymously. *Future contracts* are standardized forward contracts traded in organized exchanges that, contrary to forward contracts, are typically of a purely financial nature. *Option contracts* confer the buyer, against the payment of a fee, the right, but not the obligation, to purchase ("call options") or to sell ("put options") a specified quantity of electricity at a specified time in the future, at a predetermined price (the strike price). Option contracts are purely financial contracts.

For example, The New York Mercantile Exchange (NYMEX) offers financially settled monthly futures contracts for on-peak and off-peak electricity transactions based on the daily floating price for each peak day of the month at the PJM western hub. In Table 7.1 an extract of the specification of one of the contracts traded in this market is shown.

On these organized exchanges, transactions are concluded by settling the difference between some contracted strike price and an index price (typically the day-ahead spot market price). The exchange takes the position of central counterparty to all operations it has registered, guaranteeing the fulfilment of obligation of both parties. Once an operation is registered the exchange manages the resulting positions, through its interposition as (central) counterparty of the operations, becoming the buyer in relation to a seller and a seller in relation to a buyer, and

Table 7.2 Characterization of over-the-counter and organized contracts

Properties	Trading method	
	Over-the-counter	Power exchange
Anonymity of trading	No	Yes
Counterparty	Bilateral	Central counterparty
Counterparty risk	Yes, unless cleared	No
Trading method	Continuous trading	Either continuous or central auction

therefore allowing contract fungibility and eliminating counterpart credit risk.[18] The exchange requires counterparties to advance an initial amount of cash, the margin. Every day the contract is "marked to market": the exchange's clearing house records the price to value the contract, in order to reflect its current market value rather than its book value. If the current market value causes the margin account to fall below its required level, the trader will be faced with a margin call. The exchange will draw money out of one party's margin account and put it into the other's so that each party has the appropriate daily loss or profit.

Exchange derivatives prices are widely and instantaneously disseminated. It follows from the latter that the index price must be credible, reflecting the price at which power may actually be bought and sold on the market (which should be "liquid") at any given time (Table 7.2).

A case example: Contracts for Differences (CfDs)

CfDs are the best-known example of a risk hedging financial instrument. CfDs are two-way financial contracts that can be traded bilaterally or in an organized power exchange. CfDs specify the amount q_c of contracted energy (either with a flat profile or with a prescribed pattern during the contract period) and the contract price (strike price) P_c. The reference price P_m is the hourly or half-hourly spot price of the power exchange. There is no option fee or risk premium in this type of contract, since both participants agree on the strike price.

The monetary outcome of the contract is that the party with the consumer role receives from the party with the generator role the amount $(P_m - P_c) \cdot q_c$ This amount is positive when $P_m > P_c$, i.e., when the market price P_m is higher than the contract price P_c and it is negative in the opposite case.

A CfD has interesting consequences for both parties. Let us assume first that one of the parties is a true consumer with an expected flat demand q_c. Then, if the actual demand q of this consumer is precisely q_c, this consumer will be fully hedged with an acquisition price P_c against any volatility in the market price, since, if it happens that the spot market price P_m is higher than the contract price P_c, so that the consumer will have to pay $q_c \cdot P_m$ in the market, the CfD will make the generator to pay $(P_m - P_c) \cdot q_c$ to the

[18] See for instance www.omip.pt.

consumer, with a net value of $P_c \cdot q_c$ regardless of the market price. A similar reasoning can be made if $P_m < P_c$, with the net result that the consumer is fully hedged *if its consumption is precisely the contracted amount* q_c. What happens if the consumer deviates from the scheduled demand q_c, for instance with an actual $q > q_c$? Since the CfD only covers the amount q_c, any extra consumption $q - q_c$ will have to be paid at the current spot market price P_m. And, if $q < q_c$, the consumer will pay the actual consumed energy q at the market price Pm, although the CfD will hedge this purchase—so that the net charge will be $q \cdot Pc$—and the remainder of the CfD will result in a payment by the generator to the consumer of the amount $(P_m - P_c) \cdot (q_c - q)$. This amount may be positive or negative, depending on the actual values of P_m and P_c. Note that, if the market price happens to be very high ($P_m \gg P_c$), the consumer has the incentive to close all non-essential loads (e.g., shut down an industrial production facility) and reduce its actual load q as much as possible, of course depending of its utility function. On the other hand, if the market price is low ($P_m < P_c$) the consumer has the incentive to increase q as much as possible, exactly as it would be done in the absence of the CfD.

A parallel discussion can be expounded for a true generator that has signed the CfD. If the generator produces the contracted amount q_c, the CfD will provide a complete hedge against any variability of the market price P_m and q_c will be sold at the price P_c. However, any production above q_c will be priced at the market price P_m and, if the production q is lower than q_c, the actual production q will be hedged at the price P_c but the generator will be financially exposed to the risk of having to pay to the consumer the amount $(P_m - P_c) \cdot (q_c - q)$, which could be positive or negative. The decision-making process for the generator requires to include also the variable cost of production VC into the picture. The net income for the generator—including the market remuneration, the CfD, and the variable production cost—is:

$$q \cdot P_m - (P_m - P_c) \cdot q_c - q \cdot \mathrm{VC} = q_c \cdot P_c + (q - q_c) \cdot P_m - q \cdot \mathrm{VC}$$

meaning that the generator is subject to the market price P_m for any deviations with respect to the contracted value q_c. Since any production has per unit cost VC, the generator must produce as much as possible if $P_m > \mathrm{VC}$ and shut down otherwise. The economic incentives are the same as in the absence of the CfD, but now the generator has a hedge for a production equal to q_c.

In summary, both the consumer and the generator are perfectly hedged against the volatility of the market price P_m, but only for the contracted amount q_c; any deviations are valued at the current spot market price. Therefore, a purely financial CfD has the property of hedging against the market price for the contracted amount, without impairing the incentive of the market agents to react to the real-time value of the spot market prices.

The situation is very different for speculators who have signed a CfD, perhaps because they thought that they had some a priori knowledge that the future market price will be higher or lower than the strike price P_c, or for any other reason. When $P_m > P_c$ the speculator in the generator role will have to pay $(P_m - P_c) \cdot q_c$, while an actual generator that produces q_c will receive $q_c \cdot P_m$ from the market to compensate that loss. The same happens with the consumer when $P_m < P_c$, so that the consumption of electricity ends up being priced at P_c. In principle, speculators have no intrinsic hedge and they are subject to the full risk of the CfD.

Now, we present an alternative and much simpler view of the same situation. Since the CfD happens in parallel with the spot market, the economic implications of the market and the CfD can be examined separately. In the spot market the true consumer has to pay $P_m \cdot q$, whatever the value of q, while the CfD will provide an amount $(P_m - \mathrm{P_c}) \cdot q_c$, positive or negative, *which the consumer cannot modify* once the CfD has been signed. It is therefore straightforward to see that the consumer is always subject to the incentive of the spot market price, with the consumer being able to respond by decreasing or increasing consumption, regardless of the economic outcome of the CfD, which is independent of the consumer's actions. The same reasoning applies to the generator. This simple and direct way of analyzing this situation is one of the countless applications of Coase's Theorem. A folk version is presented in Annex B of this book, under the name of "Grandma's inheritance theorem". It is the experience of the authors of this book that the need for application of this theorem appears very frequently in regulatory decisions, rendering them much easier than when grandma's advice is ignored. Numerous regulatory authorities have adopted wrong decisions by ignorance of this basic economic principle.

7.3.3 Day-Ahead Markets

The core activity of the wholesale electricity markets is the day-ahead market (DAM), where trading takes place on one day for the delivery of electricity the next day. Market members submit their orders electronically, after which supply and demand are compared and the market price is calculated for each hour of the following day.

The design of DAMs has been permanently evolving over time. Different countries have developed a variety of market models, in such a way that it is difficult to classify them all into a small number of categories. Next, instead of trying to propose a classification of the known-to-date designs, we will review the different elements and features that characterize the design of a DAM.

7.3.3.1 DAM Organization: Role in the Wholesale Market

Attending to the type of implementation, wholesale markets have traditionally been classified into two major groups, corresponding to two different conceptions on how full liberalization of the generation activity can be carried out (originally represented by the two pioneering models previously described): the so-called Electricity Pool model and the one we will denote as Power Exchange (PX) model.

The Electricity Pool model

The Electricity Pool model (as used here) means a highly centralized market (originally although not necessarily compulsory[19]) run either by the System Operator or by a Market Operator. At this stage, one of the key differences with the PX model is that these institutions operate on a cost recovery basis, recovering their operating costs through fees (administratively approved by the regulator) paid by market participants. PXs also charge fees, but in principle they are just subject to the regulator supervision and no cost recovery is guaranteed.

As previously mentioned, this alternative was the one originally implemented in the UK market back in the early 1990s. At the time of this writing, some of the main examples that can be included under this category are the wholesale power markets in North America (both in the US and Canada), Australia (e.g., the Australian Energy Market Operator), New Zealand, and in Europe the Irish market and to some extent the Iberian one, since, although most of its features are closer to the PX model, its fees are still determined by the regulators.

The PX model

Power Exchanges operate in an open trade context, in which the generating units' scheduling is decentralized, so market agents can either bilaterally engage into any type of agreement for the delivery of energy (in the so-called bilateral market) and then declare their production/consumption schedule directly to the System Operator at the market gate closure or can submit bids for buying and selling power to the trading platform of a PX. These organized markets are optional and anonymous and accessible to all participants satisfying admission requirements in exchange of a fee.

Ideally, the main objective of power exchanges is to ensure a transparent and reliable wholesale price formation mechanism on the power market, by matching supply and demand at a fair price, and to guarantee that the trades done at the exchange are finally delivered and paid.[20]

---

[19] As pointed out by Boisseleau [7], this model has been mainly the result of a public initiative and the participation has been usually mandatory (or highly encouraged). In some cases these markets are not strictly but de facto compulsory, since generators with capacity obligations are required to submit bids to the day-ahead market. At the time of this writing (end of 2011), some of the main compulsory pools left were the ones implemented in Ireland, Alberta, Australia, and New Zealand.

[20] See for instance the web page of the European Power Exchange, www.epexspot.com.

7.3.3.2 Bidding, Clearing and Pricing

The manner in which the different market mechanisms are cleared and the products are priced typically differs from one electricity spot market to another.

Roughly speaking, in organized short-term electricity markets the day-ahead market prices are, in principle, determined by matching generators offers and consumers bids. However, this can be achieved in a number of different ways.

We find three major features that characterize short-term electricity auctions:

- Whether they use complex bidding or simple bidding;
- Whether the pricing rule is discriminatory or non-discriminatory;
- Whether single, zonal, or nodal prices are computed.

A number of other aspects could also be distinguished [4]: the trading intervals used (hourly, half hourly, or even every five minutes), if portfolio bidding is allowed or not (i.e., if no link is required between bids and units or on the contrary each bid must refer to a particular unit), if is there a limited number of bids for each portfolio or unit per time interval, if price caps are implemented, if negative prices are allowed, etc. However, next we will focus on discussing the three ones previously highlighted as most relevant.[21]

Complex versus simple auction

Since electricity is a very complex commodity, and its production is subject both to inter-temporal constraints and to the existence of a number of non-convex costs, the format of the generators' offers can range from the so-called simple one (a series of quantity-price pairs per time interval) to a gray scale of more complex alternatives, in which inter-temporal constraints and/or multidimensional cost structures can be declared. We build our brief review of the main alternatives around the two extremes (complex and simple auctions), and then we introduce the hybrid alternatives implemented to amend these latter simple designs.

Complex auctions

In a complex auction generation agents submit offers, representing the parameters and costs which define best their generating units' characteristics (fuel cost, start-up cost, ramp up limit, etc.). For example, in the case of PJM, some of the technical parameters a generating unit declares are [28]: Turn Down Ratio,[22] Minimum Down Time, Minimum Run Time, Maximum Daily Starts, Maximum Weekly Starts, Hot Start Notification Time, Warm Start Notification Time, and Cold Start Notification Time.

With all these data, the market operator clears the market using an optimization-based algorithm that maximizes the net social benefit. This optimization

[21] The discussion that follows is based on Rodilla et al. [29].

[22] Turn Down Ratio is defined as the ratio of economic maximum MW to economic minimum MW.

algorithm shares most of the characteristics of the traditional unit commitment (see formulation in Sect. 7.3.1.1), but with the only difference that the data considered are market agents bids instead of costs.

Usually, market prices are obtained as a by-product of the complex optimization-based algorithm. In the next section, the way these prices and the agents remuneration are computed is introduced.

Simple auctions

The downside of the complex-auction approach is the associated complexity of the market clearing process. This factor has been the key argument held by (mainly) generators to move toward a much simpler auction, where the efficiency of the economic dispatch that results from the market clearing is sacrificed in favor of the transparency of the price computation process.

In the so-called simple auction scheme, the format of the offers does not explicitly reflect the generation cost structure (e.g., an offer component for the start-up cost) or imply any inter-temporal constraint. Instead, market agents submit simple offers/bids, which exclusively consist of price-quantity pairs representing the willingness to sell/buy the underlying product (one MWh in a certain time period of the day, e.g., an hour).

In this type of auction no optimization algorithms are needed to clear the market in such a way that net social benefit is maximized. Matching the market and obtaining the volume of electricity that is traded in each time period of the day is straightforward when offers and bids are simple: generation's offers are sorted in order of increasing prices and the demand's bids are sorted in order of descending prices. By finding independently for each time period the interception point between both aggregate curves (demand and generation) the operator directly determines the total volumes sold. Thus, a generator's offer will have to be accepted if and only if the market price in that particular hour equals or falls above the price offered, and analogously, a demand's bid will have to be accepted if and only if the market price equals or falls below the price bid.

This simple clearing procedure ensures that the net social benefit is maximized both in each particular hour and also along the whole day.

Summarizing, if complex conditions are disregarded, the pricing algorithm is as follows.

- Bids and offers for each delivery period are submitted by a specified deadline.
- A merit order is established.
  - Bids are ranked by price in descending order.
  - Offers are ranked by price in ascending order.
- The (equilibrium) market outcome is defined by the equilibrium market price (EP).

- The EP is the bid price for the amount of energy that corresponds to the cumulative amount of energy demanded.
- Bids specifying a price not lower than the EP are accepted.
- Offers specifying a price not higher than the EP are accepted.

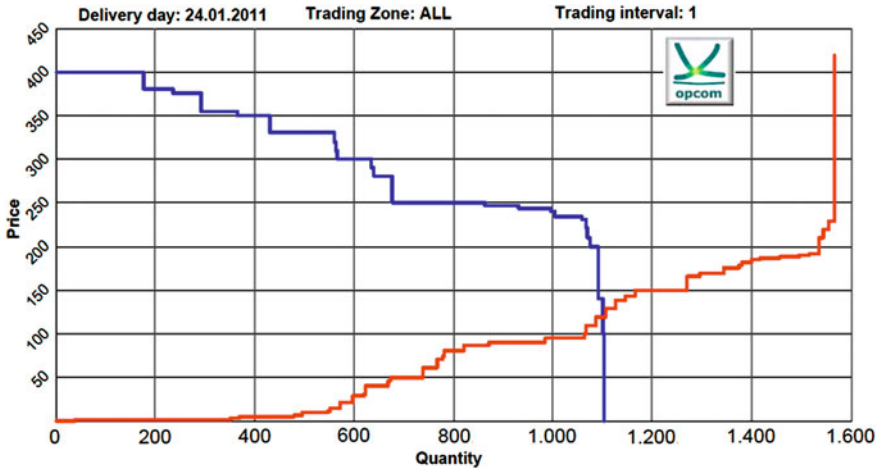


Fig. 7.2 Aggregated curves in Opcom. *Source* www.opcom.ro

Figure 7.2 represents the aggregated curves (demand and supply curves) for a Romanian electricity market auction back in year 2011. As explained above, offers made at a price above the equilibrium price (the system marginal price, SMP) are accepted and are by the generators bidding at prices lower than the EP.

Fully simple offers/bids do not imply any inter-temporal constraint. This means that, for instance, the offers of one thermal generating unit in the day-ahead market could be accepted in the third, fifth, and seventh periods, leading to a resulting unit schedule which could be highly uneconomical or simply infeasible from the technical perspective. As we later further discuss, the main drawback of this approach is that it entails that to some extent generators have to anticipate (based on conjectures) the dispatch so as they properly internalize all cost in the hourly price component.

Due to this need to internalization that has just been introduced, simple offers lead to a significant lack of offering transparency. This fact severely complicates the market monitoring task, since it is difficult for the regulator to determine whether the offers represent actual costs or not, for there are many cost components represented in a simple hourly price. Thus, paradoxically, simple auctions

offer a more transparent and replicable clearing process (the interception of two curves) based on not-so-transparent generating offers.

Hybrid or semi-complex auctions

In principle, the previous inconvenience could be partially fixed either by means of subsequent secondary trading (in the so-called intraday markets, in the EU context, or in the real-time market, e.g. in the US, see below) or closer to real time later in the balancing mechanisms/markets managed in most cases by the System Operator. However, in an attempt to combine the advantages of the complex and the simple auction design, EU PXs have opted for implementing hybrid alternatives, allowing linking semi-complex conditions to their offers.

The common idea behind the design of these semi-complex designs is simply to introduce as few complex constraints as possible in the auction, so as to not to complicate the matching process in excess while at the same time removing the huge risk at which agents are exposed in the simple auction context. Obviously, there is a whole continuum, between the extreme of including all potential constraints and the extreme of including none of them. The larger the number of constraints allowed, the closer the offers can represent the cost functions of the generating units.

In practice, this trade-off has been achieved either by introducing some of the most relevant (most difficult to be internalized) constraints, as it is the case with the ramp-up constraint (used in the Iberian day-ahead market) or by allowing some heuristic-based inter-temporal constraints in the offers format, in most cases not necessarily representing actual constraints or cost components, but rather a mixed effect of many of them. This is the case of the EU PXs semi-complex block-bids. These block-bids, as understood in the EU PXs context (for the term is also used with other meanings in some North American markets), allow agents to submit on one hand a certain interval of consecutive hours where they are willing to produce, and on the other hand the average price they require to be committed in that very period.

Some of the complex conditions and offers used in semi-complex auctions are for example user-defined block-bids (implemented, among others in the Nordpool, EPEX Germany, and EPEX France), meaning that a market agent can offer/bid a price/quantity pair for a set of consecutive hours (three as a minimum), flexible hourly bids (Nordpool and EPEX France), i.e., price/quantity pairs with no pre-defined hourly period assignment or the so-called minimum income condition implemented in OMIE, enabling a generating unit to include a minimum income condition expressed as a fix (expressed in euros) and variable term (in euros per MWh) associated with the whole set of hourly bids corresponding to one particular unit.

Figure 7.3 illustrates the impact of these semi-complex conditions for the case of the Iberian day-ahead market[23]: the resulting market price turns to be higher than the equilibrium price that would have resulted if no conditions would have activated.

[23] It can also be observed how in this market at that time a 180 EUR/MWh price cap was in force.

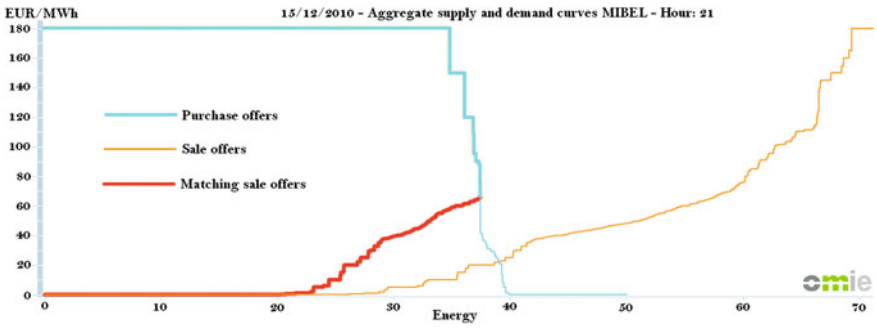


Fig. 7.3 Impact of the semi-complex conditions on the resulting market price

Pricing rules: discriminatory versus non-discriminatory payments

The computation of market prices as well as the related determination of the generating units' remuneration is a quite controversial and still open issue in the context of complex auctions. We can classify these approaches into two large groups:

- nonlinear pricing rules (also known as discriminatory pricing schemes), according to which, on top of the hourly prices, some additional side-payments are provided on a differentiated per unit basis;
- linear (or non-discriminatory) pricing rules, according to which the same hourly price is used to remunerate all the hourly production and individual no side-payments exist.

It is straightforward to observe that total charges to consumers and payments to generators will be different under the two rule systems, and individual generators will get differentiated treatment under the first rule.

Nonlinear pricing

In the context of complex auctions, nonlinear (or discriminatory) pricing is undoubtedly the most extended pricing rule (especially in the US markets). This mechanism translates into each generator having a remuneration consisting of:

- first, a set of (nondiscriminatory, i.e., common[24]) prices which serve to remunerate all production in each time period,
- and then, some additional—individual—side-payments (in practice computed as a lump-sum daily payment) which are calculated on a per unit basis.[25]

[24] The term "discriminatory" should not have here any derogatory interpretation; it just means that the pricing rule results in different charges or payments for the several agents.

[25] A very closed-related technique is the alternative one proposed by O'Neill et al. [26]. This approach entails considering start (and other non-convex) costs as additional commodities different from energy, and thus, needing to be priced independently from this latter.

The set of non-discriminatory prices are computed as the dual variables (shadow prices) associated to the generation-demand balance constraint of the linear optimization problem that results when the commitment decisions have been fixed.[26] As a consequence of the method used to compute the marginal prices, these prices do not include the effect of non-convex costs (e.g., start-up or no-load costs). This is the reason why additional payments are considered on an individual basis in order to ensure (if necessary) that every unit fully recovers its operating costs. Roughly speaking, the price determination could be understood as a two-step process in which:

- First, the unit commitment optimization problem is solved considering all the convex and non-convex costs. The solution of this problem provides, among other results, the binary variables (e.g., start-up decisions) that result from the economic dispatch.
- Second, the same unit commitment problem is resolved, but in this case fixing those binary variables. The dual variables associated to the generation-demand balance constraint are taken as the non-discriminatory price.

In practice, there is not a single way in which this two-step problem can be addressed and solved. The most common approach consists of solving a mixed integer problem (MIP) by means of the branch and bound technique.

Linear pricing

Although the nonlinear pricing approach is the most extended alternative in the context of complex auctions, linear pricing is also a possibility, see e.g. Ref. [3]. Linear pricing in this type of auctions entails computing non-discriminatory hourly prices in such a way that all generating units fully recover their operation costs (thus avoiding the need for discriminatory side-payments of any kind), so in each time period (e.g., hour) every MWh produced is remunerated with the same hourly price.

Finally it is important to remark that we have just focused on the complex auction context. The reason is that the linear versus the nonlinear pricing discussion has been less relevant in the context of simple auction. This is mainly because the question on whether or not the single price should internalize the effect of non-convex costs (such as the start-up cost or the no-load cost) makes no sense in the simple auction scheme. In the simple auction context generators have to internalize all types of costs in their price-quantity pairs offers. Once submitted, there is no way for the market operator to make distinction on which part of the price corresponds to convex and which part of the price corresponds to non-convex costs.

[26] These prices in the vast majority of cases correspond to the variable fuel cost of the most expensive unit committed in each time interval.

Prices and transmission constraints: nodal, zonal, and single pricing

A number of regulatory options are open to deal with the allocation of limited transmission capacity for transactions among players under normal market conditions. One way to differentiate the main categories of options is to gather them in two main groups: those pricing algorithms that involve a detailed representation of the transmission network, and those other which consider a simplified one.

Nodal pricing

Nodal pricing applies security constrained economic dispatch to derive a bus by bus Locational Marginal Prices (LMP), the prices paid for the energy consumed or generated at a given transmission node, as used for instance in PJM, ISO-NE, or ERCOT.

LMPs reflect the value of energy at a specific location at the time that it is delivered. If the lowest-priced electricity can reach all locations, prices are the same across the entire grid, except for the losses effect. When there is transmission congestion (heavy use of the transmission system in an area), energy cannot flow freely to certain locations. In that case, more expensive and advantageously located electricity is ordered to meet that demand. As a result, the LMPs are higher in those locations.

LMPs result from the application of a linear programming process, which minimizes total energy costs for the entire interconnected power system, subject to a set of constraints reflecting physical limitations of the power system. The process yields the three components of LMPs: LMP [$/MW] = Energy component + Loss component + Congestion component. The energy component is the same for all locations. The loss component reflects the marginal cost of system losses specific to each location, while the congestion component represents the individual location's marginal transmission congestion cost.

Nodal energy pricing provides an accurate description of the technical and economic effects of the grid on the cost of electricity. They implicitly include the effect of grid losses and transmission congestion, internalizing both effects in a single value (monetary unit per kWh) that varies at each system node (Fig. 7.4).

Zonal pricing

Zonal pricing consists of using a single market price except where significant grid constraints arise frequently between a limited number of sufficiently well-defined zones of the power system. Once the most frequent points of congestion are identified, the grid nodes affected by internodal congestion are grouped into areas or zones. As defined by O'Neill et al. [25], in this context "a zone is a set of nodes in geographical/electrical proximity whose prices are similar and are positively correlated over time". This pricing mechanism distinguishes energy prices by zone in lieu of by nodes, and the same price prevails at all nodes within a given zone. Figure 7.5 shows an example of the application of zonal pricing in the NORDEL market.

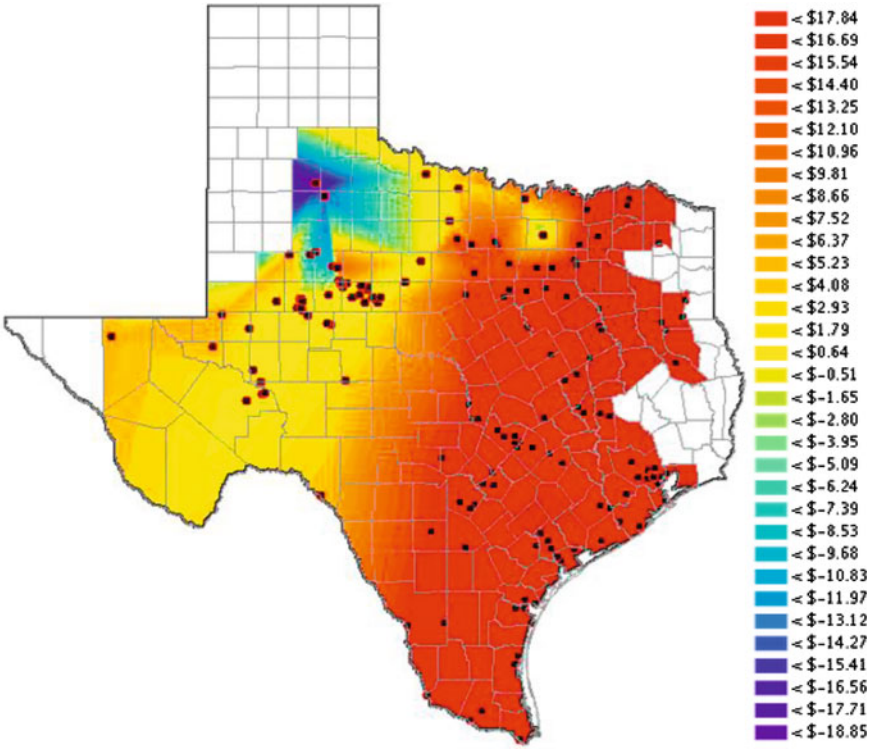


Fig. 7.4 Locational marginal prices (LMPs) in ERCOT. *Source* www.ercot.com

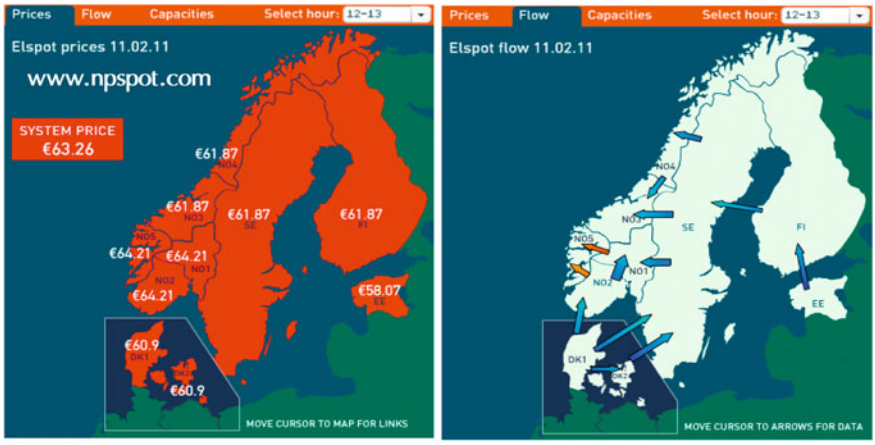


Fig. 7.5 Zonal prices of the Elspot market in the Nordic region in Europe

At the other extreme of nodal pricing we have the so-called single pricing model, whereby any transmission congestions are fully ignored when the electricity market is cleared. This alternative is implemented in those markets where supposedly no systematic or structural congestions occur.

Thus, the market is first cleared in the day-ahead PX considering the simplified representation of the network (e.g., taking into consideration predefined theoretical interconnection capacities between the zones or directly ignoring transmission congestion in the case of single pricing).

In (supposedly) few cases in which grid constraints are detected, the System Operator re-dispatches the system, determining which players must withdraw from the system and which are to be included. Energy removed to solve the network constraint may be paid at the respective agent's bid price (if a specific bid related to the constraint solving mechanism is in place), at the opportunity price (energy market price less the price of the agent's bid), or not at all. When additional energy is requested, it is normally paid at the respective agent's bid price.

7.3.4 Adjustment Markets and Tools

Once DAM closes, additional shorter termed tools have to be implemented to enable participants or the SO in their behalf to improve the day-ahead market results where sub-optimal, and afford the opportunity to assimilate new information (e.g., an unplanned outage). As previously introduced, two main models can be distinguished: the most common one in the EU context, built around intraday markets organized by the Power Exchange followed by balancing markets run by the System Operator, and the other one implemented in the US markets, centralized in the so-called Real-Time Markets.

7.3.4.1 Real-Time Markets

In the US model, real-time energy markets balance deviations between the day-ahead scheduled quantities of electricity required and the actual real-time load needs. These markets are spot markets in which current Locational Marginal Prices are calculated at five-minute intervals based on actual grid operating conditions. Security Constrained Economic Dispatch (SCED) is the real-time market evaluation of offers to produce a least-cost dispatch of online resources.

As a way of example, let us take the brief description of the real-time market operated by PJM [28]:

The real-time energy market is based on actual real-time operations. Generators that are PJM capacity resources and Demand Resources that are available but not selected in the day-ahead scheduling may alter their bids for use in the real-time energy market during the Generation Rebidding Period from 4:00 PM to 6:00 PM (otherwise the original bids remain in effect for the balancing market). Real-time LMPs are calculated based on actual system operating conditions as described by the PJM state estimator. Load Serving Entities (LSEs) will pay the Real-time LMPs for any demand that exceeds their day-ahead scheduled quantities (and will receive revenue for demand deviations below their scheduled quantities). In the energy market, generators are paid the Real-time LMPs for any generation that exceeds their day-ahead scheduled quantities (and will pay for generation deviations below their scheduled quantities).

7.3.4.2 Intraday Plus Balancing Markets

Intraday markets

Power exchanges generally hinge on a day-ahead market, where electricity is traded the day before the day of delivery. Most also provide adjustment markets (so-called intraday markets) where additional trading can take place when supply or demand situations change with respect to the estimates cleared on the day-ahead market. Market participants can modify the schedules defined in the DAM by submitting additional supply offers or demand bids.

As a way of example, a couple of these intraday markets are outlined below.

- Nord Pool: Elbas is an intraday market for trading power operated by Nord Pool Spot. Elbas supplements Elspot and covers the Nordic region, Germany and Estonia. At 14:00 CET, capacities available for Elbas trading are published. Elbas is a continuous market, and trading takes place every day around the clock until one hour before delivery. Prices are set based on a first-come, first-served principle, where best prices come first—highest buy price and lowest sell price. The products traded are one hour power contracts. Nord Pool Spot AS acts as counterparty in all contracts traded on the Elbas market and all trades are physically settled with the respective transmission system operator (TSO).
- Italy: The *Mercato infragiornaliero* (MI) takes place in four sessions: MI1, MI2, MI3, and MI4.[27] The sessions of the MI1 and MI2 take place after the closing of the DAM. They open at 10:45 a.m. of the day before the day of delivery and close respectively at 12:30 p.m. and 2:40 p.m. of the same day. The sessions of the MI3 and MI4 open at 4:00 p.m. of the day before and close at 7:30 a.m. and 11:45 a.m. of the day of delivery. The results in the four cases are made known half an hour later the closure.

Balancing markets

In real-time power system operation, the SO or TSO must ensure that supply matches demand at all times. Consequently, competitive electricity markets

[27] http://www.mercatoelettrico.org

generally feature a balancing mechanism that enables SOs to take measures geared to maintaining the supply/demand balance for which they are responsible. Imbalance pricing arrangements can be used to encourage market players to maximize their efforts to match supply and demand. Balancing markets therefore form an integral part of overall wholesale electricity trading arrangements and timetables.[28]

Notification and bidding deadlines differ depending on the country. Gate closure is a concept common to almost all SOs, although the details may differ, primarily in two respects [12].

- Gate closure may be a rolling deadline programmed at fifteen-minute (Netherlands), half hour (England and Wales), or hourly (Sweden) intervals, or a deadline set at specific times during the day (France, Germany, Spain).
- The time lapsing between gate closure for a given period and its start time also varies from one country to the next. In Sweden it is 1 min (i.e., the gate for the 10:00–11:00 period closes at 09:59). In Netherlands, Norway and England and Wales gate closure is one hour before the period begins.

After gate closure, market players may not vary in the expected physical position of their generators. In real time, the SO may change the physical positions of generators (or demand) to balance the system or manage congestion. If an interconnector links two systems that use different gate closure times, the earlier gate closure is generally applied.

Furthermore, actors may submit bids and offers on the balancing market specifying the extent to which they are willing to be paid to deviate from these positions and what they will charge for this service. These markets just remunerate the energy delivered but do not provide for any payment for availability.

In many countries a balance responsible party (BRP) is designated for this purpose. BRPs are market players that are financially responsible for balancing injections and withdrawals (including possible purchases and sales) of electric power. For instance, if one of the generating plants for which a given BRP is responsible fails to deliver electricity as scheduled in the generation plan, the imbalance cost is charged by the SO to that BRP. In a way, each balancing responsible party is like a "virtual network" that must keep its schedules balanced.

For the SO to exercise full control over system stability, all wholesale market players connected to the transmission system are usually either compelled to be a balance responsible party or to trade through an aggregator with balance responsible party status. Market participants are then either directly or indirectly bound by balancing market rules.

Following gate closure, the SO calls for generation bids and load offers to balance the system at the lowest cost. Where intraday markets exist, SOs need to take further bid and offer restatements into consideration when extending these requests.

[28] Part of this discussion is taken from ERGEG [11].

The purpose of balancing markets is to provide short-term operational security of supply (security of grid operation) from a market approach and settle energy imbalances. Hence, balancing markets must be economically efficient. SOs purchase balancing power using market criteria.

Balancing markets are generally designed to encourage market participants to manage their exposure to imbalance. For example, generators who run short of their notified position are usually required to purchase the difference between energy notified and delivered at a price established by balancing mechanism rules. This price is likely to reflect the costs to the SO of acquiring the energy shortfall and may be higher than the price on the intraday or day-ahead energy markets. Such arrangements are designed to minimize the amount of balancing power needed and lower overall balancing costs. Similarly, SOs themselves may be the object of incentive mechanisms designed by the regulator to encourage them to conduct balancing in a least-cost manner.

Market power may be a real risk in balancing markets, even where the market is relatively "non concentrated". Relatively small players may be able to impact the market heavily when the supply/demand margin is small, or where they cover a specific geographical position or have unique technical characteristics, particularly where demand parties or other generators are already committed or unable to respond to price signals on very short notice. An efficient balancing market should be as resistant as possible to any such exercise of market power, given market structure, and concentration. Transparent market operations enable all players as well as regulators to identify and discourage any unfair behavior.

Imbalances may occur only during operating hours and they are balanced using the balancing power provided by balance responsible organizations such as SOs. The costs of handling imbalances may be distributed across all users, allocated to the market participants involved in the imbalance, or a combination of the two. Parties incurring imbalance are, in any event, subject to some manner of 'imbalance charge'.

Imbalance arrangements and pricing must be simple and transparent so that the underlying principles are easily understood and justified in ways that enable market participants to readily assess economic risk. Imbalance arrangements must enhance balancing market and wholesale market efficiency. Balance settlement arrangements must provide for swift and accurate settlement and invoicing.

In principle, balancing markets must:

- deliver short-term operational security
- operate to economically efficient standards
- use market methods
- further effective competition
- not contribute to market power
- be non-discriminatory
- have clearly defined roles and responsibilities
- ensure transparency.

Imbalance pricing

In most European power systems, imbalances are presently settled via dual imbalance pricing, in which positive and negative imbalance volumes are priced differently, depending on the hour. A fictitious but representative pricing scheme is used in the following discussion. The price of imbalances in each half-hour period is assumed in this example to be determined by two factors:

- the absolute value of the imbalance for each balance responsible agent,
- the sign on the imbalance (positive or negative) for the system as a whole (the opposite sign is often referred to as the balancing trend, which is upward when the volume of energy involving upward balancing operations is greater than the volume of energy involving downward balancing operations, and vice versa).

In a given half-hour period, the settlement price for balancing agents whose imbalance has the same sign, positive or negative, as the system as a whole is computed as the average weighted cost of the balancing actions adopted by the System Operator to correct the imbalance. This value is adjusted by a factor of $1 + k$, which raises the price for negative (and reduces it for positive) imbalances, further penalizing imbalanced parties. The day-ahead price is applied to imbalances, positive or negative, which bear the sign opposite the sign for total system deviation.

This dual imbalance pricing, reinforced by the use of the factor k, is designed as an incentive for market agents to try to avoid deviating from their scheduled programs. A measure intended to improve system security, it nonetheless has certain adverse effects for market operation and overall efficiency.

The ultimate purpose of this measure is to safeguard system security, on the assumption that single pricing methodology reduces the incentives for agents to avoid deviating from their day-ahead programs. In fact, imbalances that bear the sign opposite the sign of overall system imbalance reduce net system deviation and contribute to mitigating the problem. In theory, then, the agents involved should be entitled to receive the balancing price, like any of the producers presenting bids to the system operator to solve the imbalance problem. Such uninstructed deviations should not be encouraged, however. The fear is that, on the one hand, rewarding these unintended imbalances that incidentally solve the system problem may create an incentive for some agents to deviate intentionally from their schedules to capture this additional income; and on the other that the single-price mechanism may weaken the incentives to maintain a balanced schedule.

This reasoning is sound, in principle. The dual price methodology induces costs, however, which may be significant in some cases. The dual pricing approach allocates security costs asymmetrically, for instance. Since imbalances are measured for BRPs only, netting is allowed for deviations

within the same balancing party, but not for deviations between balancing parties. Take, for example, one generator that produces 10 MW more and another one 7 MW less than expected. If they are owned by the same firm (and thus under the same balancing umbrella), only a 3 MW imbalance is recorded. But if the generators are owned by two small companies, two imbalances are computed, one for 10 MW and the other for 7 MW. When the overall balancing results are calculated for the two generators, the larger firm is better off by an amount of 14 MW times the difference between the two balancing prices. The larger the spread between these two prices, the larger is the penalty for the small firms.

More dramatically, assuming no changes in the unit operating efficiency in a firm owning a portfolio of generators, if, for imbalance settlements purposes, the firm is artificially divided into two, the resulting payments for imbalances would be larger. Therefore, if no technical improvement is made, the impact on the firm is greater.

This effect is highlighted by the fact that intra-firm netting might be a result not only of accidental imbalances with opposite signs attributable to two generators belonging to the same firm, but also of each firm's ability to self-adjust its schedules, i.e., to compensate for a potential problem in one of its plants with its own generators.

New entrants or, more generally, generators owning small portfolios, are far more sensitive to imbalance prices, and therefore more intensely affected by the asymmetric nature of the dual pricing methodology. In other words, this regulatory measure allocates a higher portion of the implicit costs of security to smaller agents, creating an entry barrier for potential new actors and consequently hindering power market development. In most European countries, where a small number of agents control an extremely high percentage of system generation resources, this should be viewed as a matter of major concern.

Due to the peculiar strategic characteristics of electric power supply, it would not be difficult to defend the premise that no market efficiency objective is justified if it implies a cutback in security standards. At the same time, however, the impact and effectiveness of this kind of measures depend significantly on the particular characteristics of the power system where they are implemented. In a fully developed, mature and quasi-perfect market, the asymmetric impact on agents might be less troubling, since in principle the agents would be more size-comparable and the costs of market intervention less significant. But the opposite is true when agents' market shares differ radically.

7.4 Generation Ancillary Services

7.4.1 Introduction: Ancillary Services Products

The foregoing sections reviewed the sequence of market-like transactions that take place prior to gate closure, i.e., when the final production schedule is determined for all participants after bilateral, spot or day-ahead market trading and possible schedule adjustment on the intraday market. From this time on, responsibility for generation scheduling and dispatching is transferred to the System Operator, which makes all the pertinent decisions and defines and manages a series of procedures to ensure the delivery of electric power to suitable quality and security standards.

Generation and network ancillary services are the services associated with the production, transmission, and distribution of electric power necessary to guarantee the quality, security, and efficiency of supply. Quality of supply is understood to mean maintaining voltage and frequency within acceptable margins for the system, security of supply to mean non-interruption of supply in the short term (not to be confounded with long-term reliability of supply), and efficiency to mean supplying electric power at the lowest possible cost.

Ancillary services are very closely associated with power production: traditionally, the provision, purchase, and remuneration of such services were fully integrated into basic power generation, from which they were considered to be inseparable. In the context of liberalization, a need is gradually being identified to establish separate provision and remuneration mechanisms for these services to minimize anticipated operating costs. While it is incumbent upon the regulator to develop the regulatory models for these services, this institution often delegates the responsibility for their implementation to the System Operator, as an independent agent with an in-depth understanding of system operation.

Following [20], operating reserves are defined as the real power capability that can be given or taken in the operating timeframe to assist in generation and load balance, and frequency control. There is also need for reactive power reserve, but it will not be discussed here. The types of operating reserves can be differentiated by: (a) the type of event they respond to, such as contingencies, like the sudden loss of a generator or a line, or longer timescale events such as net load ramps and forecast errors that develop over a longer time span; (b) the timescale of the response; (c) the type of required response, such as readiness to start quickly a plant or fast response to instantaneous frequency deviations; (d) the direction (upward or downward) of the response.

In spite of their relatively small financial significance (from 1 to 10 % of total generating costs), ancillary services, listed below, are absolutely necessary for system operation. Terminology and subdivisions into different services may differ

from one country to another. A European-oriented way to classify these services is[29]:

- Frequency control,[30] which consists of the following three elements:
  - Primary reserve: automatic maintenance of the balance between generation and demand, using the rapid response governors built into generators to handle frequency deviations.
  - Secondary reserve: centralized and automatic function whose objective is to regulate generation output in a control area to:

exchange energy with other control areas at the programmed levels, and
return the frequency to its set value in case of a (major) deviation, thus restoring primary control reserve.

  - Tertiary reserve[31]: automatic or manual change of the generator operating point (mainly by re-scheduling) to restore an adequate level of secondary control reserves.

- Reactive power for voltage regulation is essential to establish and sustain the electric and magnetic fields of alternating-current facilities and has a direct effect on system voltage.
- Black-start capability (power restoration) is the ability of a generating unit to start up and deliver power without external assistance from the power system.

Most of the ancillary services discussed in this chapter fall under one of these three headings. But as stated, it is also possible to come up with different classifications. A more US-oriented approach can be found in Milligan et al. [20], where all types of reserves are classified into five categories, in decreasing order of quickness of reaction:

(i) *frequency response reserve* (to provide initial frequency response to major disturbances; also called primary control or governor response, acting in seconds);
(ii) *regulating reserve* (to maintain area control error within limits in response to random movements in a timeframe faster than energy markets can clear; also termed frequency control or secondary reserve, acting in seconds);

[29] See UCTE [38] for a collection of operating principles and rules for transmission system operators in continental Europe.

[30] As explained in Chap. 1, electric power systems have regulation mechanisms to keep frequency within an acceptable range around the nominal value. The aim of such controls is to maintain the equilibrium between the mechanical power delivered to the generators and the electric power demanded by the system.

[31] Tertiary reserve is sometimes classified under a different heading, "balancing energy". Its inclusion in the upper hierarchical level of the process that controls system frequency is believed to be preferable, for this also avoids confusion with the balancing market explained below.

(iii) *ramping reserve* (to respond to failures and events that occur over long timeframes, such as wind forecast errors or ramps; also termed deviation reserve, balancing reserve or forecast error reserve, acting in minutes to hours);
(iv) *load following reserve* (to maintain within limits area control error and frequency due to non-random movements on a slower time scale than regulating reserves; also named tertiary reserve, acting in several minutes); and
(v) *supplemental reserve* (to replace faster reserve to restore pre-event level reserve; also called tertiary reserve and replacement reserve, acting from minutes to hours).

Regulating and load following reserves are used during normal system operation. Frequency response and supplemental reserves are used during contingencies. A mix of spinning and non-spinning reserves can be used for the slower reserves (ramping, load following, and supplemental) while the faster reserves (frequency and regulating reserves) require strictly spinning reserves.

It is incumbent upon the System Operator to establish the volume of service to be provided. The operator asks generators to provide the system with a certain reserve capacity (active power and reactive power reserve capacity or power with autonomous start-up capacity). Since the proportion of this capacity that will be used cannot be known in advance, the volume of service required must be established on the grounds of probabilistic criteria such as used in centralized planning decisions to install generating capacity.

7.4.2 Frequency Control

Frequency control is conducted at three levels, distinguished in terms of response time, as explained below.

Primary regulation
This is the automatic, local regulation provided by generating unit speed regulators. These regulators control frequency in the unit terminals by adjusting the turbine mass flow control valve: when frequency drops to below its established value the regulators widen the valve opening and narrow it when frequency rises. The variable that relates frequency variations to increases in power is known as speed regulation or droop and is a characteristic parameter of generating unit regulators. This level of regulation sustains frequency levels, preventing large deviations from the scheduled value. The response time for this type of regulation is measured in seconds. This type of regulation is sometimes expanded to include the natural response of the system as a result of the inertia of the generator rotors, i.e., the kinetic energy released or absorbed when network frequency varies. It also covers the automatic load shedding control performed by minimum frequency relays installed by distribution companies on certain loads under SO supervision.

Secondary regulation
This is the automatic, regional regulation provided by automatic generation control (AGC), which sends signals from the control center to certain generators to re-establish the nominal frequency value and restore the primary reserve capacity (and power exchanges with adjacent control areas to their original values). The response time for this type of regulation ranges from 5 to 15 min.

Tertiary regulation
This manual, regional regulation, provided by generating units and controlled by the System Operator, attempts to restore the secondary reserve capacity by dispatching the units that entail the minimum incremental operating cost for the system. The response time for this type of regulation is upward of 15 min.

Ranking regulation by response time optimizes the use of the available resources. Figure 7.6 summarizes frequency and power regulation in the event of an abrupt generation outage.

Procurement and payment of frequency control services
Europe's Transmission System Operators (TSOs) can purchase ancillary services in several different ways, depending on the type of service [12]. Ancillary services are acquired by the SO or TSO both before and after gate closure in short-term markets. Prior to gate closure, the SO may conclude long-, medium-, and short-term agreements, whereby generators commit to availability for providing primary, secondary, or tertiary reserves. These agreements may be awarded in the context of a monthly or annual market, or negotiated bilaterally. They typically specify the technical characteristics of the service, the availability level required (a certain number of hours, or between specified time periods), and a price.

Primary control services may be obtained either commercially or imposed as an obligation. In Germany, Poland, Sweden, and Denmark, all primary frequency control services are purchased commercially and generators are under no obligation to provide the service. In other countries, all the major generators are obliged to provide primary frequency control services, although payment arrangements vary. In France and the UK mandatory frequency response is remunerated on a

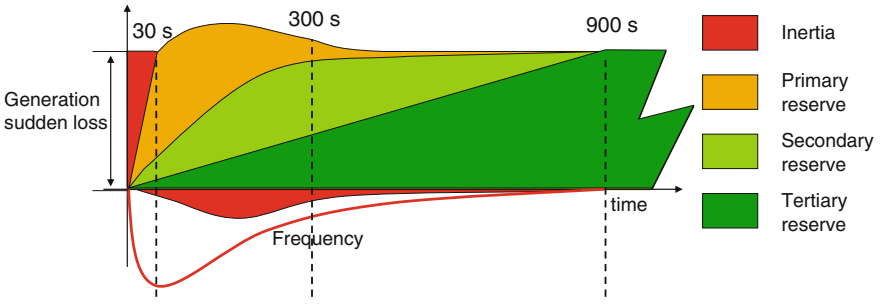


Fig. 7.6 Frequency and power after a sudden loss in generation

cost reflective basis. In Norway, mandatory primary frequency response is partly remunerated under an agreement negotiated yearly and partly in accordance with a market-based arrangement in shortage periods. The service carries no explicit charge in Austria, Spain, Italy, Norway, Netherlands, Switzerland, or Slovenia.

Secondary control is generally a commercial service with no obligations attached. The exception is in France, where secondary control is required of generators exceeding a certain size and is paid on a cost reflective basis. In Germany, frequency control services are purchased under semi-annual tendering arrangements. Pre-qualified generators submit one price for primary control, and two prices for secondary control—one for reserve capacity, and the other for the energy delivered when the generator is called upon to provide the service.

As described in NYISO [24], the NYISO selects Regulation Service in the Day-Ahead Market from qualified Resources that bid to provide Regulation Service. Market Participants may submit bids to the NYISO for Regulation Service up to the Real-Time Market market-close time (75-minutes prior to the operation hour). Bid information includes:

- Regulation response rate, in MW/min, with the exception that Limited Energy Storage Resources (LESRs), such as flywheels and batteries, are not required to provide a regulation response rate.
- Regulation availability/price, in $/MW.
- Regulation Availability MW—regulation capacity available in one direction. For example, a bid of 5 MWs is a bid to provide 5 MWs of regulation up and 5 MWs of regulation down.

As it has been described in the previous section, some manner of ad hoc market is frequently used to purchase tertiary frequency control services: the real-time and balancing markets.

In many systems, to complement these markets, the SO purchases additional reserves, in line with the ones enumerated in the previous section.

For instance, ISO New England acquires three different types of these operating reserves:

- 10-minute non-spinning reserve (TMNSR)—A form of operating reserve provided by off-line generation that can be electrically synchronized to the system and increase output within 10 min in response to a contingency.
- 10-minute spinning reserve (TMSR)—A form of operating reserve provided by on-line generation that can increase output within 10 min in response to a contingency.
- 30-minute operating reserve (TMOR)—A form of operating reserve provided by on-line or off-line operating reserve generation that can either increase output within 30 min.

Additionally, ISO-NE runs a Locational Forward Reserve Market (FRM). FRM provides revenue to peaking resources that operate infrequently. Participants commit in the FRM through auctions held twice a year. The auctions set the prices

and procure the reserve capacity to meet system-wide TMNSR and TMOR requirements and local TMOR requirements

7.4.3 Other Ancillary Services

Reactive power and voltage regulation

Another element of the electricity system, known as voltage–reactive power (V/Q) regulation, maintains the voltage in the transmission grid nodes within an acceptable range. This type of regulation is likewise phased to optimize the existing resources.

Primary voltage regulation is the automatic, local regulation conducted by an automatic voltage regulator (AVR). This regulator controls the voltage in the generating unit terminals, adjusting the excitation current to restore nominal voltage values by varying the amount of reactive power generated. This type of regulation is virtually instantaneous.

Secondary voltage regulation is the automatic, regional regulation involving the calculation of a reference voltage value which is sent to the generating units by the AVR every few seconds. The reference value is calculated on the basis of criteria that ensure system operating security. Unlike secondary frequency and power regulation, this mechanism is not installed in most electric power systems.

Tertiary regulation is centralis\zed, manual regulation, in which the System Operator sends reference voltages to generators at intervals measured in hours. These reference voltages are calculated as a compromise between technical and economic constraints.

Both of the above types of regulation, frequency/power, and voltage/reactive power, range from quick and less "intelligent" control (primary regulation) to slower control actions that require more computation time (tertiary regulation) but which yield solutions that ensure nearly optimum system operating conditions.

Black-start capability

The other traditional ancillary service is the restoration of the supply of electric power after outages. This is the system's capacity to return to full operation after a massive failure or blackout involving generation resources. Certain generators have independent start-up capacity, for which purpose the transmission grid is equipped with synchronizers.

7.5 Market Power in Power Markets

Once a comprehensive view of wholesale electricity markets has been provided, this section will extend the general introduction to market power in Sect. 2.5, by focusing on more specific features of power systems. The reader is advised to consult Sect. 2.5 before reading the present section.

7.5.1 Market Power and Market Structures

Market power fundamentally depends on the structure of the market rather than the rules, under sound market designs. Regulators should understand that when the market structure is not adequate for competition, the solution is not a change in the market rules, but a change in structure.[32]

As it was mentioned in Chap. 2, prior to the restructuring that led to the implementation of wholesale markets, most electric power systems were organized around a rather small number (often just one) of vertically integrated monopolies. These traditional electricity utilities were frequently fully government-owned (this was the case for instance in the majority of cases in Europe, being Spain one of the exceptions).

Turning this existing structure into another one that would be suitable for the implementation of a sufficiently competitive market happened to be tall order. When the utility was not fully controlled by the government but by private investors it was legally difficult to mandate the company to divest part of its assets to allow for the creation of a sufficiently significant number of relevant competitors. In the majority of cases utilities could not be legally required to divest, except when divestment is established as a condition for authorizing mergers and acquisitions (M&A). Therefore the most common outcome was the negotiation of complex agreements, frequently involving the acceptance of stranded generation costs by the regulators.

In those cases in which the government fully controlled the traditional utility, in principle it would have been possible to get to a sufficiently competitive market structure: in theory, it would have been enough to break the traditional company in a sufficiently large number of pieces (generation portfolios), in order to later sell them to different shareholders to guarantee an atomized and thus very competitive market structure, but experience shows that this was not always the case. In some power systems, the governmental decision was either to keep the full control of the company with no significant structural changes[33] while in others the decision was

[32] As we shall see later, interventionist changes in the market rules—i.e., new rules that limit the freedom of the market agents, like a price cap—do mitigate market power. But this happens in detriment of the freedom of the market that it was intended to promote, in the first place.

[33] Take for instance the case of France, in which EDF remains under the Government control and whose market share in the French wholesale market can be qualified as huge, above 85 % according to Eurostat [14], and the case of Vattenfall, still state-owned and whose market share is larger than 40 % in Sweden according to the same source, Needless to say that these market shares cannot be only judged in the context of the national market. The Nordpool in which Vattenfall is embedded turns to be a rather competitive and liquid regional market while the interconnection capacity of the French system with its neighboring markets is still far from being enough to deter the local market power of EDF.

to turn the utility into a publicly owned company within the new competitive market, divesting in some cases part of the generation assets (e.g., Enel in Italy) or a large portion of the shares in the market, but still maintaining a large size of the company (with the purpose of keeping a high market value of the company with a view to a possible future privatization, or to maintain the position of a "national champion" with a strong international presence).[34] One of the pioneer experiences was in the UK, where the former state-owned company's non-nuclear generation capacity was split into two companies, National Power and PowerGen, which were subsequently required to sell some of their assets (see Newbery and Pollit [22] and Thomas [37] for a description of the process and further examples in the UK).

It can be thus stated that in the majority of cases electricity markets are characterized by a concentrated structure, see for instance Eurostat [14] and US Environmental Protection Agency [39].[35] As a result, these oligopolistic schemes are prone to abuse of market power. Thus, looking for the means to identify the existence and detect the abuse of market power, and then to design prevention measures, has always been one of the main concerns of electricity regulators. Next, we introduce the main regulatory measures for the mitigation and control of the exercise of market power in electricity markets.

When markets are interconnected, it is not obvious to define the geographical scope of the market to which monitoring and mitigation measures should be applied. This leads to the concept and definition of "relevant market". According to the ruling of the European Commission, "A relevant product market comprises all those products and/or services which are regarded as interchangeable or substitutable by the consumer, by reason of the product' characteristics, their prices and their intended use. The relevant geographic market comprises the area in which the undertakings concerned are involved in the supply and demand of products and services, in which the conditions of competition are sufficiently homogeneous and which can be distinguished from neighboring areas because the conditions of competition are appreciably different in those areas".

The US Department of Justice & Federal Trade Commission in the "US Horizontal Merger Guidelines (1997) proposes the Small but Significant and Non-transitory increase in price test (SSNIP) to identify the scope of a relevant market: "A market is defined as a product or group of products and a geographic area in which it is produced or sold such that a hypothetical profit-maximizing firm, not

[34] The market share of Enel in Italy, still under the control of the Government (who owns more than 30 % of the shares) has evolved from 70 % in 1999 (the year in which it sold 15 GW) to below 30% [14], In Belgium, the market share of Electrabel, owned by GDF Suez, is above 75 %, while in Chile the market share of Enersis, bought by Endesa now part of the Enel group, as for 2010 was around 35 %.

[35] An excellent review of market power issues in the US is provided by Helman [16]. An easy to read, tutorial text is authored by Rose [30]; another useful reference is The Brattle Group [35]. FERC Order 697-A establishes the conditions to allow market-based rates, depending on market power mitigation issues. Although very technical in legal terms, this document illustrates the terminology and the issues involved when examining market power in actual systems.

subject to price regulation, that was the only present and future seller of products in that area likely would impose at least a "small but significant and non-transitory" increase in price, assuming the terms of sale of all other products are held constant. A relevant market is a group of products and a geographical area that is no bigger than necessary to satisfy this test". In the US a "small" price increase is normally defined as 5 %, while in the EU is 5–10 %.

7.5.2 Monitoring the Existence and Exercise of Market Power

Market power may be exercised in short-term electricity markets in two equivalent ways. In the first one, the price that is bidden for a given amount of energy is higher than what it would have been under competitive conditions. In the second one, the amount of energy offered at a price is lower than what had been competitively offered. The effect of market power in this second approach is usually the withdrawal of supply from the market.

Two long-term effects of the higher prices that result from the exercise of market power can be identified.

Generation firms have a natural incentive to install new capacity. Any additional capacity will ultimately place downward pressure on prices, therefore reducing the incentive for the exercise of market power. The easier it is for potential agents to install new capacity, i.e., the lower the entry barriers, the greater this effect is. In the absence of any entry barriers (i.e., when the cost of market entry is nil), the "virtual competition" from possible newcomers may ultimately suffice to ensure that prices are close enough to the competitive level. An oligopolist would not charge overly high prices for fear of attracting new competitors. Although incumbents on electric power markets often maintain that this is the case and that potential competition from possible new entrants guarantees competitive prices, significant entry barriers usually exist that are difficult to avoid. A long construction time for new power plants is an obvious one.

The exercise of market power eventually leads to lower electricity consumption than on a perfectly competitive market. The greater the elasticity of long-term demand, the lower the consumption.

Market power metrics

Certain special characteristics distinguish electric power markets from other traditional markets. First, the underlying asset is a practically irreplaceable good that cannot be economically stored. Generation side elasticity is very low and demand-side elasticity is presently nearly nil in the short term in most markets. Production capacity is exposed to high operating risks and investments are very capital-intensive. The transmission grid introduces complex market constraints. Market power is difficult to analyze because of all these reasons.

Concentration measurements in the context of power markets have been thoroughly studied. Newbery et al. [23] provides a good review. Another excellent reference is Hope [17]. Traditionally, simple general-purpose indicators have also been applied to power markets. The reader can find a comprehensive description in Sect. 2.5.4 of this book.

Market power monitoring

Proper market monitoring is therefore a necessity in the context of electricity markets,[36] although it faces an almost insurmountable obstacle: the calculation of the (supposedly) correct cost (i.e., competitive bid) of a generation unit is a difficult, fuzzy, subjective and thus, in most cases, pretty arbitrary task.

As stated above, the electricity supply cost/bid/price formation functions depend on such a variety of intricate drivers, that it is almost impossible for the regulator to properly supervise this process. In electricity markets the monitoring duty of the regulator turns to be a permanent pursuit of smoking guns. As stated by Barker et al. [6]: "When structure is not conducive to competition, the regulator and pool operator will find themselves unsuccessfully chasing after conduct. The solution is not a better rule, but a change in structure". If feasible, it would always be more efficient to implement an explicit *ex ante* regulatory measure to tackle the structural problem when detected. The objective of these solutions should be to mitigate the market power of dominant players while at the same time affecting the least possible the right marginal signals that lead generators behavior and system operation to the maximization of economic efficiency.

7.5.3 Measures to Tackle Market Concentration

It is remarkable, the difference in the high level approaches to market power mitigation that have been adopted in the US and in the European Union. The US FERC has a statutory obligation under the Federal Power Act of 1935 to ensure that individual State Regulatory Commissions manage liberalization to ensure that wholesale prices remain "just and reasonable". Therefore, before an electric utility could be allowed to sell at wholesale market prices, any market power has to be adequately mitigated, and the authorization can be withdrawn (& regulated prices would be used instead) if "there is any change in status that would reflect a departure from the characteristics the Commission has relied upon in approving market-based pricing".

On the other hand, articles 81 and 82 of the EU Treaty examine mergers and acquisitions, as well as anticompetitive behavior, but they do not limit market power *ex ante*. The EU Electricity Directives aim to remove the barriers to create a competitive market, without (apparently) seeing the need to ensure that the

[36] See The Brattle Group [35] and Adib and Hurlbut [1] for two good descriptions of the role and function of market monitoring units, mostly from the US perspective.

resulting market structures were sufficiently competitive before introducing liberalization. The Electricity Directive 2003/54/CE establishes that "Member States will create adequate instruments to prevent abuses of dominant position".

The regulatory measures for mitigation of market power can be categorized into two main sets: short-term (price or bid) caps and long-term structural interventions.

Short-term caps

A straightforward and very common way (particularly at the start of electricity markets back in the 1990s) to try to tackle market power has been the imposition of limitations in the generators' bids (the regulator may, for instance, through the Market Operator, amend or even eliminate certain offers considered to reveal anti-competitive strategic behavior) or (more often) in the resulting market prices.

Price caps can be "hard" (a maximum price that demand can pay) or "soft" (when the price limit is defined as a function of some index, as for instance fuel prices in international markets).

Roughly speaking, these caps are far from being a good idea. The main and first reason is that, since it is not easy to differentiate between high prices due to scarcity and due to market power exercise, such a rule directly affects the key market marginal signal for generation and demand. This leads to an inefficient economic dispatch in the short term and a significant distortion of the long-term signal that should attract investment to achieve a well-adapted generation mix.

In addition to affecting market efficiency, limiting the maximum (hourly) price according to which generating units can be remunerated curbs the potential exercise of market power in an extremely partial way. Price caps are set at reasonably high levels to avoid market power abuse under a scarcity event. But experience suggests that market power in electricity markets (if present) is exercised at all price levels (not necessarily in the peak hours) in a moderate way, in order to avoid the regulator intervention.

Long-term structural interventions

The most efficient way to face market power consists of directly intervening on the market structure in the long term. This intervention can be irreversible (e.g., mandating the incumbents to divest) or transitory. Roughly speaking, this alternative entails forcing dominant players to sell part of their generation capacity, or part of their output for a sufficiently long period of time, under the expectation that, after that time, a sufficient number of new entrants will have reduced the concentration levels of the market.

Divestitures

As mentioned earlier, divestment, a traditional and one of the most drastic approaches to excessive horizontal concentration, has often been adopted. This is obviously the most traumatic procedure for companies. Although this method introduces new agents into the market, it also dramatically reduces the presence of the incumbent companies, which are traditionally more deeply committed to the system and, in principle, guarantee investment continuity.

Certain prior considerations must be addressed when implementing this measure. One of the keys to successful divestment is to ensure that timing is sufficiently flexible. A stable regulatory framework and an established market are likewise essential to attracting possible bidders.

A variation on this theme is to prohibit dominant operators from increasing their market share, a measure applied for instance in the case of Spain, in which Endesa, the former publicly-owned utility, was not allowed to grow in the market for a number of years during the early 2000s.

Long-term contracting and market power

The existence or the obligation to sign long-term agreements that fix the revenues earned by dominant firms for part of their capacity, also reduces the impact of excessive market concentration, see for instance Wolak [40]. Thus, in principle, long-term contracts are a helpful tool for mitigating market power. If a utility (with a large wholesale market share) concludes an agreement to sell part of its output in the long term, its incentive to raise prices by withholding part of its capacity declines. Since raising the market price would only benefit the firm for the amount not sold under the agreement (because it obviously receives the fixed price set in the agreement for the part already sold), the larger the amount sold under a long-term agreement, the lower is the utility's incentive to increase the price.

Suppose that a generator owning a large portfolio of plants and typically producing a quantity q in the market, enters into a long-term contract for a portion q_1 of its market share. The immediate impact of this contract is the reduction of the generator's incentive to raise market prices. Indeed, if the generator attempts to manipulate the market price by withholding (or by bidding at a high price) a portion q_2 of its energy output, the contracted energy q_1 will not be affected by the price increase. The strategy of the generator should therefore consist of evaluating if the increase in income resulting from multiplying the quantity $(q - q_1)$ times the price increase outweighs the income loss due to withholding q_2. The long-term contract mitigates the incentive to increase the price, since it reduces the amount of energy that would benefit from the price increase resulting from withholding part of the generator's output (from q to $q - q_1$).

This is only true, however, if the price laid down in the agreement does not depend on the spot market price: in other words, if when a forward contract is concluded, the spot market price is not suitable grounds for setting forward prices. But if the contract duration is not sufficiently long, the generator will clearly perceive that the short-term market price will be the reference price at the time of renewing the contract. Thus, the incentive to exercise market power for the generator will not be actually affected, since increasing short-term market prices will allow for a higher contract price

A pending issue, to be discussed later, is the way to properly price these long-term contracts that are mandated by the regulatory authorities.

Energy release or virtual power plants auctions

This method is based on auctioning, not generation capacity, but the right to manage production. Under this scheme a dominant company is required to auction the commercial management of part of its capacity for a limited but sufficiently long period of time, e.g., three to five years. This may materialize as a call option on a fraction of a company's (pre-defined) blocks of energy at a set price or by linking the contract to the performance of specific plants.

Energy release auctions have been used on a handful of occasions to limit the market power of dominant companies in markets such as Alberta, Canada, or The Netherlands. In other cases, such as in France, it has been used as a condition for authorizing merger and acquisition (M&A) processes to prevent significant erosion of competition levels.

This method is obviously less drastic than capacity divestment, provided that the physical facilities do not change hands and that the process is completely reversible when the contract expires. The advantage is that it attracts new agents, which can help to strengthen the retail market. It also reveals the market price of generation assets, which may be useful under certain circumstances.

As in divestment, these auctions should be spaced at sufficiently long time intervals and the maximum amount of energy auctioned should be limited to prevent prices from collapsing. A stable regulatory framework and an established market are also highly advisable prerequisites. Although to a lesser degree, this method also discourages the dominant companies from investing further in the local market.

Administratively priced long-term contracts

Very often, in line with the point just raised in the previous paragraph, the lack of market maturity and thus liquidity does not allow selling in the market the significant energy amount that should be released to properly mitigate the market power of the dominant player. The number of credible potential buyers in the auction could be too small, turning an oligopoly problem into an oligopsony situation, in which buyers are allowed to exert a great deal of control over the seller and can effectively drive down prices.

If this is the case, the only way to mitigate market power through forward contracting is by fixing the price externally. The regulator establishes the total volume of production to be contracted (and even the load profile), a term, and the price per MWh. This option has been implemented for instance in the Irish market, under the name of Directed Contracts (DC) and in Singapore. In the case of Ireland, according to the Single Electricity Market Committee [33] the prices of the DC are "determined by regression formulae that express the DC strike price in a given quarter and for a given product (baseload, mid-merit or peak) as a function of forward fuel and carbon prices". In the case of Singapore, the Energy Market Authority [10] states that the contract price is set based on the long-run marginal

cost of the most efficient generation technology that accounts for more than 25 % of the total electricity demand and taking into consideration the key policy objective.

Voluntary long-term contracts

Voluntary long-term contracts would in principle have the same mitigating effect on dominant positions in power markets as virtual contracts, for under such arrangements the revenues received by the company for the volume of energy involved are independent of market price. The longer the duration of the contract, the more intense is this effect, which vanishes when the contract term is overly short (less than three years, for example). For such a contract to be accepted by the regulator as a reduction of effective capacity, it must prove to be unrelated to parallel commitments that imply uncompetitive conditions. The establishment of a set of satisfactory transparency guarantees would be no easy task, however.

Other measures

Another market power mitigation measure consists of facilitating the entry of new producers by removing any regulatory difficulties or uncertainties that might serve as deterrents. This measure is very important if the long-term aim of lower concentration is to be reached and is, in any event, a measure required to ensure that a market remains competitive in the long term.

Similarly, competition can be enhanced by furthering interconnections between neighboring electric power systems to heighten competition between adjacent markets.

Finally, a more elastic demand, able to react to high prices, would also reduce market power. This can be achieved by demand response programs, information, or educational activities, and it can be facilitated by the new telecommunication technologies.

These mitigation measures are unable, in any event, to fully eliminate market power. Depending on the type of instrument adopted, the ability to manipulate the market price may remain intact, although the economic incentive to do so will be smaller. More significantly, the maneuvring by the oligopolistic firm that may be required to benefit substantially from the abuse of market power may be much more readily detectable.

The stringency of market power mitigation measures should be reduced as the level of concentration declines or other regulatory instruments are introduced, such as improved market supervision mechanisms, more active demand response, elimination of barriers for new entrants, a broader margin of available generation capacity over demand, reinforcement of interconnection capacities, an adequate level of market information for all agents, and enhanced competition on operating reserve markets.

References

1. Adib P, Hurlbut D (2008) Market power and market monitoring. In: Sioshansi FP (ed) Competitive electricity markets—design, implementation, performance. Elsevier, Amsterdam
2. AEMO (2010) An introduction to Australia's National Electricity Market
3. AIP (2007) The bidding code of practice. A response and decision paper. AIP-SEM-07-430. www.allislandproject.org. Accessed 30 July 2007
4. Baíllo A, Cerisola S, Fernández-López JM, Bellido R (2006) Strategic bidding in electricity spot markets under uncertainty: a roadmap. Power Engineering Society, IEEE general meeting, 2006
5. Batlle C, Barroso LA, Pérez-Arriaga IJ (2010) The changing role of the state in the expansion of electricity supply in Latin America. Energy Policy 38(11):7152–7160. doi: 10.1016/j.enpol.2010.07.037
6. Barker J, Tenenbaum B, Woolf F (1997) Governance & regulation of power pools & system operators. An international comparison. Papers 382, World Bank—Technical Papers
7. Boisseleau F (2004) The role of power exchanges for the creation of a single European electricity market: market design and market regulation. PhD Thesis, University of Paris IX Dauphine, Delft University Press
8. Bogas JD (1998) La convergencia de gas y electricidad en el mercado energético global (Gas and electricity convergence in the global energy market, in Spanish). Economía industrial, ISSN 0422-2784, no 321, pp 111–122
9. EFET European Federation of Energy Traders (2007) General agreement concerning the delivery and acceptance of electricity. www.efet.org. Accessed 21 Sept 2007
10. Energy Market Authority (2011) Vesting contracts. http://www.ema.gov.sg
11. ERGEG European Regulators' Group for Electricity and Gas (2006) ERGEG Guidelines of good practice for electricity balancing markets integration. Ref: E05-ESO-06-08. www.ergeg.org. Accessed 6 Dec 2006
12. ETSO European Transmission System Operators (2003) Current state of balance management in Europe. Balance management task force. Available at www.etso-net.org. Accessed Dec 2003
13. European Commission (2010) Interpretative note on Directive 2009/72/EC concerning common rules for the internal market in electricity and Directive 2009/73/EC concerning common rules for the internal market in natural gas. The unbundling regime. Commission staff working paper. Brussels, 22 Jan 2010
14. Eurostat (2012) Electricity market indicators. Available via http://epp.eurostat.ec.europa.eu
15. Green R (2010) Are the British electricity trading and transmission arrangements future-proof? Util Policy 18(4):186–194
16. Helman U (2006) Market power monitoring and mitigation in the US wholesale power markets. Energy 31:877–904
17. Hope E (2005) Market dominance and market power in electric power markets: a competition policy perspective. Norwegian School of Economics, Bergen
18. Joskow PL (2008) Lessons Learned from Electricity Market Liberalization. Energ J, Special Issue in Honor of David Newbery, pp 9–42
19. López-Velarde R, Valdez A (2008) "Mexico". In: O'Donnell EH (ed) Electricity regulation in 31 jurisdictions worldwide. Published by Getting the Deal Through. www.gettingthedealthrough.com
20. Milligan M, Donohoo P, Lew D, Ela E, Kirby B et al (2010) Operating reserves and wind power integration: an international comparison. Preprint. Conference Paper NREL/CP-5500-49019 Oct 2010
21. Newbery DM (1997) Pool reform and competition in electricity. DAE Working Paper 9734, Cambridge

22. Newbery DM, Pollitt MG (1997) The restructuring and privatisation of Britain's Cegb. Was it worth it? J Ind Econ 45(3):269–303
23. Newbery D, Green R, Neuhoff K, Twomey P (2004) A review of the monitoring of market power, for the European Transmission System Operators, ETSO, Nov 2004
24. New York Independent System Operator (2012) Ancillary Services Manual. Available at www.nyiso.com
25. O'Neill R, Helman U, Hobbs B, Baldick R (2006) Independent system operators in the United States: History, lessons learned, and prospects. In: Sioshansi FP, Pfaenberger W (eds) International Experience in Restructured Electricity Markets: What works, what does not, and why? Elsevier, 2006
26. O'Neill RP, Sotkiewicz PM, Hobbs BF, Rothkopf MH, Stewart WR Jr (2005) Efficient market-clearing prices in markets with nonconvexities. Eur J Oper Res 164(1):269–285. doi:10.1016/j.ejor.2003.12.011
27. Pérez-Arriaga IJ, Meseguer C (1997) Wholesale marginal prices in competitive generation markets. IEEE Trans Power Syst 12(2):710–717
28. PJM (2012). PJM Manual 11: energy & ancillary services market operations. Revision: 51. www.pjm.com. 8 Aug 2012
29. Rodilla P, Vázquez S, Batlle C (2012) Auction bidding, clearing and pricing schemes in day-ahead electricity markets: a comprehensive review. IIT Working Paper IIT-012-073A
30. Rose K (1999) Electricity competition: market power, mergers and PUHCA. NRRI, Ohio State University, 1999
31. Smith A (1776) An inquiry into the nature and causes of the wealth of nations. In: Cannan E (ed) 1904 Methuen and Co., Ltd., London. 5th edn. www.econlib.org/library/Smith/smWNCover.html
32. Schweppe FC, Caramanis MC, Tabors RD, Bohn RE (1988) Spot pricing of electricity. ISBN 0-89838-260-2, Kluwer Academic Publishers, Boston
33. Single Electricity Market Committee (2011) Directed contracts—2011/2012. Quantification and pricing. Decision Paper. SEM-11-045. www.allislandproject.org. Accessed 17 June 2011
34. Stoft S, Belden T, Goldman C, Pickle S (1998) Primer on electricity futures and other derivatives. Ernest Orlando Lawrence Berkeley National Laboratory, LBNL-41098, UC-1321, University of California, Jan 1998
35. The Brattle Group (2007) Review of PJM's Market Power Mitigation Practices in Comparison to Other Organized Electricity Markets. Prepared for PJM Interconnection, LLC. September 14, 2007
36. The Electricity Pool (2001) The electricity pool. www.elecpool.com
37. Thomas S (2006) The British model in Britain: failing slowly. Energy Policy 34:583–600
38. UCTE Union for the Co-ordination of Transmission of Electricity (2004) UCTE operation handbook. Version 2.5, level E. www.ucte.org. Accessed 24 June 2004
39. U.S. Environmental Protection Agency (2010) Electric generation ownership, market concentration and auction size. Technical Support Document (TSD) for the Transport Rule. Docket ID No. EPA-HQ-OAR-2009-0491. www.epa.gov. Accessed July 2010
40. Wolak FA (1999) An empirical analysis of the impact of hedge contracts on bidding behavior in a competitive electricity market. Fourth Annual POWER Research Conference, March 5, 1999

Chapter 8
Electricity Tariffs

Javier Reneses, María Pía Rodríguez and Ignacio J. Pérez-Arriaga

La tarifa no se fija, se calcula.[1]

White Paper for the Reform of the Spanish Electricity Market, 2005.

This chapter defines the theoretical objectives that regulation should pursue in electricity tariff design (also called ratemaking) and introduces the reader to the issues surrounding that task. Satisfactory tariff design is essential both to promote optimal short-term system usage and to guide efficient long-term demand response. This is because sound electricity tariffs convey information on responsibility in the incurred supply costs to the actors involved. The design of electricity rates is, then, of major importance both in liberalised and traditionally regulated systems.

The pursuit of greater efficiency has been the main driver of the important regulatory change that has shaken the electric power industry during the last two decades. But a number of questions arise around electricity tariff design. What is meant by efficiency? What are the implications of seeking such a desirable

[1] "Tariffs are computed, not decreed". This quotation from The White Paper on power sector reform, prepared by Ignacio Pérez-Arriaga for the Spanish Government in 2005, warns against tampering with electricity tariffs, a common practice of many governments, unfortunately. The role of governments is to establish a sound regulatory framework, so the activities necessary to supply electricity are efficiently performed, but not to interfere in the process of computation of the resulting tariffs.

J. Reneses (✉) · M. P. Rodríguez · I. J. Pérez-Arriaga
Universidad Pontificia Comillas, Instituto de Investigación Tecnológica,
Alberto Aguilera 25, 28015 Madrid, Spain
e-mail: javier.reneses@iit.upcomillas.es

M. P. Rodríguez
e-mail: piaror@gmail.com

I. J. Pérez-Arriaga
e-mail: ipa@MIT.EDU

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_8,

objective? Is efficiency the sole principle that should govern electricity tariffs? Is it even the most important criterion? What others might be taken into consideration? How can tariff design enhance efficiency? Can marginal costs be used as optimal economic signals?

A word of clarification is needed on the meaning of the term "tariffs". Some costs incurred in the supply of electricity correspond to regulated activities (mostly networks, plus other regulated charges) whose remuneration is determined by the corresponding regulatory authority. This regulatory authority also determines how these costs will be allocated and charged with a regulated tariff that is called the "access tariff". Therefore "tariffs" are regulated charges and they apply both in a traditional or a competitive regulatory situation. Under traditional regulation, also the costs of electricity production and commercialisation are regulated and the regulatory authority determines the corresponding charges to the end consumers, which are included in the comprehensive "integral tariff". On the other hand, under a competitive regulatory framework, consumers freely choose a supplier and each one pays the agreed price for the energy and the commercialisation service. In this case, all consumers and (typically) generators pay the common access tariff (implicit in the integral tariff under traditional regulation) and consumers pay the agreed market price with the chosen supplier. Sometimes, under a competitive regulatory framework, consumers are allowed to opt for a regulated integral tariff called the "default tariff", instead of having to select a supplier. And, in all systems where retail competition exists, a "last resort tariff" must exist to be applied in those emergency situations where a consumer may be left without a supplier. All these options will be covered in this chapter, and completed in the next chapter on electricity retailing.

This chapter is divided into six sections. The introduction is followed by a section on the theoretical fundamentals that should govern tariff design, along with the main theoretical approaches that have been adopted over time. That review of the approaches adopted in the past provides insight into the evolution of tariff design to what are presently regarded to be the most advanced solutions. Section 8.3, which focuses on the determination of the access charge to be paid by all system users, describes the methodologies that can be applied to the cost items defined and relates them to the theoretical fundamentals discussed in the second section. Section 8.4 addresses integral (or default) tariff design, which includes the access tariff (or use of system charge) plus the cost of the energy consumed. Section 8.5 reviews a series of miscellaneous issues that, while not pivotal to ratemaking, need to be borne in mind, while Sect. 8.6 sets out the conclusions.

8.1 Introduction

As noted, although electricity tariff design is a question of cardinal importance, not all of the factors involved have been studied in suitable depth, as this chapter will show. This is somewhat surprising because all electricity systems need to establish

rates for electric power in one way or another. The possible explanations for this situation include the lack of transparency that has been and continues to be customary in many countries and regions, whereby electricity ratemaking can be used as a very valuable political tool.

In any event, before listing the primary objectives that should be taken into consideration by regulators when designing electricity tariffs, a proposal for a full ratemaking procedure would appear to be in order. The notion of what rates or tariffs are appears to be clear, for ratemaking has been going on for many years. The steps that should be taken to define a final tariff structure, the visible result of that procedure, are not always obvious, however.

Tariff design can be divided into three fundamental steps [27]: (i) choice of remuneration methods and levels for each business activity (generation, transmission, distribution, retailing, system operation); (ii) definition of the tariff structure applicable to end consumers and lastly (iii) allocation of allowed costs to that structure. When presenting each phase, an attempt should be made to comply with general ratemaking principles, which are discussed in this section.

The first step covers two phases: choice of the remuneration scheme and calculation of the allowed costs. Both depend on the type of business and have been discussed in depth in the preceding chapters.[2] Mention should nonetheless be made of the fact that the term "tariff design" is often erroneously used (even in academic articles) to mean this phase only, ignoring the importance of the other two. This chapter, then, focuses on the other two steps: the establishment of a tariff structure and cost allocation. The starting point for this discussion is the assumption that the regulator, in compliance with relevant ratemaking principles, has established the total recognised costs for each regulated business or has the means to estimate the costs of businesses subject to competition, if necessary.

The second step calls for defining the tariff structure. Its design entails establishing consumer groups or tariff categories, the time intervals or periods subject to billing and the terms under which each category and period are to be billed. The choice of one or the other gives rise to a structure on which all possible customers appear. In addition, the design criteria should, as far as possible, comply with all the regulatory principles laid down in the following section. As discussed below, some of the principles limit or at least establish certain guidelines for tariff structure design.

The last step consists of allocating costs to the tariff structure, i.e. distributing the allowed cost items among each and every term on the structure. Like all other phases, cost allocation must be consistent with fundamental regulatory principles.

It might be thought that each of the above three steps can be conducted separately. Tariff design is not always a linear process, however. Recognised costs can be calculated more or less separately, but the definition of the tariff structure and the calculation of the amount to be recovered under each term in the structure are

[2] Except for the retailing activity, which is intimately related to tariff design and will be presented in the following chapter.

closely related, for the cost allocation process itself can raise the need for new categories or terms in the tariff structure.

Conceptually speaking, electricity tariff design must meet two main objectives.

- The first is to raise the money needed to pay for the costs of the activities whose remuneration remains under the regulator's control. Which activities are incorporated in the tariff depends on the adopted regulation, as explained in Sect. 8.2.4. Generation and retailing are only included in the tariff under cost-of-service traditional regulation, or when the regulator sets a default tariff that allows certain categories of consumers to remain under a regulated "integral" tariff, instead of finding a suitable offer from a retailer. This objective links structure design per se to the first step in the ratemaking procedure: the choice of remuneration schemes and levels for the businesses involved.
- The second objective is to send the right economic signals to each customer to favour the optimal socio-economic use of electricity. Consumer behaviour is logically influenced by the current price and its possible future changes. A customer not paying for electricity on a time-of-use basis fails to perceive the incentive to transfer part of the power used from peak to off-peak times. A customer not suitably charged for peak capacity sees no need to attempt to flatten his/her load curve. Such tariffs clearly fail to convey the right signals to ensure resource use at close to the social and economic optimum.

As discussed below, a balance between the two objectives may be difficult to reach, however. The revenues associated with an economically optimal signal may differ from the revenues needed to satisfactorily remunerate industry businesses.

8.2 Theoretical Tariff Design Fundamentals

As noted, electricity ratemaking consists of a series of steps taken with a view to reaching two basic objectives: to raise enough money to cover the allowed costs of the businesses involved and to send consumers the right economic signals.

Suitable decision-making criteria based on regulatory principles and the theoretical approaches presently in place are requisite to each step in tariff design. This section, which aims to describe the principles and approaches that are essential to such design, contains also an in-depth discussion of tariff structures.

8.2.1 Regulatory Principles

The three phases into which tariff design has been divided can be presented in many different ways depending on the objectives pursued with regulation. These objectives or principles define the characteristics that ideal or optimal tariff design should meet wherever possible.

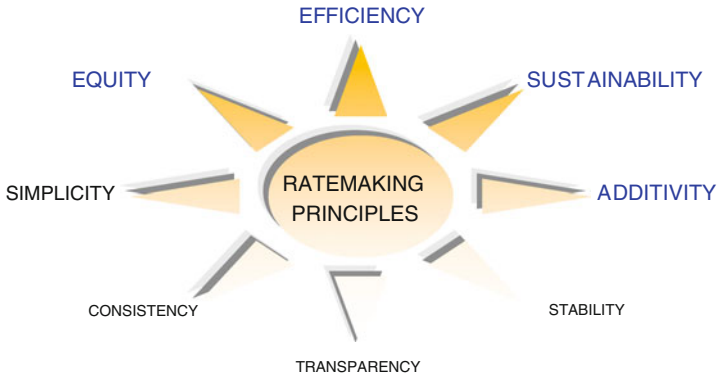


Fig. 8.1 Regulatory principles

Theoretically speaking at least, a general consensus has been reached on the regulatory principles that electricity rates should observe. While these principles provide guidelines for establishing tariffs, they also afford a certain degree of freedom or leave room for more than one interpretation. Thanks to this freedom, a number of design options can be envisaged.

The laws, directives and regulations enacted in each country as the basis for electricity tariff regulation and design almost invariably cite a number of fundamental principles [5, 6, 10, 14, 26, 27] (Fig. 8.1).[3]

- Economic *sustainability* or revenue sufficiency. This principle is the essential point of departure for tariff design. Any company that conducts a regulated business must be able to finance its businesses as well as any new investment required to be able to continue to operate in the future. This principle is directly related to the first phase of ratemaking, namely the calculation of allowed cost. The adopted tariff design must ensure recovery of this allowed cost, and payment of the electricity market price, when the tariff also includes this component.
- Economic *efficiency* in resource allocation. From the standpoint of economic theory, efficiency means that goods or services should be consumed by whoever benefits most from them. Efficiency so defined can be achieved by establishing a price signal (optimal for both the short and the long term) that will prompt each consumer to use the amount of the resource that is most efficient for the system as a whole (see Chap. 2 for a more detailed discussion). Economic theory sustains that this objective is reached by ensuring that prices are close to the marginal costs of providing the service. Therefore, costs must be distributed to reflect as closely as possible the amount that each customer costs the system, so that consumers can perceive the "electric" consequences of their decisions on power use. This efficiency principle affects all the phases of tariff design: the remuneration scheme

[3] The electricity tariffs in place in European Union countries must comply with the provisions of Directive 2003/54/EC, according to which they must be non-discriminatory and cost-reflective.

should further efficiency; allowed costs should be efficient (otherwise, consumers will not use resources appropriately); and the methods used in tariff design, both as regards the tariff structure and the way prices are established, should further efficient consumer behaviour. In short, efficient rates send consumers the most appropriate signals and constitute a powerful tool for efficient energy use [23]. In practice, this "marginal pricing principle" can be only implemented as a broad guide, because of the complex nature of the considered activities: networks with lumpy investments and strong economies of scale, or electricity markets with significant nonlinear operation costs, for instance.

- *Equity* or non-discriminatory access to the service and cost allocation. As a rule, non-discrimination is agreed to mean that equal power consumption should be charged equally, regardless of the nature of the user or the use to which the energy is put. Equity does not mean, then, that the same costs should be allocated to all grid users (to make this perfectly clear, this principle is often referred to as the fairness rather than the equity principle). From the standpoint of the electric power business, this principle ensures that the rates applied do not provide a given competitor (in this case, customers) any advantage over any other within the electricity system. Each country has established more or less restrictive measures regarding the implementation of this principle, which in some cases has been considered to be compatible with certain types of price discrimination [14], depending on the laxity in the interpretation of the term equity. One example is low income consumers: is it equitable (fair) for them to be deprived of electric power, defined as it is to be a basic service, because they are unable to afford it? Or is it more equitable (fairer) for this group of consumers to pay lower electricity rates?
- *Transparency*, complementary to the non-discrimination principle. The aim is to ensure transparency in the definition of ratemaking methodology, its specific application to each business and publication of procedures and results. The publication of tariffs and a clear and understandable description of the method used to establish them is the sole instrument available to verify whether or not the other principles (sustainability, equity, efficiency…) are being honoured.
- Tariff *additivity*, an outcome of the principles of sustainability and transparency. This means that the end rates should be the result of the sum of all the cost items applicable to each group of consumers. Rates should be calculated from the bottom up, beginning with an analysis of all cost items. In addition, the tariffs from which consumers are allowed to choose must be coherently structured. Hence, the sum paid by all consumers for each item should be equal to the total recognised cost of that item. Further to this principle, the impact of each remuneration item on the tariff can be analysed individually. This is consistent with the principle of economic efficiency and can be used to show consumers how the rates they pay are itemised. Additivity is requisite to objective and transparent tariff design [2].

Other criteria that may appear to be obvious but which must also be borne in mind are listed below.

- *Simplicity* of the methods proposed, as far as possible, while attempting not to forfeit other more important principles. The aim is to facilitate comprehension and acceptance.
- *Stability* of the methodology used, so that regulated actors are subject to the lowest regulatory uncertainty possible. Companies must be able to make their forecasts with some certainty. Overly high risk may deter investment, to the detriment of satisfactory electricity system operation.
- *Consistency* with liberalisation and the regulatory framework in place in each country at any given time. Specifically, the degree of industry liberalisation affects the choice of the allocation method and even the tariff structure itself.

In addition to the foregoing, other, higher ranking criteria laid down in each country's regulations may be applicable. Examples in this regard include the protection of low income consumers, application of uniform rates throughout a given region or a country or environmental considerations.

One relevant observation is the difficulty (actually, the impossibility) of simultaneously meeting all the above principles, at least in their full dimension. This is sometimes attributable to a lack of know-how, but it is often due to conflicts among the principles themselves. For instance, the principle of efficiency may at times clash with the principle of sufficiency (marginal prices, particularly when dealing with networks, do not provide for full cost recovery), equity (efficient cost allocation need not necessarily be socially equitable[4]) or simplicity (where very complex processes are involved). That notwithstanding, all these principles should be borne in mind to know why certain decisions are made, what aims are pursued and what is to be forfeited to reach them. The ultimate objective is to reach a reasonable balance among all the principles discussed here.

8.2.2 Tariff Structure Design

This section focuses on tariff structure design, viewed in light of the principles described above. It first analyses the meaning of cost drivers and their relationship with the variables that can be billed. The following discussion explains how space and time affect electric power consumption. Lastly, it describes the tariff structure, along with the elements that guide its design.

[4] The efficient allocation of some regulatory costs with no obvious criterion for allocation of cost responsibility (such as stranded generation costs, subsidies to domestic fuels or renewables or some network costs) should be efficiently made by Ramsey pricing (allocation of cost in inverse proportion to price elasticity), which is obviously discriminatory.

8.2.2.1 Cost Drivers

Economic theory uses the cost function as an indispensable element to analyse any process. In the pursuit of optimal tariff design, it shows that establishing a single rate for all customers is unsuitable. The tariff structure should reflect system costs and customer behaviour. It should be designed by grouping demand side actors as well as cost elements whose behaviour within the system is assumed to be similar. The tariff structure attempts to reconcile the information available on costs with the variables responsible for those costs, through which they can be recovered. The structure should consequently seek to reflect which system elements or variables generate costs and the degree to which they do so. These elements or variables are called cost drivers, because they can be used to explain or estimate the costs incurred. Their definition is of major importance in tariff design and must be studied judiciously to suitably reflect the costs associated with each regulated business [2, 14].

In the case of electric power, the greater share of system costs is determined by two fundamental variables: a customer's installed capacity (typically, the peak demand that can be handled by the facility in question) and the energy consumed, at a given connection point and time.[5] On these grounds, a very large number of cost drivers could be defined, based on the capacity available and energy consumed at each connection and during each hour of the year. In actual ratemaking practice (in keeping with the principles of simplicity and transparency), however, a manageable number of factors is defined from which the cost function can be estimated as accurately as possible.

Part of the system cost may be reasonably assumed to be linked to a charge that is a function of the capacity available to the consumer, which reflects the instantaneous capacity desired. This, from the ratemaking standpoint, would entail introducing a capacity charge. An essential factor in this regard is the specification of which capacity parameter is to be used as a cost driver. The alternatives are the peak power consumed (if measurable), the capacity defined in the contract or the installed capacity. The type of capacity chosen depends on the type of meter installed as well as the type of contract concluded with the consumer.

Costs linked to the energy or power consumed by each customer may also be defined, charging users for energy consumed as evaluated during a pre-defined period (from several months to a single hour, if smart meters, also known as interval or time-of-use meters, are installed).

In addition to these main charges, each customer is also billed a fixed amount to cover point of supply related costs (such as customer management costs), i.e. costs that depend on the number of customers. Lastly, most regulations established a one-off connection charge to cover the costs of providing a new customer with electric power.

[5] Depending on place and time, the system cost of consuming a given unit of energy may vary widely.

Table 8.1 Cost drivers

Cost driver	Associated costs
Energy	Costs associated with the amount of energy consumed during a given period of time
Capacity*	Costs associated with peak demand or the potential to reach that demand
Connection	Costs of connecting a consumer to the supply grid
Consumer	Costs associated with the number of consumers connected to the supply grid

*Installed, in contract or peak

Table 8.1 summarises the most common cost drivers and associated costs.

While several of the drivers defined are used in standard practice, certain authors [23] propose bundling as many of these charges as possible into the energy or power consumed charge.

Space- and time-based differentiation

One factor that needs to be taken into consideration in tariff design is that the cost of supplying electric power depends on when and where it is consumed. The cost of supplying 1 kWh in an urban area at night is obviously not the same as delivering that same kilowatt hour in a rural area during the day. Consequently, customer location and time of day constitute cost drivers.

As far as spatial differentiation is concerned, the grid needed to carry 1 kWh depends on where it is consumed, as do the associated costs. This variation in cost has primarily to do with the amount invested in the grid, but also depends on the energy lost during transmission and distribution, the congestion, the maintenance, the service quality required and so on. Kilowatt hours (kWh) can therefore be differentiated in terms of the location of the service connection. That differentiation also depends on the grid businesses involved and how they are conducted. A consumer in a rural area, for instance, typically generates high distribution costs. But if that consumer is located within a mostly exporting area—i.e. with excess generation over demand–, the associated transmission costs are very low or even negative.

The cost of supplying 1 kWh also varies with the time of day, as a distinction must be drawn between peak and off-peak demand. On the one hand, when power is consumed during peak demand, the peak grows, necessitating higher investment in grids and power plants to cover consumption at such times. On the other, during peaks, electric power is produced by plants with higher marginal costs (fuel costs) that are used at peak times only, raising production costs. Furthermore, since grid losses rise quadratically with intensity, the heavier the load on the lines, the greater are the losses per kWh. As in spatial differentiation, time-related costs also depend heavily on grid factors. For instance, a user may consume power in system off-peak times (when generation costs are lower) and still have a negative impact on grid costs if his/her use concurs with a local peak.

Both space and time differentiation must be applied to cost drivers to establish a tariff structure. This means that a single charge for energy or power consumed cannot be defined; rather, this charge depends on the consumer location and the time of day consumption takes place [27]. No single fixed charge may be defined either, because of the same reasons.

The tariff structure, in short, is meant to be a simplified version of reality. Highly detailed rate structures reduce unfairness, subsidies, inefficiencies and discrimination. But at the same time, greater detail implies greater complexity in a context in which for the time being that may not be socially acceptable or advisable. An overly complex tariff structure may also involve excessive and unjustified rate calculation and billing costs (including meters and information processing).

The structure should, then, reflect the existing complexities, but only to a certain extent. Customer categories should be established, along with the various terms in each category, bearing in mind that reality needs to be simplified. The customer groups defined must be carefully designed and with the smallest possible error (given the consequences of that error in terms of non-compliance with ratemaking principles such as efficiency and discrimination).

Bearing all these considerations in mind, implicitly at least, each country has adopted a tariff structure in keeping with the characteristics of its electric and regulatory systems. Some general traits follow:

- Different tariff categories are established for consumer groups regarded to originate similar costs that can be distinguished from the costs generated by other categories of users. The first consideration here is the voltage connection.[6] But in addition, further to the most widely used cost drivers, consumers in the same category should have both similar consumption patterns and similar location-related circumstances.
- While arguments have been wielded to prevent overly aggressive differentiation, groupings can be established that send the right location-related signals. At this time, the existence of both nodal prices (see Chap. 6) and of several distributers offering consumers different prices is socially and politically acceptable in a considerable number of countries. This is one type of spatial discrimination. In other countries, by contrast, with a single rate in effect everywhere, any two users having the same connection voltage are grouped in the same tariff category.
- Tariffs may differ depending on the hour of the year. As a rule, hours are grouped by periods[7] with similar consumption patterns, with all hours in a period charged at the same rate. The aim is to reflect the differences in the economic impact of consuming power (or being able to consume power) in different time periods. Tariff periods group hours by level of responsibility for the costs incurred. Such clustering consists of two steps: seasonal blocks, if any, are first defined and then the hourly intervals within each season are established. Periods are normally determined in accordance with energy prices or total system demand, even though the most accurate procedure would be to define

[6] The voltage levels that are to define different tariffs must be established. As a rule, not all the voltages existing in the system are taken (for that would run counter to the simplicity principle). Rather, they are grouped in three or four levels.

[7] The implementation of advanced (hourly, or interval or time of use) meters would make it possible to establish hourly differences. No consensus has been reached on their universal application, although the trend towards its generalised deployment is clear.

intervals by the cost of the various activities in each hour [3]. Generally speaking, prices and demand are very closely related to generation and transmission costs, and perhaps less to distribution costs, where grid saturation patterns at certain voltage levels or in certain areas may differ substantially from the overall system profile.[8] Lastly, the definition of hourly intervals calls for a practical compromise to ensure that each rate interval always covers the same hours: on weekdays, all the hours from 4:00 to 10:00 p.m., for instance, should be regarded to be peak time. Furthermore, the tariff periods defined for any given category must adjust to the characteristics of the meters installed at each type of connection.

8.2.2.2 Tariff Structure

Tariff categories and periods as well as cost drivers are used to build the tariff structure. This is a table in which tariff categories are normally shown in the rows and the tariff periods in the columns. Each resulting cell contains the cost assigned to capacity, energy consumed and number of customers. The costs of each business must also be assigned to these three drivers, as explained in greater detail in Sects. 8.3 and 8.4.

When costs are not defined for all the hourly intervals in a given rate category, several columns are merged into one. A conceptual example of a tariff structure is given in Table 8.2.

8.2.3 Marginal Cost-Based Approaches

The tariff structure with its tariff categories, periods and cost drivers is then used in the third stage of ratemaking: the allocation of the cost items to the cells on the tariff structure. This is a very complex task for which no universally accepted procedure has been found. By way of introduction to the specific analysis of the allocation of each cost item (in Sects. 8.3 and 8.4), this section describes the possible theoretical approaches that can be adopted. The specific use of these approaches to allocate costs to each business is illustrated below, in the aforementioned sections.

Historically speaking, as far as the authors are aware of, the first theory-backed methodologies to be used in electricity tariff design were based on corporate analytical accounting. For that reason, they are known generically as the

[8] Electricity energy prices and demand are not necessarily well correlated. For instance, in Costa Rica the highest electricity demand coincides with the cold season, which also happens to be the wettest one, with a high level of hydro production and lowest electricity energy prices. This forces to consider more time periods to account for electricity energy price levels and also demand levels for the allocation of network costs.

Table 8.2 Sample tariff structure

		Winter			Summer		
		Peak	Shoulder	Off-peak	Off-Peak	Shoulder	Peak
LV < 1 kV	$P^* = 1$	€/kW €/kWh €/customer					
	$P = 2$	€/kW €/kWh €/customer			€/kW €/kWh €/customer		
	$P = 3$	€/kW €/kWh €/customer		€/kW €/kWh €/customer		€/kW €/kWh €/customer	
MV > 1 kV and <33 kV	$P = 3$	€/kW €/kWh €/customer		€/kW €/kWh €/customer		€/kW €/kWh €/customer	
	$P = 6$	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer
HV > 33 kV and <72 kV	$P = 6$	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer
EHV >72 kV and <220 kV	$P = 6$	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer	€/kW €/kWh €/customer

* P number of time periods in the considered tariff. This table does not correspond to any existing tariff structure; it is only meant to show how a tariff structure could look like

accounting approach. These methods were meticulously developed and implemented in most USA states—and are still in use today in some of those states—as well as in some countries with vertically integrated electricity systems. The main objective of the accounting approach is to recover all the cost items posted in companies' accounts, to which end each item is allocated, ad hoc, to the cells on the tariff structure. While the method constituted a significant step forward in its time,[9] it does not send consumers the most suitable economic signals (because the cost distribution criteria are not optimal) and from the standpoint of sound tariff design theory, its use is not recommended.

Since the mid-twentieth century, in some countries, progress began to be made in the application of classic economic principles to electricity tariff design. Hence, the use of marginalist theory, and specifically long-term marginal costs (LTMC), began to be explored as the basis for suitable economic signals in electric power pricing. The earliest studies were conducted in France by Électricité de France (EDF) staff, especially Marcel Boiteux. The idea underlying the application of marginalist theory for this purpose is that the resulting rates are fairer and send economic signals that maximise social welfare. The earliest studies found that the most suitable signal would be provided by LTMC, which is defined as the increase in the network infrastructure and associated operation and maintenance costs attendant upon a sustained rise in demand over time.[10] Both the operating costs and the costs of adapting the existing facilities, i.e. investment costs, would rise with demand. Most of the proposals put forward along these lines, which focused primarily on generation, can be found in Joskow [12] and Munasinghe [21]. For applications involving electricity systems, LTMC is often replaced by long-term incremental cost (LTIC). The reason is that finite increments are more appropriate for calculating marginal cost in this industry, where decisions to invest in new facilities are discrete. The main advantage of using marginal cost-based rates is they constitute an attempt at making that each consumer defrays the system costs incurred by his/her own use. One problem is that marginal rates do not generally recover allowed costs completely (especially for the networks, due to economies of scale), and consequently call for significant adjustments that may ultimately distort the economic signals sent to users. Another problem is that LTMC or LTIC are difficult to estimate and are based on questionable assumptions. That notwithstanding, these approaches are still in place in many countries even today, essentially for grid tariff design, as discussed in Sect. 8.3.

With the advent of restructuring and liberalisation of the power sector, electricity is traded in wholesale markets with short-term energy prices. It has been

[9] One of the most prominent advantages is its theoretical simplicity, and the need to have standardised regulatory accounting in place, which is particularly useful from the regulatory standpoint to determine the remuneration for regulated businesses. For background to this approach see the description of the traditional regulation in Chap. 3, and in particular in Annex A of this book.

[10] Long-term marginal costs can, by definition, accommodate space differentiation. In other words, increases in cost depend on where demand rises.

shown in Chap. 2 that the application of short-term marginal costs is the most efficient economic signal for power system operation [29]. The short-term marginal cost (STMC) is defined as the increment in operating system costs resulting from a per unit increment in demand at any given time and network node. In most countries, wholesale electricity market design has evolved towards the use of STMCs (normally hour by hour, but also every half hour or even every 5 min) as the optimal economic signal for energy trading. Therefore, STMCs for electric energy are readily available without ambiguity in space and time. The primary advantage of using STMCs in operation is that the resulting economic signal optimises consumers' response to the cost of supplying electric power [13] as well as the response of generators. Since STMC differs by location, when applied to energy injected and retrieved at the different nodes, they result in some rents that could be used to pay for a fraction of the total network costs. Note the essential difference between LTMC of networks and STMC of energy, as described here: while LTMC refers to network total costs, STMC refers to electric energy costs. Therefore, these two approaches do not only differ in being short or long in time, their underlying concepts have nothing to do with one another. While LTMC of networks is an attempt to assign responsibility of the increment of the demand at each node in the development of the network, the STMCs of energy look at the earnings that the network could make in the short term by purchasing cheaper electricity in some nodes and delivering it (minus losses) at other nodes at a higher price. Under ideal conditions of continuity in the investments, absence of economies of scale and others (see Chap. 6 for transmission), both approaches would be able to allocate the complete network costs to its users. In practice, none of them can, unless they are tweaked somehow. Regarding STMCs, as shown in Chap. 6, their use for grid remuneration in general leads to the recovery of only a small fraction of the total cost, making it unsuitable for tariff design.

The network cost allocation problem has only become relevant in parallel with the process of restructuring and liberalisation. Under traditional regulation only consumers had to pay for network costs and almost universally the end user regulated tariffs are uniform (i.e. no geographical differentiation) for the same class of consumers that are connected to the same voltage level. On the contrary, when the transmission activity is unbundled and there is competition at wholesale level, transmission cost allocation matters, as it is now of essence to send sound locational signals to prospective new generation investors, and sometimes also to large consumers. And in the anticipated future distribution networks, teeming with distributed generation and demand response capabilities, locational signals might be also an advisable regulatory instrument.

Since the strict application of marginal principles is unable to give an adequate response to the network cost allocation problem for the reasons described above, it has to be complemented or replaced by other cost causality or beneficiary pays approaches, as explained in Chap. 6 (for transmission) and also Chap. 5 (for distribution). The starting point in these approaches is that the costs to be paid by the network users should be allocated in accordance with each user's responsibility (or associated benefit) for each cost item. The possible methods derived from this broad

principle have been already developed in Chaps. 5 and 6. The main difficulty in applying these approaches lies in practical issues, since they require a detailed understanding of the planning function used by the companies involved in each area of the electricity network business, to be able to accurately allocate costs [14].

8.2.4 Access Charge and Integral Rate

As discussed in this and preceding chapters, the regulations in place for each electricity industry business have a direct effect on tariff levels (allowed revenues). But the regulatory factor with greatest impact on the methodology adopted is the degree of industry liberalisation, which affects not only rates, but the choice of the allocation method and the very structure of the tariff structure.

Two broad types of electricity tariffs can be distinguished in any electricity market. In a wholly liberalised market, tariffs need not be designed to take generation and retailing businesses into consideration. Consumers pay the network activities with an access tariff, while purchasing power from the supplier of their choice at a freely established price. In this case, the only regulated price is the access tariff, which covers the costs of all the activities whose cost is determined by the regulator, such as the networks, system operation and different types of subsidies. The design of the access tariff is dealt with in Sect. 8.3.

When the electricity market is not liberalised, the price of electric energy and customer management costs must be included in the price of electricity, conforming to what is known as the integral tariff. In practice, this tariff often co-exists with the access tariff when the electricity market is in the process of liberalisation or even after it has been fully liberalised, if the regulator decides to establish a default tariff, as shown in Chap. 9. The integral tariff is addressed in Sect. 8.4.

8.3 The Access Tariff

The access or use of system (UoS) tariff covers the electricity system cost items that must be paid by users inescapably, either separately (if participating on a liberalised market) or as part of the integral rate. These cost items, which are determined by the regulator, include the costs associated with regulated businesses along with any other cost the regulator deems should be paid by all consumers.

This section analyses the methodology for allocating the cost items that form part of the access tariff. These costs can be classified into three distinct categories.

- Costs for using the transmission and distribution networks. These normally account for the larger part of the access charge. No universally accepted methodology for allocating grid costs is in place, and a variety of criteria have been adopted for this end.

- The allowed customer management costs incurred by distributors to attend to the customers connected to their grids.
- Other regulated costs, including the costs inherited from preceding regulatory systems and present costs that are deemed to be attributable to power system agents, regardless of whether or not they are liberalised market actors. Some of the costs most commonly included under this heading are listed below.

  Functioning costs of the System Operator, the Regulatory Commission or Market Operator (this cost might be recovered with some charge applied to market agents).

  Stranded costs. In systems undergoing substantial regulatory change, the regulator may consider that it is fair to compensate those former system agents that might be hurt by the change. These costs must be assumed by all consumers.

  Costs associated with environmental and energy diversity policies. This may include the cost of subsidising renewable energy, local sources of energy and energy efficiency programmes.

  Positive or negative deviations over the previous year's revenues with respect to allowable revenues. This issue is discussed more fully in Sect. 8.5.1.

  Other costs stemming from industry-specific regulation.

8.3.1 Methodologies for Allocating Network Costs[11]

As noted above, no universally accepted methodology is presented in place to allocate transmission and distribution costs to the tariff schedule. It, therefore, continues to be an open and much debated issue.

In most of this section transmission and distribution costs are considered together, despite the existence of specific proposals for each of these businesses. The reason is that here it will be assumed that the initial task of allocating the cost of the transmission network among countries or regions, and even to specific areas or nodes when this is the case, has been done already following the procedures explained in Chap. 6 (Electricity Transmission). Therefore, the job that is left for this section is to design the most adequate format for this component of the tariff for the network users in a certain area or node, once the total cost of transmission corresponding to that area or node has been predetermined. Consequently, for most intents and purposes of methodology, the transmission grid can be regarded to be simply another voltage level on the distribution grid. In many countries, the transmission grid is paid for partly by consumers and partly by generators [11], for

[11] The reader is advised, before studying this section, to review what has been already said about network cost allocation and pricing in Chap. 5 (distribution, in particular Sects. 5.2.3—network charges—and also 5.7—impact of distributed generation—) and Chap. 6 (transmission, in particular Sect. 6.5 on transmission cost allocation, as well as Sect. 6.5.1 on the requests of connection to the grid).

it benefits and is used by both. The only change in allocation methodology required under such circumstances is that the part of the transmission grid cost defrayed by generators must be deducted from the total and only the difference allocated to consumers. Also in distribution networks there is connected generation and in some power systems already in significant amounts. Specific tariff for distributed generation will have to be developed, but this issue is only briefly discussed in this chapter, in Sect. 8.5.5.

The accounting approach

The initial attempts to establish a methodology based on the use of the principles of economic efficiency to set grid rates were made in the framework of the accounting approach. As explained in Sect. 8.2.3, in that approach business accounting was used to allocate cost items to the cells on the tariff schedule. One of the major contributions of this type of design (implemented primarily in the USA) was the development of a procedure for systematising ratemaking. That procedure can be translated into a three-stage cost allocation process. Since some of these steps are necessary in causality principle-based methodologies (Sect. 8.3.1.2), they are described more fully under that item.

The billing variables through which costs are to be recovered are defined in the first step. As discussed below, this question is far from having a single solution, although a number of proposals have been put forward. The earliest approaches to the problem (accounting approach) made provision for the fact that network cost is not entirely a result of the capacity for which it is designed (this issue is dealt with in greater depth in Sect. 8.3.1.2). They consequently proposed allocating all grid costs to the capacity charge, with the exception of the so-called minimum distribution grid, a fictitious grid that interconnects all consumers, but carries no electric power. The cost of this latter grid was to be allocated to the number-of-consumers variable, i.e. to be recovered by means of a fixed charge billed to each consumer.

The second step in this cost allocation approach consists of distributing costs across the rate periods defined. Under the accounting approach, this entails merely dividing the cost allocated to capacity. Nonetheless, a large number of complex methods of particular interest arose that are still being used in many of the methodologies applied today. One of those methods, the probability of contributing to the peak (PCP) method, distributes the cost across tariff periods in accordance with the likelihood that the hours in each period will concur with grid peak hours.

Lastly, the third step entails cost allocation (by cost driver and period) to tariff categories. Again, a substantial number of methodologies were developed to solve this problem. Many remain current, the coincident and non-coincident peak demand methods being among the most prominent.

As noted, while these approaches led on occasion to scantly efficient cost allocation based on ad hoc criteria, they constituted the point of departure for tariff design and more specifically for cost causality methods. They will not be further explained here.

Marginal cost-based approaches

The application of STMC (so far only implemented at transmission level in some power systems) at every node may result in the collection of some revenues, which could be used to partly cover the total network charges, as shown in Chap. 6. Note that the practical application of this methodology to transmission grids has shown that cost recovery was normally under 20 % [24, 25]. Any "suitably modified" STMC that could recover the network costs completely would distort the STMC message completely. Therefore, even in those systems that have implemented STMC at transmission level, another method for network tariff design must be used. The application of STMCs to distribution networks is even less suitable, since cost recovery may be even lower and the signals sent to users would violate a number of tariff principles, such as equity and stability. Indeed, the STMCs for two very similar users may be totally different because of the existing grid layout. If, for instance, a user is located at the beginning of a feeder line, his/her STMC may be much smaller than a user located at the end of that same line. This situation may be due to completely arbitrary grid planning or network operation criteria and might change if a new line were to be laid, or a network reconfiguration takes place, making these economic signals very unstable.

For this reason, the application of marginal cost-based principles to grid tariff design will be limited to the use of LTMCs. Section 8.3.1.1 describes the application of this approach to grids in detail.

The other approach most widely applied in grid ratemaking—which in essence is not that much different—is based on the cost causality (or beneficiary pays) principle. That approach is addressed in depth in Sect. 8.3.1.2.

8.3.1.1 Long-Term Marginal Cost-Based Network Rates

The earliest papers that proposed the use of LTMCs in grid tariff design were authored by French economists [7] in the mid-twentieth century. They justified the approach on the grounds that consumers must pay the costs they would generate in a perfectly adapted network, for they are not responsible for grid planning.

While the LTMC approach is conceptually attractive in grid tariff design, its practical application poses several problems. First, LTMC must be properly defined, since its meaning is not obvious in the context of electricity grids. In fact, while a number of methodologies based on this approach have been put forward, no consensus has been reached on how to calculate LTMC. No wonder, since the lumpiness of network investments and the fact that strong economies of scale exist for network investment costs, make a strict implementation of marginal costing to be impossible. Moreover, the application of such methodologies to actual networks normally calls for very complex planning models to perform the calculations. Lastly, as shown in Chap. 6, the use of LTMC for grid tariffs usually entails the non-recovery of the revenues allowed for the businesses involved, necessitating revenue reconciliation or adjustments to attain such revenues. Depending on their magnitude, these adjustments may significantly distort the signals.

For grids, the LTMCs are normally replaced by LTICs to mitigate the investment lumpiness problem. LTICs can be calculated with planning models which, defining the existing grid as the baseline, optimise an expansion plan for a given demand-side trend.

The new problem with LTICs is that the results are highly dependent on how close or far is the current network from being "perfectly adapted" (i.e. optimal, for a given demand level). With LTMCs this can be avoided by taking as the reference network the one that is perfectly adapted to the current demand. But this network could be very different from the actual one, and the results may be very questionable.

A review of the most relevant methodologies that have been proposed for calculating LTMC-based grid tariff design are described below, followed by a discussion of possible revenue adjustment procedures.

LTMC- or LTIC-based methodologies

The paragraphs that follow summarise the main proposals found in the literature on the use of LTMCs in electricity network rate design.

Williams and Strbac [31] introduced a methodology used by English and Welsh distributors to establish distribution rates. It is based on the DRM (Distribution Reinforcement Model, also called the Generic Pricing Model or 500 MW model), an expansion model whose baseline is the existing grid, on which demand is raised by 500 MW at each voltage level.[12] LTICs can be calculated with this model. Costs are allocated either entirely to capacity (although some consumers are charged for energy used, based on load curves) or an energy use charge is defined for upstream facilities and divided among the tariff categories in accordance with their estimated contribution to peak demand.

Marangon Lima et al. [20] addressed distribution tariff design in Brazil. Instead of marginal cost, these authors proposed an expansion plan-based mean incremental cost, calculated from an aggregated facilities model. Cost was then divided among consumers in accordance with their contribution to area peak demand and the area's contribution to upstream peak demand (for which load profiles were used).

Another LTIC-based proposal for distribution grids was put forward by Ponce de Leão and Saraiva [25]. Here also, calculations were performed with a long-term planning model, here based on operating costs, reliability (power not supplied) and investment.

Li and Tolley [16] presented an innovative way to calculate marginal or incremental costs. According to these authors, conventional LTMC or LTIC calculation (evaluating the investment needed to cover predicted increases in demand and generation) poses two problems. On the one hand, it is a passive approach, for it makes no attempt to modify growth predictions, and on the other, demand and generation forecasts are subject to a good deal of uncertainty. Consequently, they proposed a methodology that attempts to reflect the cost of anticipating or delaying investment by calculating the number of years of service life left for each facility.

[12] The choice of a 500 MW increment is based on the fact that it has a sufficient impact on the grid, without diluting the effect of using the existing grid as a point of departure.

It then assumes a variation in demand over that number of years. The difference between the present values of reinforcements with and without the variation in demand is used to calculate the LTIC associated with the facility. This methodology is applicable to both positive (consumer) and negative (distributed generation) variations in demand and the result is applied to the capacity charge only.

Taking a similar tack, Li [15] developed the same methodology but using analytical expressions to calculate LTMC instead of incremental cost. Li and Matlotse [17] also applied this methodology to calculate reactive energy tariffs, using the reinforcements required to offset voltage limit violations.

Parmesano [23] staunchly defended the use of LTMCs. This author proposed the recovery of local distribution grid costs with a fixed charge based on design demand, on the grounds that these facilities are not marginal due to minor changes in the demand. This fixed charge may be a monthly charge, a capacity charge or a surcharge on a first tranche of power consumed, paid by nearly all users. For transmission and high voltage distribution grids, by contrast, the paper proposed using charges that would vary by hourly intervals based on the likelihood with which an increase in demand in each interval would affect grid costs.

The Portuguese tariff design methodology is described in Apolinário et al. [2, 3] and specifically applied to distribution in Apolinário et al. [4]. Tariffs are based on LTMC, although what is actually calculated is LTIC. These researchers used a formula based on the yearly investments needed to meet an increase in demand at each voltage level. In addition to contract capacity (or peak capacity for metered consumers), the billing variables proposed include the mean capacity during the peak period.[13] Specifically, the charge for using the transmission grid and the central part (i.e., the part used by many consumers) of the distribution grid is included in this billing variable, because a consumer's impact on grid cost is proportional to his/her peak capacity during system peak periods. The cost of distribution grids close to the points of supply, in turn, is recovered by a contract capacity charge, for grid sizing is conditioned by a small number of consumers (on occasion, only one).

Revenue reconciliation

As noted, the use of LTMC- or LTIC-based tariffs does not ensure the total recovery of the allowed costs of network businesses, due primarily to the existence of lumpiness, economies of scale and the difference between the present network and the one that is perfectly adapted to the present demand.[14] For this reason, these rates generally need to be adjusted, a process known as revenue reconciliation.

No satisfactory approach has yet been found to this problem, in particular where reconciliation is of significant magnitude, i.e. when there is a material difference between allowed costs and costs recovered through LTMC. In fact, revenue

[13] It is actually a charge equivalent to an energy use charge during the peak period.

[14] Contrary to the STMC approach, no detailed studies have been conducted on what part of grid costs would be recovered under LTMC-based methodology.

reconciliation is not even mentioned in a substantial number of proposals. The most elementary adjustments are made by applying coefficients to the rates. Two types of coefficients can be envisaged: multiplier and additive. In the former, the rates obtained are multiplied by the ratio of the allowed costs to the LTMC-based revenues. This maintains the relationship between rates for the various consumer categories and tariff periods. In this case, the LTMCs are used as coefficients to recover total costs [4, 20, 31]. Another solution uses additive coefficients, i.e. adding the same amount to all rates, thereby maintaining the absolute difference between categories and periods, keeping consumers from shifting consumption from one period to another as a result of the adjustments [23].

Economically speaking, the objective is to minimise the effect of reconciliation on consumer behaviour. In this regard, if the consumers' utility function were known, reconciliation could be conducted efficiently, further to so-called second best methods. One example of these methods, often proposed for electricity tariff design, is to be found in Ramsey prices [14]. In this method, adjustments are allocated to consumers in inverse proportion to their elasticity to price variations. Non-elastic consumers would pay more. Despite its pursuit of economic efficiency, this methodology has two important drawbacks that make it scantly recommendable in practice. The first is that it is discriminatory, violating the principle of equity, for consumers' private data and not only the technical and economic characteristics of the power consumed are used in the allocation (two consumers with the same consumption pattern could pay different tariffs depending on the use to which the electricity is put). As a rule, Ramsey prices are detrimental to domestic consumers, who tend to be the least elastic in developed countries.[15] The second drawback is that, from a practical standpoint, reliable data on consumer elasticity cannot be readily obtained. One alternative to Ramsey prices (used in Portugal and Spain) is to define elasticities not on the basis of consumers, but of time intervals. In this approach, peak time consumption is regarded to be least elastic.

In pursuit of an efficient signal, Parmesano [23] proposed a more qualitative approach, which would involve adjusting the fixed charge or creating a surcharge on the first tranche of power consumed, which is paid by practically all consumers.

Lessons to be learnt from LTMC- or LTIC- based methodologies

By way of conclusion, LTMC- or LTIC-based grid tariff design has been and continues to be researched. At the origin, the rationale for these schemes was to design network tariffs as optimal economic signals under ideal conditions. But LTMPs and LTICs can also be seen as an attempt to search for responsibilities in network investments. This is how the method of Investment Cost Related Pricing (ICRP) was developed in the UK in 1990 and it is still used now. All are nonetheless subject to a number of practical difficulties that have acted as a deterrent to their application to network tariff design.

[15] The opposite is typically the case in some developing countries with protected local industries not exposed to competition and impoverished populations.

One of the major problems posed by these approaches is the need to reconcile revenues, a practice that ultimately distorts the signal emitted (and the pursuit of that signal is the justification for using LTMC or LTIC). If the signal obtained is ultimately distorted, calculating rates from marginal costs may in the end be fairly futile.

Another important problem is that LTMCs and LTICs are calculated with optimal expansion tools, which are potentially subject to questionable criteria or manipulation.

Lastly but no less importantly, the approaches proposed focus on high voltage transmission and distribution grids, essentially ignoring low voltage grids. Many proposals call for facility-by-facility calculation [16], which is not feasible at the low voltage level.

8.3.1.2 Network Tariffs Based on the Cost Causality or Beneficiary Pays Principle

Ideally (i.e. no lumpiness, no economies of scale, no economically unjustified reliability constraints and no planning errors), marginal pricing is a theoretically sound approach to assigning responsibility for network investment. However, the difficulties described in the preceding item on the use of LTMCs have given rise to more direct approaches to the application of principles of cost allocation based on cost causality or (equivalently) beneficiary pays. The central idea of the new family of approaches is to consider cost allocation as a by-product of network planning: Network reinforcement has to be justified as the most efficient response to demand growth or—in a market environment—because the cost of the reinforcements is less than the entailed aggregated benefit for all network users. Therefore, the planning process must provide enough information to determine how to allocate the costs.

The new family of approaches borrows from the accounting method, the emphasis on considering the totality of costs to be allocated from the outset, but now using much more economically efficient allocation criteria. These approaches feature a number of advantages that make them very promising from a practical standpoint, such as set out below.

- The economic signal emitted is efficient if the causes underlying the incurred costs are accurately identified. The causality principle can be likened to a very LTMC model that includes, in addition to grid reinforcement costs, the cost of replacing all the existing infrastructure at the end of its service life.
- These are very robust methodologies because they ensure cost recovery (sufficiency principle). In fact, they start from allowed cost, which is distributed over the various cost drivers, periods and consumer categories in accordance with their responsibility for costs. If successful in allocating all costs, one advantage of these methodologies over LTMC is that they call for no subsequent revenue reconciliation.
- These are signals that (if the right criteria for identifying the causes underlying cost are used) comply with the non-discrimination principle, for two users responsible for the same grid cost are charged the same.

- Even though the methodology is complex in theory (inasmuch as it entails identifying the causes underlying each cost, which in turn calls for a detailed understanding of the companies' planning function), reasonable simplifications can be made in practice that lead to simple and transparent tariff design.

The practical application, then, entails designing mechanisms to identify the parties responsible for each cost inherent in grid businesses. The function reflecting the relationship between costs and their causes is (rather appropriately) called the cost causality function. As in the case of the LTMC approach, no single methodology has been universally accepted and a fairly large number of different approaches to the problem have been adopted. As a rule, the necessary know-how must be drawn from grid planning experts to ascertain the reasons for investment and on those grounds to allocate grid costs to each cell (cost drivers, time period, category) on the tariff structure.

In electricity grids, nearly all costs can be regarded to be fixed, for most are the capital costs associated with grid infrastructure and operation and maintenance costs, which may be regarded to be proportional to capital costs.

The causality function for these costs is none other than the decision-making process that leads to investment in grids. The grid planning function must therefore be studied and an analysis conducted of the consumer characteristics underlying investment decisions and the respective weight of each such characteristic in each investment decision. The causality function may be formulated in greater or lesser detail, depending on the degree of differentiation and complexity desired for the final tariff schedule. The implementation of individually personalised rates would naturally be unthinkable for two reasons: First, the amount of information needed, its management and tariff handling itself would be practically impossible; and second and more importantly, given the characteristics of energy distribution, applying the principle of causality consumer by consumer could lead to highly discriminatory rates. In other words, there is no single or best distribution grid[16] and there is no way to choose one over another if both have the same costs and requirements. This means that the cost allocated to a specific charge is different in each possible grid and that the cost originated by a given consumer cannot be regarded to be unique. Rather, it depends on the grid chosen.

The tariff schedule must consequently reflect but simplify the causality function obtained. The categories on a tariff schedule must be defined taking these simplifications into consideration, and by grouping variables and consumer groups that give rise to similar costs.

As a general rule, four distinct steps are involved in cost causality principle-based grid tariff calculations. The first is the definition of the cost drivers to be used and the allocation to each driver of the part of the cost to be recovered.

[16] A number of optimal sites may be defined for substations and transformer stations. Load proximity to or distance from the respective transformer station, for instance, affects its cost to the system. Nonetheless, a consumer should not benefit from or be punished for a design criterion over which he has no say.

The second is the establishment of a grid model with which to allocate the cost of each voltage level to the actors responsible for that cost (typically, downstream consumers connected at that level). The third step is to divide the cost associated with each driver among the tariff periods. The fourth and last step is to distribute the cost by driver and period over the tariff categories. On completion of the fourth step, all the cells on the rate schedule will be filled in. These steps are described below, along with some of the most relevant methods proposed for each.

Step 1. Cost drivers: capacity charge and energy charge

The first step in calculating grid rates is to decide how to divide the total cost among the possible cost drivers. Some proposals envisage a single capacity charge,[17] although in most cases an energy use charge, or more rarely, a fixed charge, is also implemented. The reason for the single capacity charge is the belief that network investment is solely determined by the peak demand that grids are expected to withstand. This consideration is inaccurate in both transmission and distribution, however, for energy use is also an important variable and at times a determinant for investment levels due to the need to minimise grid losses and to improve reliability. On the other hand, the network charge in many systems is purely volumetric (€/kWh), because of the lack of knowledge about the individual peak loads, given the absence of individual enhanced meters, and frequently of contracted capacities or any limits to the individual peak demands.

In those cases in which a capacity charge or an energy charge is not the sole charge, most of the approaches to dividing the cost between a capacity and an energy use charge are based on ad hoc criteria. One such criterion, found in several proposals, is to assume that, for a given group of consumers, the cost of the closest grid (typically, the grid for their voltage level) is recovered by the capacity charge, for grid design is regarded to follow local peak demand. The cost of the more distant grid (higher voltage levels), by contrast, is recovered by an energy use charge in the peak period, because this peak is assumed to be responsible for cost in that part of the grid [4, 31]. Parmesano [23] proposed recovering local distribution grid costs with a fixed charge, and the transmission and high voltage distribution grid costs with a capacity charge (or an energy use charge, where smart metres are installed).

Rodríguez et al. [27] introduced a method that adopts a detailed network-based quantitative approach to determine the capacity and energy use charges. This proposal consists of evaluating the weight of the contracted capacity and energy consumed in grid development, based on a reference network model (see Sect. 5.5.4). As we already know, reference network models are tools that deliver the optimal grid design using the parameters that determine grid planning in a given region, such as different consumption patterns and their location, terrain profile, costs of the components needed to meet the contract capacity and energy

[17] Logically, this capacity charge is designed taking customer location and time of use into consideration as cost drivers. In other words, this charge depends on the tariff period and voltage level and may depend as well on geographic location (urban or rural area).

demand requirements (such as cables and transformers), security of supply and service quality criteria. Evaluating the share that should be allocated to capacity and energy use charges is a two-stage process.

- Initially, the reference network model is used to design the optimal grid considering the contracted capacity only, i.e. not the energy demanded at each point of supply. This grid is designed to quality criterion N-1 (or any other, depending on the applicable regulations). The theoretical cost of a grid, if that were the sole design principle, can then be calculated, along with the unit cost per kW of contract capacity that should be allocated to each consumer.
- The optimal grid calculation is subsequently repeated, but considering both contract capacity and the energy demanded by consumers. This grid is also designed to quality criteria (such as the SAIDI and SAIFI[18] requisites in place in each area), considering the cost of losses as well. This analysis, especially for medium and high voltage grids, yields a grid whose cost is higher than found in the preceding phase. The difference between the costs of the grids resulting from the two phases is the amount that should be allocated to consumers' energy use charge.

Step 2. Definition of a grid model

As noted earlier, network costs are generally distributed on the grounds of voltage level. Since real life grids, especially distribution grids, are built with a wide variety of voltages, the use of some type of simplified grid is normally used to calculate tariffs. With this model, the cost of each voltage level can be allocated to the actors responsible for the cost in question (typically, downstream consumers connected at that level).

The use of a cascade grid model that envisages the existence of transformers between non-consecutive voltage levels is recommended for this purpose. With this grid layout, inter-voltage level flows and their respective shares can be calculated fairly simply, bearing in mind network losses, power plant delivery at each level and consumption figures. Figure 8.2 shows an example of a grid with the voltage levels proposed by Spain's National Energy Commission. A simpler example applied to the Libyan system is given in Reneses et al. [26]. Here, the possible presence of distributed generation at each voltage level (as shown in Fig. 8.2) has been ignored.

The following data associated with the grid model are needed to allocate the various voltage level costs to tariff periods and consumer categories:

- energy E_i and capacity P_i (be it contract, peak or mean, as discussed below) for each tariff period and category,
- losses for each voltage level and tariff period,
- power plant generation G_i delivered at each voltage level and in each tariff period,

[18] System Average Interruption Duration Index and System Average Interruption Frequency Index, respectively.

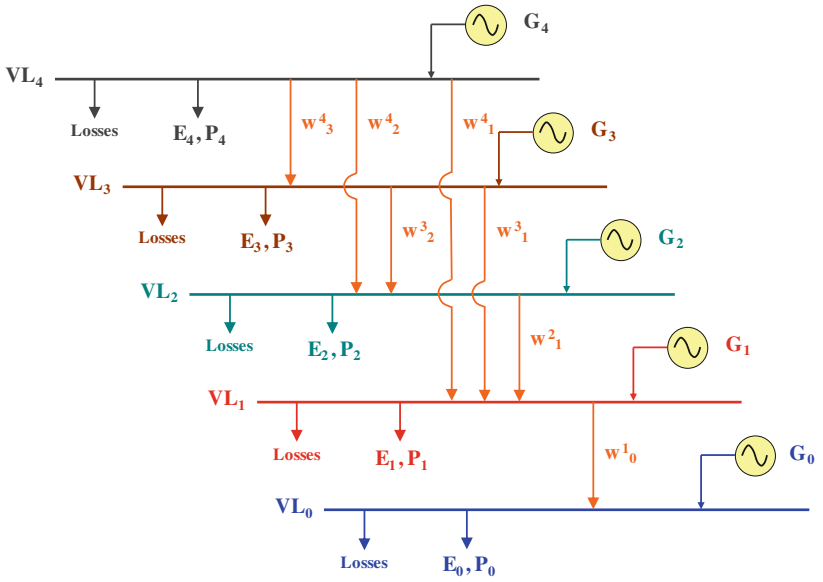


Fig. 8.2 Cascade grid model proposed by Spain's National Energy Commission

- flows w_j^i between the various voltage levels for each tariff period,
- voltage level load curves for each tariff period and
- load profiles for each customer category in each tariff period.

The grid model and associated data are used to distribute the cost of each voltage level among the responsible users, normally consumers connected at that level and the downstream voltage levels (assuming that power flows from higher to lower voltage levels).

Lastly, if different rates are to be instituted in different areas (different rates for each distributor, for instance), the grid model and associated data have to be broken down accordingly.

Step 3. Cost distribution across tariff periods

Once the decision has been made as to what portion of the cost is to be recovered under the capacity charge and what portion under the energy use charge, the cost must be distributed across the tariff periods. This is necessary because grid costs depend not only on the energy consumed or peak capacity, but on the time of day when power is consumed.

Standard practice is to distribute energy use costs in proportion to the energy consumed in each tariff period, for the part of network costs recovered with the energy use charge and attributable to each tariff period cannot be calculated a priori. This means that the unit charge is the same in all periods.

The distribution of capacity costs, which is more complex, has been discussed in considerable detail from the earliest accounting approach studies of the

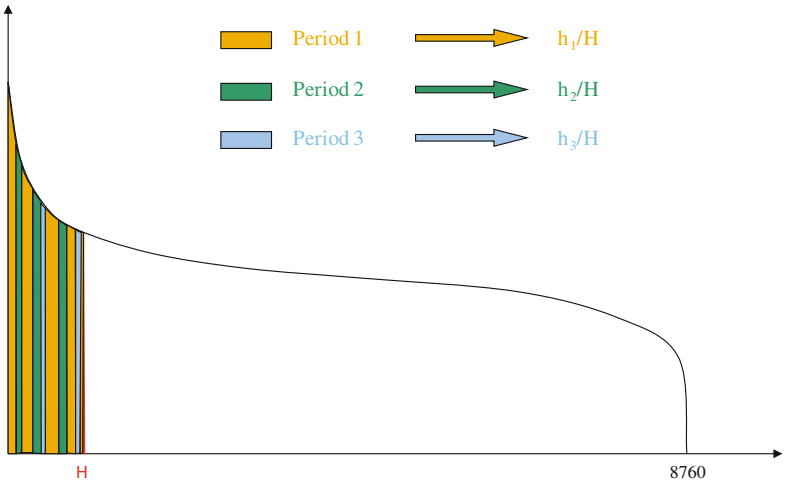


Fig. 8.3 Capacity cost distributed across tariff periods

problem. A number of very promising and complex methods put forward at the time are still used in many of today's methodologies. These methods arose out of the need to define hourly intervals and attempt to reflect the existence of times of day which, while inside the system peak periods, are actually not peaks from the standpoint of grid cost distribution (and vice versa, off-peak hours that are actually peak periods). Three reasons for this can be identified. First, periods are determined beforehand and demand is likely to shift or change sporadically; second, when defining tariff periods, clear and unchanging boundaries must be established (for instance, the weekday peak is defined to occur between 4.00 and 10.00 p.m.), even though this may vary from day to day; third, while grid costs are distributed separately for each voltage level (or even for each facility), the peak for a given voltage level (or area) may not concur with the overall system peak.

One of the most prominent methods is the PCP (probability of contribution to peak) method, that divides the capacity cost for each voltage level by the *H* hours of highest demand in that level. The capacity cost is divided on the grounds of the share of each tariff period in each level's peak demand, using the percentage of hours in each tariff period belonging to those H hours (see Fig. 8.3).

This method has been used and continues to be used (implicitly or explicitly) in a fair number of tariff schedule designs [26].

Another simpler approach consists of allocating the entire cost of a facility to its peak demand period [22]. Lastly, De-Oliveira-De Jesús et al. [9] proposed calculating hourly or load interval prices based on optimising social welfare, taking the demand response into consideration.

Step 4. Cost distribution across tariff categories

Once costs are allocated to cost drivers and distributed across tariff periods, the last step in their allocation to each cell on the tariff schedule is to divide them into the various consumer categories.

For the energy use charge, the natural and universal method is a pro rata division by the consumption recorded for each category within each tariff period. Hence, the cost of energy during a tariff period is split in proportion to the energy consumed by each tariff category in that period.

For the capacity charge, however, no obvious or generally accepted allocation method is in place. A number of approaches for allocating grid costs to consumers have been put forward in the literature, for this step has drawn more attention than the division across tariff periods. For distribution networks some authors have proposed to use the proxy of network utilisation, as it has been also used in transmission (see Sect. 6.4.2.2). These approaches include the MW-Mile method and its variations, Amp-Mile or MVA-Mile [18, 28]. In these methods, cost allocation entails geographic discrimination, whereby the costs in a given grid area (such as a substation and all downstream facilities) are distributed among downstream consumers in proportion to their consumption and the distance involved (in the example, the distance from the consumer to the substation). Zhong and Lo [33] proposed using a cross between the MW-Mile and the LTIC approaches.

In practice, a number of options have been used since the earliest analyses of the problem (accounting approach). The two most prominent are the coincident peak and the non-coincident peak methods. In the coincident peak method, cost is distributed by the value of the demand in each category during the peak for the respective tariff period. Under this method, the capacity charge of a group of consumers with a very high demand peak remains unaffected if that peak does not concur with the period peak [28]. The signals obtained are efficient for high voltage facilities used by all consumers (for they are designed for the system peak) but may not be for all the local grids that are designed to handle the peak capacity in a given area. Mutale et al. [22] proposed this exercise, but taking into account whether each facility is generation- or demand-governed (i.e. whether the flow is from high to low voltage or vice versa). Once that is determined, a peak flow charge is applied to each facility, so that positive or negative charges can be obtained for demand and distributed generation, depending on the direction of the flow. In the non-coincident peak method, costs are distributed in accordance with the peak for each category throughout the tariff period, regardless of whether or not the peak concurs with the period peak. This method provides better signals for local grids, but not for grids used by many consumers.

Rodríguez et al. [27] proposed classifying the consumers in a given voltage level by the actual cost they originate by using a reference network model and clustering techniques.

Other alternatives that have been applied include the use of the sum of non-coincident peaks (for each category, each consumer's peak is added to all the others'; this is unfair, for it fails to take simultaneity factors into account) or of the mean period demand (which fails to emit signals that would encourage peak reduction).

8.3.2 Methodology for Allocating Customer Management Costs

The second group of costs that should be allocated to calculate the access charge is the allowed customer management (or commercial) costs incurred by distributors. These are costs directly associated with technical management and the relationship with all connected customers. Most of these costs are related to metering and billing (where this task is performed by distributors) and customer support.

The solution nearly universally adopted to ensure efficient allocation of customer management costs is to establish a fixed charge per customer. This fixed, typically monthly, charge varies for each tariff category, for the costs associated with different customer categories may vary widely. Indeed, the management costs for a large industrial account may be several orders of magnitude higher than for a domestic consumer. As an alternative to the fixed charge per customer, Parmesano [23] proposed using a capacity charge or a surcharge on the first tranche of energy paid for by practically all users, so as not to distort the marginal signals that should be received by users.

8.3.3 Methodology for Allocating other Costs

The allocation of the rest of the regulated costs to the access charge is often an open issue. The most advisable general criterion is the causality principle to determine the most economically efficient allocation. That entails analysing the motivations underlying each cost item and, further thereto, to attempting to allocate them to the beneficiaries or actors responsible in accordance with each one's impact on the total cost. Below are a number of reflections on certain specific costs that are normally included in the access charge.

- Institutional costs. Since these costs normally account for a negligible proportion of the total regulated costs, applying a complex allocation criterion is not justified. The traditional solution consists of applying a uniform percentage of each customer's access charge or an energy use charge.
- Stranded costs. Allocation depends on their nature and volume, although in this case the causality principle cannot be applied, for these costs were incurred at some prior time. The solutions adopted usually include a fixed percentage of the bill, a fixed charge or, if the costs are not very high, an energy use surcharge.[19]
- Costs associated with environmental and energy diversity policies. The fraction of these costs defrayed by the electricity industry is normally allocated to an

[19] While the application of Ramsey prices would be optimal from the standpoint of economic efficiency, the arguments against their application are the practical difficulties estimating the sensitivities and the violation of the equity principle entailed.

energy use charge, which is usually the clearest cost driver. The target pursued by renewable energy incentives, for instance, is for this type of energy to attain a certain share of the generation mix. Hence, the more energy consumed, the greater the production and associated investment needed in renewable energy, and consequently the greater the associated cost. However, distortion of the actual marginal energy cost should be discouraged. And there are sound reasons to charge a significant fraction of the financial support to renewables to other non-electric energy consumptions, like heating or transportation, which also must participate in the efforts to reduce carbon emissions [5].

8.3.4 Determination of Final Tariffs

All the steps described in this section lead to the allocation of each cost item to the various cells on the tariff structure matrix. Application of the additivity principle yields the total cost allocated to each cell on the schedule, which is defined by each tariff category (consumer group), tariff period (hourly interval) and cost driver (capacity, energy or number of consumers). The final rates are calculated by simply dividing each of these total costs by the respective cost driver.

Those costs that only depend on the number of consumers must be divided by the expected number of consumers in each category, yielding a charge per consumer, and typically month. The cost of energy is divided by the volume of energy estimated for each consumer category and period, yielding a charge per kWh. Lastly, the capacity cost is divided by the estimated capacity that has been used to assign costs (contract capacity, coincident peak capacity, non-coincident peak capacity or mean capacity) to calculate a charge per kW.

For categories of consumers with metres able to provide the data needed to bill all these variables, this completes the ratemaking process. Most consumers do not have such metres today, however, and even those that have them may not be subject to such complex tariffs. Consequently, certain cells on the tariff schedule need to be grouped to determine the rates for such consumers, which become simpler.

If a tariff structure is designed with six rate periods, for instance, but domestic consumers only have two, the costs in the cells affected need to be combined. This operation is normally based on load profiles for the category in question, whereby the costs to be recovered in each original cell are calculated, the values found for the combined cells are added and the sum is divided by the sum of the cost drivers.

This type of adjustments to determine the final rates is likewise applicable to determine the integral tariff described in the following section.

There is still one more point to be made regarding the design of the final format of the tariffs to be applied to the end consumers. Not only the amount of network charges to be paid by the end consumer matters, also the format of the charge (€/kWh, €/kW or annual charge) and the process itself of computing the charges matter, as different formats and computation methods result in different economic

incentives for the end consumers. Once the amount to be charged to each end consumer is computed as described above, the format of the final charge could be chosen, so that any potential distortionary impacts are minimised. This issue was already discussed in Sect. 6.4.3 in the context of transmission charges and it will appear later in this chapter too in Sect. 8.5.7.

8.4 The Integral Tariff

In most countries, some manner of integral (or comprehensive) tariff is needed that includes all the cost items to be paid by regulated consumers. That would cover the access charge items as well as the energy use items, namely the cost or price (depending on the regulatory framework) of electricity production and any customer management (commercialisation) costs. Integral tariffs may be needed to cover either of the following two situations.

- Regulated integral tariffs, in which the cost of electricity production is one of the several regulated costs included in the rate paid by non-eligible consumers (all consumers in non-liberalised markets).
- In most liberalised electricity systems, eligible consumers, at least for some time, are allowed to opt to continue to buy electricity at a regulated tariff termed default tariff), provided by a retailer legally assigned this responsibility (and often related in some way to the distributor to which the consumer is connected). This tariff is discussed in detail in Chap. 9.

An integral tariff often co-exists with the access charge, when the electricity industry is in the process of liberalisation (i.e. in which not all consumers participate in the market yet) or even after full liberalisation (when a default tariff is available as a voluntary protection for some subset of consumers). Irrespective of the circumstances, the regulator must determine the methodology for allocating two further cost items to the access charge to establish the integral tariff, namely generation cost and customer management costs. The present section deals with this issue.

8.4.1 Methodology for Allocating Generation Costs

The first step for including generating costs in the tariff schedule is to quantify the total cost to be distributed among users. The way this sum is determined depends essentially on the regulatory model in place.

- In wholly liberalised markets, it can be obtained directly as the cost of purchasing electricity on that market. Criteria must nonetheless be established to define which price is to be used for this purpose: the spot price, the forward price

in any organised power exchange, some measure of the bilateral contract prices or any combination of the three. Some manner of incentive should also be instituted to encourage default retailers to purchase power efficiently.

- Where the market is liberalised but not all consumers have retail access, the price of energy on the wholesale market (spot, organised forward or bilateral contract) may also be used. In this case also, incentives should be established for the efficient purchase of power by retailers acting on behalf of regulated consumers.
- Lastly, where the market is not liberalised, the revenues allowed for generation are determined as the foreseeable total production cost (or allowed cost) for the period in which the tariffs are to be in effect. The following discussion assumes the existence of an organised electricity market. Section 8.4.1.2 focuses on the methodology to be implemented where no such market exists.

The earliest methodologies for allocating allowed generation costs were based on ad hoc criteria, along the lines of the accounting approach. One of the most widely used criteria for dividing costs between the capacity charge and the energy use charge was to allocate the variable costs of generation to energy component of the tariff (€/kWh) and the fixed costs (capital costs and generation independent operating and maintenance costs) to the capacity component (€/kW). In the specific case of hydroelectric plants, instead of assigning the total cost to capacity, part of the investment was regarded to be earmarked for storing fuel and consequently allocated to energy. The energy use costs were distributed more or less intuitively across tariff periods, based on the volume of demand in each period weighted by the cost of the fuel used. One of the methods used for capacity costs was the BIP (base, intermediate, peak) method, which classifies generating technologies in one of three categories: base (operating all year), intermediate and peak (only operating during the hours of highest yearly demand). The peak technology costs were allocated to the winter peak, the intermediate technologies equally between the winter and summer peaks, and the base costs were distributed equally over the winter and summer peaks and the off-peak period.

The greater availability of information and the enhancement of IT tools have logically allowed the development of more sophisticated and efficient methodologies. More specifically, in the 1970s and 1980s[20] a great deal of progress was made in establishing the price of generation based on the LTMC approach. Under that approach, marginal cost is calculated as the increase in total generation cost attributable to a sustained increase in demand over time, including both operating costs (essentially fuel and maintenance costs) and the cost of investment in new power plants [12, 21]. Normally, these LTMCs are calculated with a system planning and operation tool able to deliver costs broken down by time and even geographically. Typically LTMCs are used as a surrogate of STMCs and therefore they are allocated to the energy component of the tariff.

[20] Although the first proposals to address the problem arose around mid-century, they did not begin to gain popularity until several decades later.

The main advantage of the LTMC approach is that it sends consumers a stable long-term signal, encouraging long-term economic efficiency. It has two drawbacks, however. On the one hand, since the short-term signal (real costs at any given time) is lost, consumers do not perceive deviations from the usual pattern of prices, and in particular exceptional situations (such as drought, high demand or steam plant outages) with very high costs. Their concomitant failure to react to such situations threatens system security of supply and is not economically efficient. On the other hand, it does not generally ensure recovery of allowed generating costs (depending on the specific rules for remunerating generation, revenues may be higher or lower than allowed), necessitating adjustments that distort the signal.

In any event, the use of LTMCs to set generation tariffs began to wane with electricity industry liberalisation, although some authors continue to propose their implementation, particularly for transition economies [19, 30].

Industry liberalisation and the appearance of electric power wholesale markets led to the widespread use of STMCs as the optimal signal for remunerating the generation activity and establishing generation tariffs[21] [1, 29]. These marginal costs (or spot market prices) are transferred directly to consumers as an energy use charge, whereby the cost is distributed among them in proportion to the energy consumed at any given time. Moreover, these marginal costs can be used as a basis for both geographic (since spot prices can vary from one node to another) and by time of use (since prices are normally set by the hour) differentiation. The main advantage of using STMCs is that this approach sends consumers the most economically efficient signal, for account is taken of the cost of supplying power at any given time (and if the market is nodal or area based, in any given place). The scheme therefore reflects the situation actually prevailing: demand, hydroelectric reserves, fuel costs, outages and so on. Moreover, in systems with a wholesale market it is the most transparent signal and the simplest to implement.

Some people argue that a drawback to STMCs is their volatility. It is true that the wholesale market price, which represents the real cost of energy at any given time, may vary abruptly and reach very high or low values in certain circumstances.[22] Other people claim that today most users are not prepared to react to real-time pricing (RTP). In reality all consumers—with advanced meters on not—are equally exposed to these volatile short-term energy prices and pay them. The only difference is that those with advanced metres have the information about real-time prices and may react to them, while consumers with simple metres are subject to the same prices but do not have the possibility of doing anything about it.[23]

---

[21] See Chap. 6 on Electricity Generation and Wholesale Markets for further details.

[22] Consumers could, of course, conclude agreements to protect themselves from volatility, while still being responsive to real time prices if the right format of hedging contracts is chosen, see Sect. 9.3.3.

[23] It is assumed here that, at the end of the year, there is some sort of adjustment in the tariffs for next year, so that any deviation between the estimated tariff for those consumers not directly exposed to the short-term prices and the actual prices during the year is accounted for. See Sect. 8.5.1.

Improved information and communication technologies afford the opportunity to conduct large-scale experiments, and a substantial number of articles have attempted to analyse the effect of applying tariffs based on STMCs (see, for instance, Borenstein [8]). The unrelenting trend is for an increasing number of users to receive some manner of time-of-use signal, in response to the pursuit of economic efficiency and improvements in demand-side management. For low voltage consumers, these signals are usually time-of-use (TOU) rates, which are calculated ex ante with electricity system operating models that estimate the expected STMC values.[24] One factor that should be borne in mind is that while more stable price signals lower the consumer exposure to hourly STMCs, these more stable signals are nonetheless less efficient from the standpoint of system operation. In some countries, regulators require default suppliers to conclude agreements for part of the energy they need to buy for their customers (for further detail, see the chapter on the retail business) to reduce the exposure to the short-term prices, as any load serving entity would do to hedge against uncertain future short-term prices.

8.4.1.1 Capacity Payment

The regulator may wish to ensure a certain reliability level for the power system (in simple terms this requirement may be visualised as a constraint to install a certain minimum quantity of firm generation capacity above the estimated peak demand). Under a market-oriented regulatory framework, this reserve margin can be attained by means of some regulatory mechanism that translates this requirement into some sort of capacity remuneration for generation investments. Since such mechanisms vary much from one electricity system to another, their reflection in rates should take the design criteria into consideration, see Chap. 12.

As a general rule, the charge to cover the cost of the adopted capacity instrument is applicable to all consumers, regardless of the supply scheme and in keeping with their responsibility in generation investment. Since the concern is ensuring generation availability in critical periods (peak system demand, typically), the most suitable variable for allocating this cost should be the peak capacity of each consumer that is concurrent with the system peak.[25] In practice, the billing variables are usually contracted capacity or actual power consumption (kW) during the peak period. This charge should logically be part of the capacity (€/kW) component of the consumers′ tariff.

---

[24] The adjustments required once the market price is known are addressed in Sect. 8.5.1.

[25] In the presence of large volumes of intermittent generation (wind or solar PV), or with a significant presence of hydro storage, the power system is more stressed when there is scarce wind and sun and there is little water in the reservoirs, even if the demand is not at the annual peak. Short-term market prices above a prescribed threshold or the exhaustion of operating reserves are more reliable indicators of power system stress than the pure demand level.

In approaches that propose using a planning model to calculate LTMCs, the capacity payment is included therein, if the model is designed to accommodate the reserve restriction. For that reason, LTMCs must be divided into an energy use charge and a capacity charge.

8.4.1.2 Revenue Reconciliation

This section only applies to power systems under traditional cost-of-service regulation. As noted earlier, where the market is not liberalised, the remuneration of generation is determined as the allowed cost for the period in which the rates are to be in effect. The allowed cost needs not, of course, concur with the revenues accruing from the (short or long term) marginal cost rate design and (if applicable) the capacity payment. In such circumstances, the calculated rate must be adjusted by means of a supplementary charge; this is known as revenue reconciliation.

It is important to note that, in power systems under traditional regulation, it is possible to send to consumers the same efficient short-term signals that consumers may receive under competitive market conditions. As we know from Chaps. 2 and 7, the spot prices of electricity in perfectly competitive markets coincide with the STMCs of generation. In the design of integral tariffs both the short-term prices and the STMCs have to be estimated ex ante, although adjustments may be applied later, see Sect. 8.5.1. And, with advanced metering, in both cases it is also possible to apply real-time pricing. The main difference is that, under market conditions, the market price is all that generators receive and consumers must pay (except, when it applies, any capacity payment). With cost-of-service regulation the generators must recover all their allowed costs, and this amount may differ from what consumers would pay under marginal cost-based tariffs. As said before, making generators whole under cost-of-service regulation with marginal cost-based tariffs is the role of revenue reconciliation.

This supplementary charge must be allocated in a way that distorts the marginal cost signals as little as possible. To this end, second best methods, such as Ramsey prices, are sometimes implemented, whereby these costs are allocated to consumers in accordance with their elasticity. As noted in Sect. 8.3, however, these methods are difficult to apply in practice and may violate the equity principle. For that reason, multiplier methods (coefficients that maintain tariff proportionality across periods) or additive methods (coefficients that maintain the absolute value of differences) are often used [2, 23].

As in the case of market-based regulation, the STMCs are wholly allocated to the energy component of the regulated tariff.

8.4.2 Methodology for Allocating Customer Management Costs

In addition to generation costs, retailers that purchase power for regulated rate users also incur customer management costs. As indicated in Sect. 8.3.2, this type of costs is nearly universally allocated by means of a fixed charge per consumer. These charges naturally vary depending on the tariff category, for each customer group is responsible for different amounts of management costs.

If the rate designed for a certain category is intended as a default tariff for a small number of consumers only, high customer management costs can be established to encourage consumers to participate in the market.

8.5 Miscellaneous Issues

This section briefly reviews a series of subjects that, while not constituting the core of tariff design, have a bearing on final rates.

8.5.1 Tariff Adjustment

As the reader will have realised, ratemaking often entails using parameter forecasts for tariff periods, including items such as the energy consumed by each category, peak demand or number of consumers in a given category. As a result, and particularly, to ensure that all allowed costs or remuneration based on actual energy prices are recovered, tariff adjustment mechanisms must be designed to revise calculations and offset any deviations once the actual values of these parameters can be determined.

Such adjustments may be made in different timeframes, depending on the item to be revised. Most are made yearly, when calculating the rates for the following year. In some cases, however, semi-annual, quarterly or even monthly revisions are made, depending on the regulations in effect. One of the most common adjustments required is the energy cost pass-through charge applied to regulated consumers.[26] Lastly, special adjustments may be implemented in the event of substantial deviations with an impact on costs that is too heavy to be postponed.

[26] In California, in the summer of 2000, two retailers-distributors went bankrupt when wholesale market prices rose abruptly and no mechanisms had been established to transfer these rises to end consumers.

8.5.2 Connection Charges

Grid connection charges have been already introduced in Chaps. 5 (distribution, Sect. 5.2.3) and 6 (transmission, Sect. 6.5.1.1). These are one-off charges for system connection. Two main philosophies are in place with respect to connection charge design: so-called deep cost and shallow cost tariffs.

In deep cost tariffs, the new consumer is charged for the cost of his/her own service connection and of all the upstream grid reinforcements required to supply the contracted capacity. If, for instance, that entails enlarging a substation, reinforcing a line or changing a relay system, those items are included in the connection charge. This solution carries a very strong location signal for new customers, for connection costs may constitute a very high barrier for connecting to a saturated system. If incorrectly designed, however, it may discriminate inordinately among consumers connected at one and the same part of the grid, depending on when they connected. For instance, once the grid has been reinforced (bearing in mind that reinforcement is discrete), a consumer with exactly the same characteristics as the preceding user may connect to the same point without having to pay for the reinforcement already paid for by the latter.

With shallow cost tariffs, the customer pays the service and grid connection costs only. All other reinforcements are regarded to form part of the grid costs recovered under the access charge. This rate is not discriminatory, although the location signal is not as strong as in the deep cost approach. An alternative even more beneficial for users is the super shallow cost-based connection charge, which includes only the equipment up to the connection line, but not the line itself.

Intermediate models are also possible, such as the so-called shallowish charges, which depend on customer size, location, loads in the surrounding area, connection voltage and so on [32].

In any event, the connection charge chosen and the access charge must be consistent to ensure that all the costs incurred by the distributor are acknowledged, and acknowledged only once.

8.5.3 Utilisation Factors

Even without advanced metres, it is possible to distinguish among small and medium size consumers connected at a given voltage level on the basis of their amount of energy consumption or, more precisely, their utilisation factor.

The utilisation factor (UF) of a consumer is defined as its total consumed energy divided by its contracted capacity. The maximum possible value of UF is 8,760 (or 1.0, if we normalise all the utilisation factors by dividing them by 8,760), corresponding to consumers whose demands, at every hour of the year, are equal to their contracted capacities.

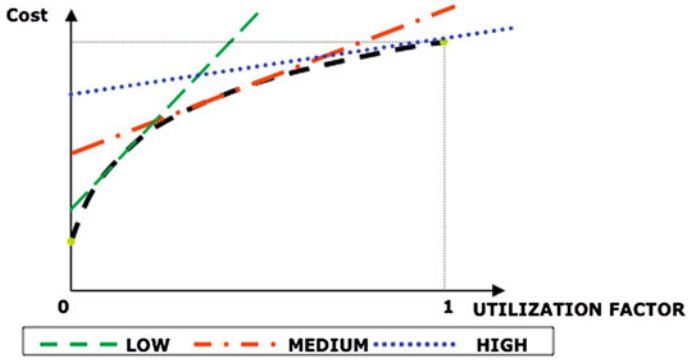


Fig. 8.4 Network cost functions in terms of the value of the utilisation factor

If the network charges are accurately computed for many real consumers individually over the entire range of values of utilisation factors, it should be possible to fit a curve to these experimental data. Of course, this curve should be verified and adjusted for each situation, but typically it should have the shape that is shown in Fig. 8.4. Since an experimental curve cannot be easily used to define tariffs with a simple format, the curve in Fig. 8.4 has been approximated by three straight lines: one fits well the network charges for low UF consumers, other for medium UF and a third one for high UF consumers. These straight lines are now used to define three sets of simple tariffs, since they only have two components:

- A capacity component (in €/kW, applied to the contracted capacity or the estimated peak demand of the considered consumer): This is the value of the straight line at its intersection with the vertical axis: the incurred cost even if the energy consumed is zero, just for being connected.
- An energy component (in €/kWh, applied to the consumed energy, which is proportional to the utilisation factor).

Note that consumers with a low utilisation factor, who typically consume preferentially at peak time, have a lower capacity charge, while they have a higher energy charge than other consumers with higher values of UF. The opposite applies to consumers with a high utilisation factor. The explanation should be clear: Even though low UF network users concentrate their consumption at peak times, their coincidence factor with peak demand necessarily has to be inferior to that of consumers with UF close to 1.0, since they consume at almost all times. On the other hand, the energy consumed by low UF consumers is, as an average, more expensive than the energy consumed by high UF consumers, which basically corresponds to the average energy price over the entire year.

The set of tariffs that have been defined in this way has an interesting property (inferred from inspection of Fig. 8.4): If the consumers know their expected utilisation factors for next year and they are free to choose whatever of the three

tariffs, they will chose the one corresponding to their true UF values, since this is the one that minimises their payments.

Utilisation factors are frequently used in practice in the design of tariffs for consumers connected at low and medium voltages. It allows the tariffs to be better adapted to the load profiles of the individual consumers without having to resort to advanced metres.

8.5.4 Reactive Power Rates

Reactive energy affects energy losses and voltage regulation, two of the keys to satisfactory system operation. Reactive power impacts grid dimensioning. Most reactive power is consumed in users' facilities. Consumers can and should, then, participate in controlling this power, with a view to maintaining voltage levels and minimising system losses. Signals should consequently be sent to consumers in the form of a specific charge.

When designing reactive power charges, the signal must be adapted to consumer typology. The keys to good reactive charge design are, on the one hand, metre specifications and on the other, consumers' capacity to correct reactive power. Reactive charges are often not applied to low voltage domestic consumers, in light of their scant ability to vary their behaviour.

For large-scale consumers, however, reactive power metres are usually installed, to measure consumption only or generation and consumption. Some metres also feature time of day recording. Charges are commonly established in terms of the consumption cosine or tangent when electric power is expressed as a phasor, although this parameter oversimplifies the reactive power issue, particularly for large accounts. Charges that penalise reactive power consumption in peak periods and its generation during off-peak periods are often implemented [4].

Some authors have attempted to develop specific methodologies to design reactive energy tariffs. Foremost, among these methods are the adaptation of the MW-Mile approach [18, 28], and the use of LTICs on the grounds of the reinforcements required due to voltage limits at nodes (Li and Matlose 2008).

8.5.5 Subsidies

As maintained throughout this chapter, from the standpoint of economic efficiency, good tariff design ensures that the tariffs charged reflect the costs incurred. This induces efficient system use, guarantees long-term viability and investment and avoids cross-subsidies among consumers in different rate categories. Certain communities of users (low income consumers), however, may not be able to pay the system costs they originate. Subsidies may be in order to enable them to access

the service. From a social-economic standpoint, these subsidies are justified by the fact that for certain basic needs, electric power is regarded to be a universal right.

Two forms of subsidies are implemented in standard international practice.

- Subsidies integrated in the tariff, which is often progressive, with higher consumption brackets paying a higher price,[27] and designed, so that basic needs (lowest bracket) can be covered at a very low price. This practice is generally unadvisable from the standpoint of economic efficiency, since the same electric energy is simultaneously subsidised for some consumers and charged extra for some others.
- Tariff independent, explicit subsidies, identifying the beneficiary communities and establishing direct payments to cover the cost of their electricity. The main advantage is that they do not distort the economic signals emitted by rates and their primary drawback is the pitfalls involved in identifying the groups involved.[28]

The funds needed to cover the subsidies may be sourced externally (public subsidies) or internally (recovered as a regulated charge in the electricity tariff).

Another type of subsidy that is unavoidable in ratemaking (in the absence of individual hourly metering) is the cross-subsidy among consumers in the same category. These subsidies are inherent in cost distribution based on behavioural similarities among consumers in a given tariff group. Regardless of how refined the rate categories are, consumption patterns will always vary within each one. Consumers with load profiles less coincident with hourly price profiles subsidise consumers whose demand is very aligned with the periods of high prices.

8.5.6 Distributed Generation[29]

The conventional generation plants (large thermal, hydroelectric or nuclear plants) are normally connected to the transmission grid, driving energy flows from higher to lower voltage levels to successively feed demand at each level. For a number of years and due to a variety of reasons, a host of different generation technologies have been appearing, which share two features: small size compared to conventional plants and connection to the distribution rather than the transmission grid.

In some electricity systems and more specifically in certain areas in some systems, distributed generation (DG) is changing the grid planning and operating paradigm, and giving rise to flows that circulate from lower to higher voltage levels. The regulation of such facilities is a highly topical and open issue whose impact on tariff design should also be studied.

---

[27] Another aim of this practice is demand-side management to further energy savings.

[28] See Chap. 9 for further details in this regard.

[29] The reader is advised to review Sect. 5.7, for an introduction to this topic.

At the present moment, there is no generally accepted approach to determine network charges for DG and it can be considered as an open research topic. As indicated in Sect. 5.7 (distribution), the reasonable way to proceed is to analyse the cost impact on each grid of the penetration of DG with a network reference model, until the experiences obtained may allow the development of general purpose rules.

Another important factor that has been the object of considerable debate is whether DG connection charges should be based on super shallow costs, shallow costs or deep costs.[30] This is a subject for research that will have to be addressed in depth in the years to come, as DG acquires greater importance in electricity systems.

Another open topic in tariff design, already announced in Sect. 5.7, is the treatment to be given to distributed generation at residential level, i.e. distributed generation such as microcogenerators or roof-top solar panels, which are connected at the same node in the distribution grid as the residential demand. If separate metres are used for demand and generation, each one could be properly charged with its specific tariff. If a single metre is used jointly for both (this is usually termed "net metering"), a special tariff design could be developed for this new category of network user, which sometimes behaves as a load and other times as a generator. The problem appears when the distribution charge that is applied to the net metred combination of generation and demand is a purely volumetric one (€/kWh) or almost. Why a problem? Because the local generation avoids the payment of network charges to any existing demand and it is therefore being unduly subsidised. Some credit might be given to local generation for reducing network use, but not as much as 100 % reduction in the network charges to local demand per kWh of energy produced.

8.5.7 The format of the tariff

The previous discussion makes us think about a fundamental issue in tariff design, but one that the authors have never seen discussed in the specialized literature: the format of the tariff matters and, if necessary, it can be made independent from the procedure that has been used to determine the monetary value of whatever charge an agent must pay.

This applies to generators as well as to consumers, since both should be subject to network charges. This issue was already mentioned in Sect. 8.3.4 and in Chap. 6, and it will be explained here using as an example, the network charges applied to a generator located at a node n in a transmission network. The logic of cost causality, as described in this chapter, when applied to a specific generator, must result in some structure of charges, with some amount associated to energy (€/kWh), another

[30] Detailed information on connection charges for distributed generation can be found in the European ELEP (http://www.elep.net) and Green-Net (http://www.greennet-europe.org) projects.

to some measure of capacity (€/kW) and perhaps also some annual quantity as a lump sum (€/yr). The point is that we might not want to *apply the charges* following this same format. For instance, note that an energy charge (€/kWh) would be internalized by the generator in its bids into the market, therefore, distorting its competitive market behaviour in the short-term, given that the transmission charge is a long-term signal (not to be mistaken with nodal energy prices). In conclusion, the structure that results from the procedure that has been followed to compute a tariff may be taken as a valuable indication of the format of the final tariff, but other issues must also be considered. Only the total amount to be charged must be maintained. This concept will grow in importance in the complex design of network tariffs for "prosumers", i.e. entities that will be at the same time producers and consumers of electricity.

Finally, an additional point is that per unit (€/kWh, €/kW or €) network tariffs—for consumers or generators—should not be attached—as it is customary—to a given node, because network tariffs depend on the pattern of behaviour of the generator or consumer that is connected to that node. Certainly a nuclear plant, a peaking unit or a wind farm connected at the same node n should not pay the same €/kW network charge. This leads us to think that there is no good reason to maintain the fiction of nodal network charges that apply to any agent located at a given location.

8.6 Summary

This chapter addresses tariff design, covering the fundamental theory underlying the approaches presented in place as well as their practical implications. The conclusions drawn are listed below.

- Electricity tariff design is a highly complex task, theoretically and practically speaking, which has not always received the attention it merits.
- The fundamental ratemaking principles to be borne in mind for tariff design are sufficiency (cost recovery), economic efficiency and equity. Since these principles normally clash, tariff design consists of reaching a compromise among them.
- Other principles to be taken into consideration in ratemaking include transparency, additivity, simplicity, stability and consistency with the regulatory framework.
- The tariff design procedure consists of three main steps: (1) choice of the remuneration scheme and level of remuneration for each one of the activities needed for electricity supply (as presented in the preceding chapters); (2) definition of the tariff structure applicable to end consumers and (3) allocation of the allowed costs to the structure.
- The design of the tariff structure must be based on cost drivers (billing variables) and tariff periods and categories (consumer groups).

- The earliest attempts at cost allocation conformed what is now known as the accounting approach, based on business accounting. Later, in the mid-twentieth century, marginal pricing concepts began to be applied to ratemaking, in pursuit of economic efficiency. This gave rise to tariff designs initially based on LTMCs. The difficulty involved in applying marginal pricing to network costs eventually led to the advent of tariff design more strictly based on the principle of cost causality (which borrowed from both the accounting and the marginal pricing approaches).
- The cost items accounted for in tariff design can be classified into two categories: access charges (to be paid by all grid users, regardless of whether the regulatory framework is traditional or market oriented) and the remaining charges or prices, corresponding to the generation and retail activities, which can be performed either under traditional or market regulation, and complete the access charge to result (under traditional regulation only) in the regulated integral rate.
- The access charge can in turn be divided into grid costs (normally accounting for the major share of this charge), distributors' management costs and other regulated costs. The integral tariff is topped off with generation (energy purchase) and customer management costs.
- Network cost allocation has been and continues to be the object of a sizeable number of proposals, which in recent years have been narrowing into a fundamental tenet: cost allocation must be based on cost causality. This principle somehow inspired the early attempts using LTMCs and it is also behind the most recent approaches, which try to exploit the fact that network planning is caused by those agents whose increased demand or generation has required the network to expand. Therefore, at the heart of the planning process must lay the justification for cost allocation.
- Customer management costs (incurred by distributors and in the integral rate) are recovered practically universally by means of a fixed charge that varies from one rate category to another.
- The most efficient methodology to recover energy purchase (generation) costs is based on marginal pricing principles. STMCs plus some adjustment elements for revenue reconciliation (under traditional regulation) or spot market prices (when wholesale competitive markets exist) provide the efficient economic signals to be used in the design of integral regulated tariffs or to be used by retailers in their competitive offers to end consumers.

References

1. Andersson R (1984) Electricity tariffs in Sweden. Energy Econ 6:122–130
2. Apolinário I, Felizardo N, Leite Garcia A, Oliveira P, Trindade A, Verdelho P (2006a) Additive tariffs in the electricity sector. Power Engineering Society General Meeting, IEEE, Montreal

3. Apolinário I, Felizardo N, Leite Garcia A, Oliveira P, Trindade A, Verdelho P (2006b) Determination of time-of-day schedules in the Portuguese electric sector. Power Engineering Society General Meeting, IEEE, Montreal
4. Apolinário I, Correia de Barros C, Coutinho H, Ferreira L, Madeira B, Oliveira P, Trindade A, Verdelho P (2009) Efficient pricing on distribution network tariffs. In: Proceedings of the 20th international conference on electricity distribution, Prague, Czech Republic
5. Batlle C (2011) A method for allocating renewable energy source subsidies among final energy costumers. Energy Policy 39:2586–2595
6. Berg SV, Tschirhart J (1989) Natural monopoly regulation: principles and practice. Cambridge University Press, Cambridge
7. Boiteux M, Stasi P (1964) The determination of costs of expansion of an interconnected system of production and distribution of electricity. In: Nelson J (ed) Marginal cost pricing in practice. Prentice Hall, Englewood Cliffs
8. Borenstein S (2005) The long-run efficiency of real-time electricity pricing. Center for the Study of Energy Markets, University of California Energy Institute
9. De Oliveira-De Jesús PM, Ponce de Leão MT, Yusta JM, Khodr HM, Urdaneta AJ (2005) Uniform marginal pricing for the remuneration of distribution networks. IEEE Trans Power Syst 20:1302–1310
10. Green R, Rodriguez Pardina M (1999) Resetting price controls for privatized utilities. A manual for regulators. The World Bank, Washington
11. ETSO (European Transmission System Operators) (2009) ETSO Overview of transmission tariffs in Europe: synthesis 2008. https://www.entsoe.eu/fileadmin/user_upload/_library/publications/etso/tariffs/Final_Synthesis_2008_final.pdf
12. Joskow PL (1976) Contributions to the theory of marginal cost pricing. Bell J Econ 7:197–206
13. Kahn AE (1988) The economics of regulation: principles and institutions. MIT Press, Cambridge
14. Lévêque F (2003) Transport pricing of electricity networks. Kluwer Academic Publishers, Dordrecht
15. Li F (2007) Long-run marginal cost pricing based on network spared capacity. IEEE Trans Power Syst 22:885–886
16. Li F, Tolley DL (2007) Long-run incremental cost pricing based on unused capacity. IEEE Trans Power Syst 22:1683–1689
17. Li F, Matlotse E (2008) Long-run incremental cost pricing based on nodal voltage spare capacity. Power and Energy Society General Meeting, IEEE, Pittsburgh
18. Li F, Padhy NP, Wang J, Kuri B (2008) Cost-benefit reflective distribution charging methodology. IEEE Trans Power Syst 23:58–63
19. Malik AS, Al-Zubeidi S (2006) Electricity tariffs based on long-run marginal costs for central grid system of Oman. Energy 31:1703–1714
20. Marangon Lima JW, Noronha JCC, Arango H, Steele dos Santos PE (2002) Distribution pricing based on yardstick regulation. IEEE Trans Power Syst 17:198–204
21. Munasinghe M (1981) Principles of modern electricity pricing. Proc IEEE 69:332–348
22. Mutale J, Strbac G, Pudjianto D (2007) Methodology for cost reflective pricing of distribution networks with distributed generation. Power Engineering Society General Meeting, IEEE, Tampa
23. Parmesano H (2003) Rate design is the No. 1 energy efficiency tool. Electr J 20:18–25
24. Pérez-Arriaga IJ, Rubio FJ, Puerta JF, Arceluz J, Marín J (1995) Marginal pricing of transmission services: an analysis of cost recovery. IEEE Trans Power Syst 10:546–553
25. Ponce de Leão MT, Saraiva JT (2003) Solving the revenue reconciliation problem of distribution network providers using long-term marginal prices. IEEE Trans Power Syst 18:339–345
26. Reneses J, Gómez T, Rivier J, Angarita JL (2011) Electricity tariff design for transition economies. Application to the Libyan power system. Energy Econ 33:33–43

27. Rodríguez MP, Pérez-Arriaga JI, Rivier J, Peco J (2008) Distribution network tariffs: a closed question? Energy Policy 36:1712–1725
28. Sotkiewicz PM, Vignolo M (2003) Towards a cost causation-based tariff for distribution networks with DG. IEEE Trans Power Syst 22:1051–1060
29. Schweppe FC, Caramanis MC, Tabors RD, Bohn RE (1988) Spot pricing of electricity. Kluwer Academic Publishers, Boston
30. Vikitsert T (1995) Electricity tariffs in Thailand: structure, objectives and impact on system load. Thai J Dev Adm 35:205–226
31. Williams P, Strbac G (2001) Costing and pricing of electricity distribution services. Power Eng J 15:125–136
32. Williams P, Andrews S (2002) Distribution network connection: charging principles and options. DTI, London
33. Zhong X, Lo KL (2008) Development of distribution networks pricing with distributed generation. In: Proceeding of the third international conference on Electric Utility Deregulation and Restructuring and Power Technologies—DRPT 2008. Nanjing, China

Chapter 9
Electricity Retailing

Carlos Batlle

Every product has some element of service, and every service some element of product.
Aubrey Wilson (UK marketing consultant)

"Electricity" (electric energy) is a homogeneous commodity. In principle, the nature of this commodity is unaffected by scale, i.e., the same product, namely electromagnetic energy carried over a network, is delivered to large, energy-intensive factories and small consumers in other parts of the system.

However, this homogeneity in the physical product is deceiving when this commodity is examined in more detail. Every hour, electric energy is produced at different plants, whose short- and long-term costs are very different. Once produced, the energy flows to the consumer through a dense network whose costs also vary widely depending on the location of each connection: a direct connection to a high-voltage network (assuming that the consumer uses that network only) is not the same as a low-voltage connection in a central district of a large city. Thus, the cost of supplying a megawatt hour fluctuates significantly, depending on consumer location and the specific time when consumption takes place. In other words, the same physical product is delivered at different times and places, incurring different costs and this must obviously lead to different prices. Moreover, unlike demand in many other commodity markets, electric power consumers may not yet be receiving the right price signals, because of regulatory and hardware shortcomings. And the price elasticity of demand (i.e., how much the electric power supply is worth to consumers at any given time) is still poorly understood.

Under the new regulatory paradigm, electricity has become a product comparable in many respects to other consumer goods such as cars, for instance.

- Just as customers may purchase a more or less reliable vehicle, electric power customers should be able to sign on for a more or less robust power supply (for

C. Batlle (✉)
Institute for Research in Technology, Comillas Pontifical University,
Sta. Cruz de Marcenado 26, 28015 Madrid, Spain
e-mail: Carlos.Batlle@iit.upcomillas.es

C. Batlle
MIT Energy Initiative, MIT and with the Florence School of Regulation,
European University Institute, Florence, Italy

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_9,

instance, consumers should be able to pay less if they consent to having their supply interrupted on occasion).
- Just as automobile manufacturers offer less-polluting vehicles, new electricity retail companies should allow consumers to choose to pay more in exchange for electricity produced using cleaner technologies.
- Just as automobile purchases now often come with incentives such as discounts on insurance premiums, electric power purchases should come with analogous benefits such as a lower price for gas or advice on how to save on the energy bill by using power more efficiently.

Ultimately, while some users need a large vehicle for their businesses designed to run for many miles at a low cost, others can make do with a vehicle to go for short weekend outings. Similarly, some users (e.g., supermarkets) consume a large amount of energy during working hours, while others consume small amounts and at other times, i.e., nights and weekends.

Clearly, then, not all customers are equal or want the same product. This circumstance has given rise to a new kind of business, with a new breed of agents: retailers, who act as the intermediaries between producers and end customers. This new business model is based on exploiting flexibility and economies of scale, e.g., purchasing energy in the wholesale market and selling it to customers according to a variety of consumption patterns, or catering to certain types of consumers with specific concerns and needs.

9.1 Introduction

9.1.1 Unbundling of the Electricity Tariff for Consumers

To calculate the "right" price for consumption and consumer class is a complex task, necessarily subject to assumptions and approximations, as shown in the preceding chapter. The term "right" encompasses many considerations that go far beyond the mere cost of electricity, as many higher order factors often come into play, such as environmental costs, social criteria or impact on the country's competitiveness. Therefore, regulations and regulators commonly use a variety of adjectives to qualify the prices expected from sound business operation: right, fair, reasonable, suitable, equitable and efficient are a few examples.

Prior to the restructuring and liberalisation of the power industry, the activity now known as "retailing" (also as "supply" and sometimes even "commercialisation"[1]) was a part of the chain of activities performed by vertically integrated utilities and was, therefore, a public service or a regulated monopoly.

[1] Hereafter, both terms, as well as retailer and supplier, will be used interchangeably as synonyms.

The government or the appropriate regulatory authority decided who had to pay what, as well as how and when, typically under a cost-of-service rule, although the design of the individual end-use electricity tariffs was often more the result of politics, lobbying and negotiation than the application of sound economic principles. A clear exception to the lack of a rigorous method is the long ratemaking tradition in the US, with an approach that, with many small variants, has been followed for many decades by the Public Utility Commissions of the different states [31, 35]. Other countries have also for a long time applied systematic approaches to ratemaking that are based on economic concepts, like the Green Tariff in France since mid-past century. Regulated tariffs for consumers (defined by the regulator) were intended to pay for three main types of costs:

- The cost of electricity networks (high and low voltage, transmission and distribution).
- The cost of energy purchases (not only the cost of energy in the market or ancillary services or capacity mechanisms, but also the costs associated with the retailing service itself: marketing, metering and billing).
- Other costs arising from decisions taken by the regulatory authorities: subsidies for research activities, renewable energy sources, energy efficiency programmes or support for domestic fuels; the nature and existence of these costs is typically very country-specific.

One of the main changes resulting from the full restructuring of electricity systems affects the second item, i.e., the cost of energy supply: as described in Chap. 7, on the generation side, the cost of energy is determined by the market sequence (bilateral, pool or balancing, where the "cost" becomes the "price"), whereas on the demand side, a new activity has appeared, liable to be governed by free-market criteria.

The complete restructuring and liberalisation of the power industry, as described in Chap. 3, was expected to profoundly affect the activity of competitive retailing, which was an outgrowth of the regulatory reform. However, power sector restructuring does not necessarily entail the liberalisation of retailing. As discussed later in this chapter, in some countries where the wholesale market, and therefore the generation business, have been liberalised, the liberalisation of consumer supply has been postponed or limited to large consumers only.[2] In other electricity systems, supply has been gradually liberalised as both the wholesale and retail markets have matured.

In countries undertaking full reform of the electric power system, however, such as the European Union Member States where retail has been unbundled, the

[2] This may be attributed to a number of factors. One is the consequence of the technical and organisational difficulties involved in implementing a fully liberalised retail market for all consumers. Another is the belief that the benefits of opening up this market to small consumers might be considerably less significant than the benefits of liberalising generation, which accounts for the bulk of industry costs and investment.

cost of retailing for consumers buying electricity on the retail or wholesale markets is internalised in the trading process and no longer subject to regulation.

Therefore, in the liberalised context, the task of determining the rate charged to pay the costs of the retail activities is no longer the responsibility of the regulator. The regulator must calculate the network-access tariffs (which cover the items in bullets one and three, above: networks and other regulated charges) and establish the conditions to ensure that the costs of supply (prices) are set competitively.

Because of this change in paradigm, in which electricity is no longer seen as the product of a single firm, but as the final outcome of the efforts of several firms, each one in charge of a distinct activity of the liberalised electricity value chain, with transmission and distribution remaining as separate regulated businesses, retailing has been able to grow into a significant market activity. Retailing is the link between wholesale market transactions and end customers, and facilitates interaction between the two. A key ingredient of this model is the separation between the activities performed by retailers (who sell electricity) and distributors (who provide the infrastructure to physically transport and deliver power).[3]

The most prominent features of retailing are discussed below, with an emphasis on the elements characteristic of competitive supply, which generally elicit the greatest interest and entail the greatest complexity.

9.1.2 Objectives of Retail Business Liberalisation

Broadly speaking, electricity supply is a process in which an intermediary (the retailer or supplier) is responsible for buying energy for a consumer in return for remuneration. In addition, this process may entail the provision of other associated services, such as advice on how to use and purchase electricity, or implementation of energy efficiency and saving measures.

The main objective pursued with the liberalisation of electricity retail is to improve the overall efficiency of the electricity business by providing intermediation services. These services may adopt a variety of forms, resulting in a diversity of agents known by different names, depending on their primary function, although there is some overlap: brokers, traders and retailers. A *broker* is a party that arranges transactions between a buyer and a seller, and gets a commission when the deal is executed. Although this is not necessarily always the case, it is commonly understood that a broker does not act as a seller or as a buyer (merely executes instructions given to her), so she does not become a principal party to the deal. A *trader* is someone who buys and sells derivatives on the asset (e.g., electricity) and therefore takes positions in the market. Retailing may include subordinated services, see Sect. 9.1.3. A *retailer* buys electricity in large quantities from generators, and then sells smaller quantities to the end-users.

[3] This issue is discussed below in Sect. 9.3.2.1.

This chapter is specifically devoted to the retailing activity, which essentially consists of buying electricity wholesale and selling it to end consumers. Retailers must select a portfolio of electricity purchases in the short- and long-terms, and negotiate selling agreements with the end consumers. Any added value provided by retailers must be based on hedging consumers' risk, offering them suitable rates and services and promoting and developing new tools and products. The latter might include taking advantage of new metering equipment to offer a variety of contract types (such as flat rate or time-of-use [TOU] tariffs), or combining products (dual fuel, electricity and gas, for example), or catering to certain consumers' environmental concerns by supplying them with green energy.

Supply is a business with a high turnover and very narrow profit margins: i.e., the difference between the price at which suppliers sell electricity and the price at which they buy it is quite small, as is their unit profit per kWh.[4] Total revenues of a firm are, nonetheless, reasonable because of the large trading volume involved.

Efficiency should, therefore, improve in each of the processes comprising the retail business. These processes and their associated sub-processes are briefly discussed below, based on Temboury et al. [46].

9.1.3 Global Description of Retail Activities

In general, the electricity retail business encompasses a number of processes that can be classified in several ways. One criterion is to differentiate between:

- Technical tasks, such as billing and data mining (storing and using individualised consumer information) and
- Economic tasks, in particular energy purchase and commercial relationships with existing and new consumers.

Another way of classifying retailing processes would be by distinguishing between:

- Services intrinsic to supply: demand forecasting, balancing, wholesale market access (financial guarantees, settlement and information systems, personnel), network access management, metering equipment and consumer information and
- Additional services: training, energy audits, quality improvement, environmental responsibility advice, electrical equipment maintenance management, equipment purchases, network connection expansion or reinforcement, financing and multi-service offerings.

[4] For instance, according to the data provided by the Spanish Energy Regulatory Commission [5], the cost of retail in Spain as of January 2010 accounted for less than 4 % of the electricity bill.

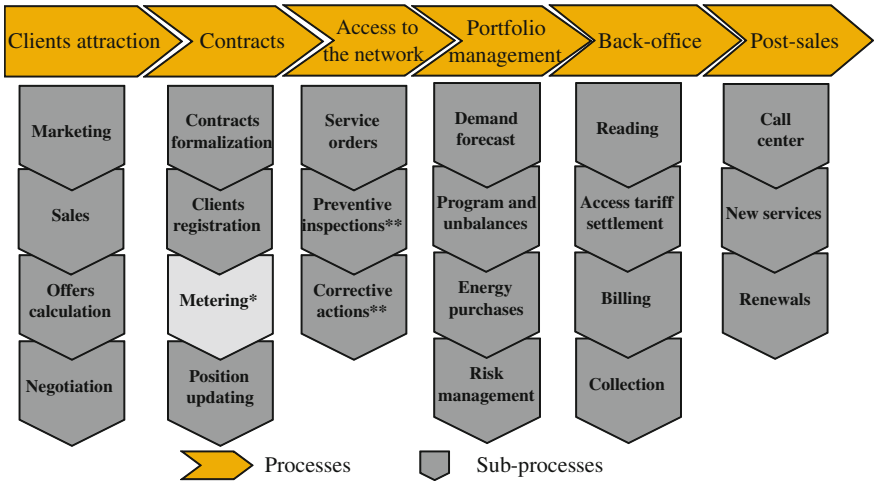


Fig. 9.1 Retail business processes [In most cases, metering is performed by distribution companies (see Sect. 9.4). These activities are associated primarily with consumers' first dealings with their liberalised retailer. While they should be incumbent upon the distribution company, they should advisably be performed by the retailer. Such activities include, for example, verifying that the capacity limiter is properly installed or that the meter is in good working order]

The tasks comprised by the retailing activity are described below for illustration purposes. Figure 9.1 shows one possible classification of the processes and relationships that constitute the retail business.

As a rule, these sub-processes are handled by different business areas in electricity retail companies. See Annex A for a detailed description.

In any case, it is important to take into consideration that the weight of the retail activity in the overall cost of electricity supply is not significant. See for instance in Fig. 9.2 and Table 9.1 the retail margins considered by two different regulatory authorities. This is a factor worth taking into consideration, particularly in the context of the discussion about the suitability of abolishing regulated tariffs and therefore fully liberalising the retail business (see Sect. 9.2.3).

9.2 End-User Price Regulation: Retail Market Development and Regulated Tariffs

The liberalisation of the power sector began with the wholesale market reform, and the liberalisation of the retail activity was frequently postponed to the later stages of the deregulation process or implemented gradually. After a good many years, a quick look at the current state of development of the retail business shows that competition and all it entails has been introduced in only a few cases. Before briefly reviewing progress worldwide, some attention should be given to the real meaning of "full liberalisation" in the retail context and exactly what "all it entails" means.

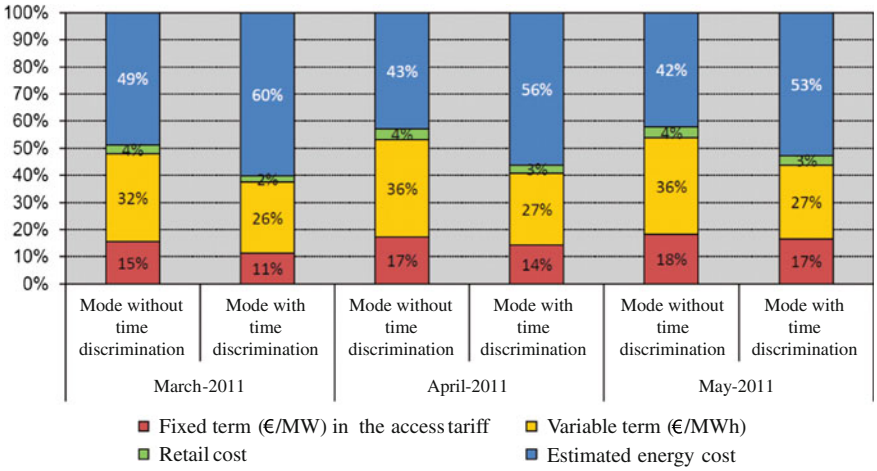


Fig. 9.2 Tariffs components in Spain for households [5]

Table 9.1 Sample cost stack for the A1 Residential Tariff in Western Australia [37]

A1 Nominal cost stack

Cost	2009/2010		2010/2011		2011/2012	
Component	¢/kWWh	%	¢/kWh	%	¢/kWh	%
Black energy costs	10.97	51	11.43	42	11.42	37
Carbon pollution reduction scheme	0.00	0	2.48	9	2.73	9
Mandatory renewable energy target	0.12	1	0.18	1	0.25	1
Ancillary services	0.14	1	0.14	1	0.15	0
Market fees	0.06	0	0.06	0	0.06	0
Retail operating costs	1.55	7	1.59	6	1.64	5
Net retail margin	0.63	3	0.79	3	0.89	3
Network costs	8.24	38	10.59	10.59	13.58	44
Total	21.71	100	27.27	100	30.72	100

According e.g. to Frontier Economics [19], 'black' energy cost measures the demand weighted average cost of new non-renewable generation plant, as well as the cost of meeting the reserve constraint, divided by each retailer's average MWh of use

In this context, the term "full liberalisation" is often identified with "full eligibility". As defined in successive European Directives, the term "eligible customers" means customers who are free to purchase electricity from the supplier of their choice. According to this criterion, an electric power retail market is fully liberalised when all customers are given that option.

This approach is overly simplistic, however, if electricity retail markets are to be analysed properly. Freedom of choice does not necessarily imply a healthy retail competition. The main differentiating factor is the existence (or persistence) of whatsoever manner of regulated price control. As briefly shown below, in most

of the retail markets established to date, the regulator still has some type of mechanism to protect small customers, who in one way or another are guaranteed a ceiling on electricity prices. Therefore, a fully liberalised retail market with "all that it entails" would be one in which customers are fully eligible, the regulator's role is limited to merely supervising and monitoring prices and market operation in general (compliance, performance monitoring and reporting and complaints) and no price caps or safeguards or administratively set tariffs are in place, or they are designed so that their impact is minimal.

The true degree of retail market liberalisation depends on the proportion of customers under the regulator's price protection and the degree of this protection, which can be readily assessed as the difference between the regulator's price caps and the market prices (an easy-to-measure and revealing indicator is the share of customers remaining under the regulator's protection).

The degree of development of retail markets in some areas of the world where electric power system reform has been implemented and information is more easily accessible is summarised below. The discussion shows how the design of regulated energy tariffs continues to be a key factor for the retail activity in many electric power systems in which full eligibility has been achieved.

9.2.1 Degree of Deregulation of the Retail Business Worldwide

9.2.1.1 European Union

The European Directive 2009/72/EC states, with regard to retail competition: "the interests of the Community include, inter alia, competition with regard to eligible customers". It goes on to say that "Member States shall ensure that eligible customers comprise: (…) from 1 July 2007, all customers". Full eligibility is therefore mandatory in the European Internal Electricity Market.

Nonetheless, the degree of retail market liberalisation and competition varies significantly across the Union. In practice, not many Member State governments are overjoyed at the thought of losing their ability to control energy prices. The French Government, for instance, has considered potential schemes to allow what has been defined as the consortium of *électro-intensives délocalisables* industrial consumers to "benefit from specific electricity purchase prices" ([6], *Le Ministère de l'Économie* 2005), a clear protection measure for these consumers. Recently, the Champsaur Commission [4] put forward a proposal whereby consumers in France would pay for their energy consumption based not on the market marginal price but on a sort of average price (i.e., cross-subsidising consumers with the inframarginal rents earned by nuclear generation). This proposal is laid out in the "Act on the New Organisation of the Electricity Market" [1].

While the European Directive establishes the EU's commitment to full eligibility, it also states that an "affordable" energy supply service should be guaranteed: "Member States shall ensure that all household customers, and (…) small enterprises, (…), enjoy universal service, that is, the right to be supplied with electricity of a specified quality within their territory at reasonable, easily and clearly comparable and transparent prices". It further declares: "To ensure the provision of universal service, Member States may appoint a supplier of last resort". While it opens the door to such suppliers, it does not specify their required characteristics, or much less what criteria must be met by the rates that they offer or which consumers should be entitled to it.

As discussed below, this wording has been interpreted by some Member States to mean that they still can decide whether energy market prices at any point in time are reasonable or not. In many other Member States, regulated last-resort tariffs (in their default format; see discussion below) protect a significant share of consumers. A number of governments still retain the prerogative to set price caps or any other manner of rules to avoid passing market prices on to some or all end-users.

9.2.1.2 The Slow-Pace or the Stagnant Development of American Retail Markets

In the US, where retailing is under state jurisdiction, it has never been deregulated in some states, while it is fully liberalised in others (see Fig. 9.3).[5] After the California crisis in particular, many considered that retail markets had been a failure in the US, at least for small consumers. For some, the main "evidence" of failure is that prices have allegedly risen more in states that deregulated than in the others (see [44]). Nonetheless, even if this assertion is accurate and prices have behaved in this way, the comparison is obviously unfair, since deregulation mostly happened in states with higher prices than average, to start with, and the quality of regulation in some of the deregulated states has been poor (below-cost tariff freezes, rising costs not passed on or default tariffs too low to encourage customers to opt for other options). The bottom line is that retail competition has less political support in the US now than it once had.

As Rose and Meeusen [44] report: "The overall status of state access has remained relatively unchanged for several years. Sixteen states and the District of Columbia have fully implemented their legislation and commission orders and currently allow full retail access for all customer groups. Nevada and Oregon allow retail access for larger customers only. Six states that passed restructuring legislation later delayed, repealed, or indefinitely postponed implementation. Twenty-six states are not considering retail access or restructuring at this time and no state has passed restructuring legislation since June of 2000, when the

[5] An up-to-date description of the situation of retail in the US can be found in http://www.eia.doe.gov/cneaf/electricity/page/restructuring/restructure_elect.html.

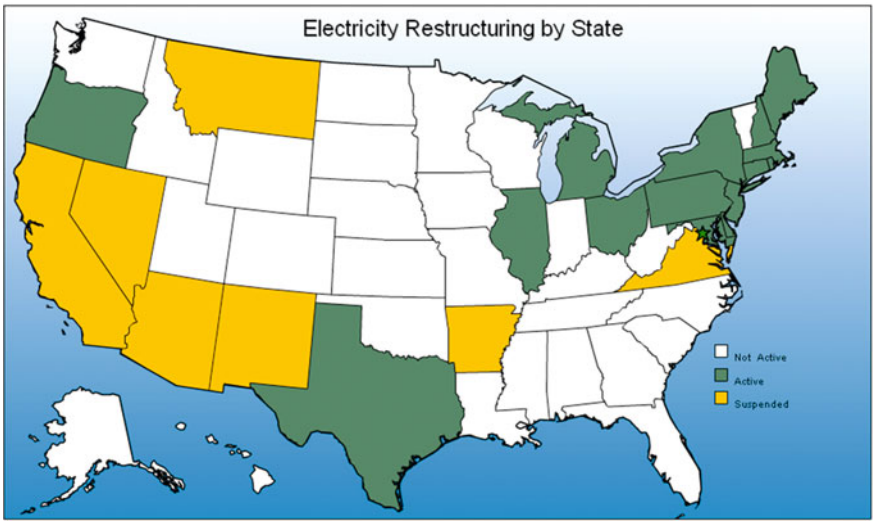


Fig. 9.3 Status of restructuring in the US (May 2010). *Source* Energy Information Administration

California and western power crisis was just beginning. A total of 34 states have repealed, delayed, suspended, or limited retail access to just large customers, or are now no longer considering retail access".

In the vast majority of the States in which retail market liberalisation (full eligibility) is still active, the retail market coexists with the basic service option (as it is called in Oregon[6] or Massachusetts,[7] for instance) or default service (as it is known in Pennsylvania[8]); in other words, traditional utilities continue to offer the choice of supply under regulated terms, in most cases acquiring their future supply needs through medium- and long-term default service auctions (see [32]).

A close look at Latin America, where market reform started back in the early 1980s (in Chile), reveals a similar situation. Nowhere in the Latin American electric power systems is the retail business liberalised at the residential level.

[6] In Oregon, the Oregon Public Utility Commission regulates three electricity options: basic service option, renewable resource options (fixed renewable, a fixed amount of electricity each month from new renewable resources or renewable usage, all of the electricity supply from renewable resources) and time-of-use option (on-peak or off-peak prices).

[7] According to the data provided by the Executive Office of Energy and Environmental Affairs, as of April 2010, around 15 % of the electric power customers in Massachusetts had migrated from incumbent to competitive generation sources.

[8] In Pennsylvania, each local electric utility has a "price to compare" (PTC). The PTC is the price charged by the local utility for the portion of the service that is open to competition. The Pennsylvania Office of Consumer Advocate (see www.oca.state.pa.us) provides electricity customers with information about the differences among the PTCs in their area and the prices offered by alternative suppliers. Data on the customers served by an alternative supplier in the various areas are also given in the Electric Shopping Statistics posted on the website.

At the same time, however, distribution companies are supposed to manage energy purchases to supply their captive demand, while competing on the retail market for large consumers. The result is a clearly undesirable regulatory situation, in which, without the necessary unbundling, distributors enter into long-term contracts to supply regulated demand with generating units sometimes even forming part of their own business group. This frequently leads to a struggle with the regulator, which often sets limits on the tariffs without further justification than its mistrust of market and contract prices. When tariffs are set below the value that would correspond to a pass-through of the market price plus the regulated charges, deficits ensue, thus eroding the financial health of the distribution companies and, as a result, the quality of supply.

Therefore, as these varied experiences show, the advisability of maintaining a default tariff once the retail market has been liberalised is still a subject of debate.

9.2.2 Regulated Energy Tariffs: First or Last Resort

In Spain, automobile insurance is mandatory for anyone who has a car. To ensure universal coverage, the authorities created an *Insurance Compensation Consortium,*[9] which underwrites the compulsory coverage of motor vehicles not accepted by any insurance company, as well as central, regional and local government-owned motor vehicles and automobiles belonging to other public bodies (upon request).

The *Insurance Compensation Consortium* is not meant to compete with market prices offered (its price usually exceeds the highest market price), but to safeguard drivers seeking insurance but unable to find an insurer willing to take their risk.

The present section analyses whether, in a new market context in which all consumers are theoretically eligible to choose their supplier, some form of regulated tariff should be made available to all consumers.

The liberalised electricity retail business must co-exist with some form of regulated tariffs to ensure that all consumers are supplied under exceptional circumstances. The regulator faces the task of establishing a method for calculating these rates, whether it be for the minimum level of regulatory intervention, meant to specifically ensure the supply for a short (but sufficient) period of time to consumers who may have been abandoned by their supplier (e.g., due to bankruptcy), or the maximum, which involves determining the rates to be applied to consumers (e.g., residential customers) that the regulator wants to protect from market vagaries. Here, we shall use the term "last resort or back-up" tariffs for the former and "default" tariffs for the latter.

[9] Consorcio de Compensación de Seguros, www.consorseguros.es.

9.2.2.1 Scope of Regulator Protection

There seems to be a consensus among energy regulators on the universal nature of electricity supply as well as on a related subject: the need for a supplier of last resort (SoLR). The question here is to decide how far such SoLR protection should go.

The various definitions of SoLR duties range from meeting supply demands in a "dire straits" situation at one extreme (which should really be the only interpretation of a "last resort" supplier; however, as this is not the case in actual fact, the term "back-up supplier" is used in the ensuing discussion) to the "default" supplier at the other.

The former approach is the minimum expression of a SoLR, intended only to ensure that all consumers will receive electricity if their existing retailer fails and to serve as a temporary hedge for a short period (the time needed by the affected consumers to find a new supplier). The latter approach, more in line with the traditional electricity tariff concept, represents a "more or less" regulated tariff for those consumers whom the regulator considers vulnerable (typically residential customers, but others may also be included). In principle, this alternative is only warranted where sufficiently competitive alternatives are lacking, for otherwise, it seems clear that the rather shaky justification of this regulatory provision is to protect "consumers who do not want to bother" or, more precisely, "consumers that the regulator, or ultimately the Government, does not want to bother". Section 9.2.3 contains a brief summary of the ongoing debate on the advisability of maintaining a default tariff.

9.2.2.2 The Real "Last Resort": The Back-Up Supplier

The back-up supplier is a safeguard mechanism to guarantee universal electricity supply, as mentioned in European Directive 2009/72/EC, but it does not necessarily have to protect all consumers in all cases (e.g., large customers are able to find a new supplier quickly).

While such a mechanism seems indispensable for resolving a lack of supply due to retailer bankruptcy, the fact is that many markets have not developed precise rules about how to proceed in this situation.[10] Exceptions include the United Kingdom, Australia and Italy, where specific regulations are in place.

Several key issues must be addressed when designing this protective mechanism, as listed below.

- Which consumers have the right to resort to the back-up supplier?
- How should the back-up supplier be designated?
- How should the failed supplier's customers be allocated?
- How long should the back-up supply situation last?
- How should the back-up supplier be remunerated?

[10] The EnergyXS bankruptcy experience in The Netherlands was a very good example of the problems that arise in the event of a regulatory void (see [25]).

An analysis of international experience shows that the answers to these questions vary.

- In some systems, the back-up supplier only covers households; in others, all consumers, no matter what kind, are temporarily protected, although sometimes under different conditions (in Italy, small consumers can revert to the default tariff, whereas only predefined transitional tariffs are available to larger consumers). For example, in Germany in 2010, there were three cases for which a default supplier was pertinent: (a) if the customer could not find a supplier on the market, (b) if the customer was moving in without having chosen a supplier (a contract starts by "switching the lights on" according to the regulations) and (c) if the customer does not choose a supplier himself. ERGEG [14], after consulting the different regulatory institutions of the Member States about the cases in which the last-resort and default supplier should be applicable, suggested the following alternatives: A. Customer does not pay, B. Customer cannot find a supplier on the market, C. Supplier goes bankrupt, D. Moving in without choosing a supplier, E. Customer does not choose a supplier (for instance when the market opens), F. Expired contract, G. Other.
- The assignment of consumers to a back-up supplier may be solved by direct allocation by the regulator (UK)[11] on a case-by-case basis, allocation to the retailer associated with the incumbent distributor in the area (e.g., the so-called "local retailers" in the state of Victoria in Australia), allocation among the programme responsible parties (network operator supplying companies) by market share in terms of number of customers, or allocation via area auctions (e.g., in Italy). In Texas, for instance, the Public Utility Commission (PUCT) designates providers of last resort (POLR) in each area of Texas open to competition. In general, the service is intended to be temporary and used only under rare circumstances when a REP (retail electric provider) is unable to provide service, or when a customer requests POLR service. In Austria, according to the regulations in force in 2011, every supplier was a potential back-up supplier and the customer could choose the supplier itself, there was no allocation method established in the regulation. Moreover, the back-up supplier was obliged to charge a tariff that did not surpass its own standard average tariff, i.e., the tariff that most of its consumers pay for their electricity.[12] In Germany, the distribution system operator (DSO) designates the supplier that has most of the stranded customers within its network area. This criterion determines which supplier will be the last-resort or default supplier.
- The period during which consumers are entitled to pay the last resort tariff ranges from 3 (e.g., the Netherlands) to 6 months (UK). In other cases no

[11] The Authority may issue a last resort direction to a licensee if it can comply without significantly impairing its ability to supply its customers and to fulfil its contractual obligations.

[12] If suppliers cannot know in advance how many consumers would be allocated to them in case of a bankruptcy, and they are not allowed to charge a higher price to them, they are somehow forced to internalise this future risk in their prices, so to some extent their current consumers subsidise the ones who eventually might contract with a "riskier" retailer.

maximum period is set, but a premium is placed on the price of electricity, as an incentive for consumers to abandon the SoLR. In Texas, for instance, the rules [43] establish that the energy charge for a residential consumer is obtained by multiplying the actual hourly market prices for the customer over the billing period by the amount of kWh used, and then applying a 20 % surcharge.
- In some cases, the remuneration paid to last-resort retailers is defined by the regulator on a case-by-case basis in light of the circumstances (UK). In others it is established by auction (as in Italy), or determined through an *ex ante* calculation designed to assess the costs that might arise if the last-resort retailer is obliged to take on a customer due to failure of a standard retailer (common in the Australian case).

9.2.2.3 Default Supplier

In systems having both a wholesale and a retail market, a default rate should not theoretically be necessary. Consumers, in addition to paying access tariffs, would purchase energy on the market to meet their needs.

As mentioned earlier, however, the actual situation is quite different. The reasons often given for this circumstance, in line with the discussion in Sect. 9.3.1, range from market malfunction (market power abuse in generation, retailing, etc.) to some consumers' (especially households) aversion to change and the need for a default tariff (which is not always considered temporary) that protects them from possible abuse by the (sometimes scarce) suppliers.

In this case, the key questions are listed below.

- Which customers have the right to resort to the default supplier?
- How should the default supplier be designated?
- How are the energy cost and the retail margin calculated? How are energy cost deviations (if any) allocated *ex post*?
- Are implicit or explicit incentives in place for consumers to give up the default tariffs? If an additional charge is established, how is this extra income set aside?

Here again, the answer to the first question varies from one jurisdiction to another, although in most cases it depends on the nature of the customer (households, as in Italy) or on their annual consumption (e.g., less than 10 kVA in Spain, 250 MWh in some Canadian states, 1 MWh in Texas).

In the vast majority of cases where a default tariff exists, the regulated SoLR is associated in some way with the area distributor. A different approach is the Italian model, in which a public institution (*Acquirente Unico*[13] or Single Buyer) was created, whose purpose is to purchase energy for regulated customers, while all other supply-related tasks are performed by the distributors.

[13] www.acquirenteunico.it

Lastly, the methodology for calculating the cost of energy and the default supplier's retail margin remains to be determined. International experience shows that any number of alternatives have been instituted, some of which are listed below.

- Rates are monitored by the regulator and sometimes, as in Denmark, compared to those offered by market suppliers. In New South Wales [23], the regulated retail tariffs and regulated retail charges that standard retail suppliers may charge small retail customers are proposed by the former and approved by the regulator for a 4-year period on a CPI-X (price cap) basis.
- Rates are determined through ad hoc supply auctions, for example, in many states in the U.S.[14] or (partially) in Spain.
- Rates are based upon the purchases of a body created for this purpose (the aforementioned Italian Single Buyer, which buys in the day-ahead wholesale market, under bilateral contracts and in short-term CfD (contracts for differences) auctions).
- Rates are determined by the regulator following a variety of criteria. They may depend on forward market prices (Alberta, Canada) or reflect that a significant percentage of energy is purchased in advance at prices determined well in advance (energy may be purchased from nuclear and hydro in Ontario; or from some of the generators of the former incumbent at regulated prices, as in Ireland's ESB). Another alternative is to use a price that would cover the estimated total costs that a new producer entering the market would incur, as determined by some prescribed procedure.

9.2.2.4 Vulnerable Customers and Energy Poverty

Beyond any further consideration about how far regulator protection should go, the fact that electricity supply is universally recognised as essential raises the issue of how regulation should deal with the so-called fuel or energy poor or vulnerable consumers.

Two of the many regulatory issues that must be carefully resolved in this field are set out below.

- Which categories of consumers should be considered vulnerable and, therefore, entitled to special support?
- What is the most efficient way to implement this support?

[14] For instance, in Massachusetts, the Executive Office of Energy and Environmental Affairs [15] states: "For residential and small commercial and industrial customers, the Department directed each distribution company to procure 50 percent of its default service supply semi-annually, for 12-month terms. As a result, default service prices for these smaller customers (for both the monthly and the six-month pricing options) are now based on an average of the results of two separate procurements". See also Loxley [32] and Tierney [47].

Definition of vulnerable consumers

One key concept that should be established before initiating any discussion on this issue is that protecting vulnerable consumers does not equate with maintaining regulated energy prices for certain categories of customers.

One of the best published definitions, which summarises the various approaches, was laid down by the Electricity Authority of New Zealand (2010):

> A domestic electricity consumer is defined as vulnerable if:
>
> - for reasons of age, health or disability, the disconnection of electricity to that domestic consumer presents a clear threat to the health or wellbeing of that domestic consumer; and/or
> - it is genuinely difficult for the domestic consumer to pay his or her electricity bills because of severe financial insecurity, whether temporary or permanent.

Therefore, vulnerable consumers are defined from two points of view, one purely economic and the other social.[15] For instance, in the state of Massachusetts, a customer meets income eligibility requirements for the Low-Income Home Energy Assistance Program ("LIHEAP") administered by the Department of Housing and Community Development if the household income does not exceed 60 % of the median income in the state.[16]

Support system design

Consumer protection mechanisms may be designed in a number of ways, namely, through a general social support system independent of the energy industry (such as tax benefits or measures favoring particularly vulnerable consumers) or as part of it (explicit subsidies through lower tariffs).

For instance, the Texas Department of Housing and Community Affairs' Weatherisation Programme[17] is designed to offer qualified low-income Texans an energy audit, or a review of the home's energy efficiency, and installation of weatherisation measures to increase this efficiency. In the UK, the winter fuel payment is an annual tax-free lump sum given to eligible people aged 60 or over to help pay their winter heating costs.[18]

Conversely, also in the UK, all energy providers came to an agreement with the Government to offer social tariffs to help their most vulnerable customers cope with electricity costs for the period 2008–2011 (eligibility criteria may vary among suppliers). In Flanders (Belgium), as of 2009, each individual in a household receives 100 kWh free of charge annually, and the household as a whole also receives 100 kWh free per year.

In Latin America, explicit cross subsidies are much more common for consumers. In Colombia, for example, consumers are ranked in six different categories

[15] See ERGEG [13] for a survey of the definition of vulnerable customer in the EU.

[16] Executive Office of Energy and Environmental Affairs, www.mass.gov.

[17] See www.puc.state.tx.us.

[18] See www.dwp.gov.uk.

depending on their income; the lowest income category pays around 50 % of the full cost of electricity (or even get a minimum consumption for free) while the two wealthiest categories among the six as well as businesses are taxed by up to 30 %.

In addition to these special tariffs and subsidies, non-economic support systems are also commonly implemented and mainly include measures to help vulnerable customers avoid disconnection. In Italy, for instance, this applies to disabled persons living in homes fitted with medical or safety equipment.

Two brief comments are in order regarding the efficiency implications of these measures.

- In theory, it would be better to outsource subsidies of this type and centralise them in a specific governmental agency able to provide more comprehensive and coordinated support to those individuals who might need it.
- To avoid minimising the impact of market price signals, support should be offered as a lump sum payment regardless of the quantity of energy consumed. Discounts on the energy tariff lowers the consumers' incentive to practise efficient consumption in proportion to the percentage discounted.

9.2.3 What Should be Done with the Default Tariff for Small Consumers?

Assuming that an efficient power market exists, the next question is what should be done with the default tariff. Is it superfluous? Does it constitute an obstacle to retail competition?

The two extreme alternatives consist of giving full eligibility to all consumers and completely eliminate the regulated tariff, versus not implementing retail competition for small consumers, but applying a well-designed regulated tariff or set of tariffs for the consumers to choose. An intermediate approach would open retail competition to all consumers and keep for all consumers an administratively calculated default option (which typically contains a surcharge to incentivise consumers to seek alternatives, the so-called shopping credit). The rationale behind these options has been clearly expressed in the literature. A particularly interesting debate took place between Paul Joskow and Stephen Littlechild, each one of them defending one of the extreme options, which is referenced and discussed below.

9.2.3.1 The Pass-Through Alternative

The first basic choice is to maintain the regulated tariffs and not to implement retail competition for small consumers. This position, defended by Joskow [26], is based on the assumption that the deregulation of the retail market will not lead to any significant efficiency improvement or added value. This author examines the areas where retailers might add value beyond the mere delivery of electricity: mainly,

selling their goods at suitable times or locations, keeping stocks or developing more efficient technologies, and concludes that these strategies cannot be readily transferred to power markets. In addition, he evaluates the retailers' potential to add value for the customer and concludes that it is questionable.

In particular, the potential retail savings on activities such as metering, billing or customer services are uncertain and their expected economic impact is too low to be significant for most residential customers. For retailers, the advertising costs and the financial risks incurred would possibly offset any potential gains. He concedes that retailers may introduce certain benefits. They may offer new billing options, improve the operation of the wholesale market by raising the number of agents involved, enhance demand elasticity or provide energy efficiency-related services. That retailers alone would induce these changes is unclear, however.

Therefore, if retailers might incur non-negligible costs and their opportunities to create added value services are doubtful, they will generally not be able to compete with a regulated tariff that has been suitably calculated. Such a tariff could be provided through a basic electricity service (BES, see Sect. 12.2.3.1) offered by the distribution company, based on an energy term calculated as a transparent pass-through of the market price.

If retailers are unable to compete effectively with the regulated tariff, they have no opportunity to create value. Consequently, the lack of retail competition for small consumers should not be viewed as a problem and the regulated tariff should be maintained indefinitely for small consumers and calculated simply by passing the market price on to them.

Joskow was not the only author to reach this conclusion. Waddams Price [49], for instance, states: "In such markets suppliers cannot differentiate themselves according to the physical properties of their product, though they can source the product from environmentally friendly origins, for example marketing electricity as 'green', and try to convince the consumer that their custom results in fewer global warming emissions. Retailers may also differentiate their product through ancillary services, such as meter reading or billing; or by appealing to social concerns (for example by enabling the firm to offer lower prices to low income consumers). All of these strategies are used by UK suppliers to differentiate their product. However the inherent homogeneity of the product raises the Bertrand paradox for potential entrants. If suppliers cannot differentiate their product from that of others in the market, competition is likely to focus on price (Bertrand) competition, so that effective competition would drive the price of the product down to marginal cost. If there are any non marginal costs (for example the cost of a computer system to administer metering and billing, or of advertising to consumers) then such a market has no prospect of being able to sustain profitable entry. The only feasible solutions seem to be either no entry, or an arrangement which will sustain prices above marginal cost after entry."

A good many other authors have drawn similar conclusions. Flaim [18] highlights (among other reasons) the absence of value added service relative to the transaction costs of serving small customers, Brennan [2] points to consumers'

reluctance to change and Defeuilley [9] underscores the inability of consumers to make appropriate choices and the lack of innovative processes.[19]

9.2.3.2 The Defence of the Retail Market Option

Littlechild [28] contends that a plain pass-through regulated tariff is neither transparent nor even desirable. Profiling and uncertainty about future prices is the reason for the first disadvantage, and international experience is the support for the second one.

Domestic customers appear to be risk averse (agreeing on a tariff with a retailer reduces the risk of a strict pass-through tariff that might need to be updated) and not indifferent to choices, and retailing is precisely the activity that can lead to products that best suit customer preferences and their low risk appetite. Moreover, introducing competition is equivalent to opening a door to innovation. The market can implement creative alternatives that a regulated framework would never consider.

From this perspective, establishing a BES makes no sense. If retailers decide to offer a direct pass-through option to small consumers, introducing an additional regulated alternative is unnecessary. On the contrary, the absence of such an offer from retailers would prove it to be unnecessary, and stands as evidence that consumers neither demand such an option nor retailers consider it appealing.

From this perspective, the advantages of competition are clear, and the retail market should be deregulated for all customers, including residential users, the implication being that the regulated tariff should disappear.

9.2.3.3 One Step Further in the Pass-Through Alternative

From the authors' perspective, the risk assignment problem mentioned by Littlechild is the weightiest objection to a BES. Nevertheless, introducing retail competition is not the only way of tackling this problem, which can also be solved in a more regulated framework. A well-designed regulated tariff should not be identified with a BES, in the sense that a direct pass-through is neither the only nor the most suitable approach. The "market price" to be passed through to the tariff should not be merely the hourly spot market price, as that price should at least be considered together with the prices in longer term contract markets. The proportion of the price drawn from each of these two markets determines the level of risk that the regulator deems that customers under the regulated tariff should bear. User risk profiles, in turn, can be taken into account through a number of regulated tariff formats.

From this perspective, if the risk issue is assumed to be the most significant objection, and a regulated framework is believed to be able to successfully address

[19] See also Littlechild's [29] response.

the question, the option to choose a properly calculated regulated default[20] tariff should remain open to all consumers. The Spanish White Paper [41, 42; extended in 48], proposed striking a balance between the volatility of relying on only the daily market spot price and requiring the purchase of this energy through long-term contracts in organised public auctions. It therefore proposed that the "energy-market price" used to compute the default tariff should be obtained as a weighted average (established by the authorities) of two energy prices: the short-term market price (40 %, for instance) and the price at which regulated tariff retailers purchase energy under 1-year contracts in organised auctions held at regular intervals throughout the year (for instance, four auctions per year, where 15 % of the energy required by each retailer is purchased at each auction[21]).

9.2.3.4 The Final Defence: Giving the Market Option an Initial Boost

As Joskow notes, retailers may not be able to compete with a well-designed regulated tariff in the current situation. International experience supports this view, as do the opinions of the main industry players, regulators and retailers. A consensus appears to have been reached about the need to add a shopping credit to regulated tariff prices to give the retail business a chance to develop. The justification for this extra cost is that a system where consumers offered regulated tariffs can never replicate the innovation generated by a competitive market.

Thus, assuming that the market may create added value, the market option is the correct choice. A shopping credit included in a well-designed default tariff should account for these expected benefits, reflecting the potential value added by retailing. Faith in the market would, therefore, mean that competition would lead retailers to develop their activity at lower costs than this extra charge.

From a Joskow-like perspective, which appears to assume that suppliers are unable to create value, the extra cost would be interpreted as wasteful, and should disappear altogether. By contrast, from a standpoint closer to Littlechild's, this shopping credit would tend to infinity, and what should be abolished is the regulated tariff. Intermediate approaches would lead to halfway solutions, where the shopping credit would be viewed as an investment that would be as large as the expected benefits of competition.

However, this approach merits additional comment: in certain systems, regulated default tariffs have been abolished only for certain consumers (usually the

[20] A default tariff would give all customers the option to purchase their energy at a well-calculated price that reflects market outcomes—not only spot markets, but also forward and contract markets—with no shopping credit included.

[21] Indeed, in Spain Order ITC/2129/2006 mandates distribution companies to purchase some 5 % of their energy at organised long-term auctions, which can be interpreted as a first step in the suggested direction. A new draft by the Energy Ministry seems to require 100 % of the energy under a regulated tariff to be purchased in open auctions, which appears to be at the other extreme.

larger accounts), while the rest are either afforded the option to choose the regulated tariff or are ineligible for any other (as in many Latin American power systems, e.g., Brazil and Guatemala). If the value that retailers are able to add does not offset the costs incurred (what happens particularly when the retail market and supplier are still immature), "liberalised" consumers would be unfairly treated, since they have to pay a higher cost. Indeed, since the decision of abolishing the regulated tariff for certain (or all) consumers is often justified as regulatory support for retail business development, a partial liberalisation of the retail business, affecting only some of them (for example large industrials), implicitly means that the costs of investing in the future advantages of retailing fully rely on the non-eligible consumers (who cannot not choose and thus defray part of the "start-up cost" of the retail market).[22]

Conversely, if the retailer were able to provide added value, non-eligible consumers would bear an extra cost. This leads to the conclusion that the same options should be available to all consumers.

Retail in the context of actual tariff design

The shopping credit is merely an extra charge to be added to a well-designed tariff, i.e., one calculated as a complete pass-through of the hourly spot prices, thereby avoiding cross-subsidies among consumer types. However, in real life, the design of most regulated tariffs appears to deviate greatly from this description.

First, cross-subsidies and design flaws that result in uneven effects on different categories of customers are common. Under these circumstances, the retail business has proven to be an activity with added value, not just for unfairly treated customers, but for the entire system, as a tool to detect these flaws. A circumstance that led to the development of independent retailers in Spain, for instance, was the existence of a flawed and asymmetrical design of consumer charges in connection with capacity payments to generators.[23]

The growing tendency to base the default tariff on long-term, often inflexible regulated contracts arising from mandatory public auctions[24] affords retailers the opportunity to attract customers. Larger scale consumers, especially, can be lured

[22] In these Latin American cases, this market segmentation is implemented to protect large (industrial) consumers from the cross subsidies embedded in the regulated tariffs designed for households. In most cases, high income consumers subsidise low income ones by paying a higher rate.

[23] This is not an isolated example by any means. One of the main factors that favoured retailing in Guatemala was that a customer could avoid paying the stranded costs resulting from the 15-year contracts signed by distributors before the system was liberalised.

[24] Brazil's energy auctions (Bezerra 2006) are becoming a benchmark in Latin America (followed, for instance, by Peru and Panama). The 'regulated bilateral contracting for the energy demand of distributors through public auctions' announced by the Spanish Government, for instance, defines just two products, a base-load contract and a single-load profile contract for the entire system.

by flexible contracts better suited to their energy needs. Retailers would profit in such cases from the lower marketing costs involved in reaching these consumers, rather than the residential ones.

Some prerequisites should be enforced to ensure the institution of sound regulated tariff design and the attainment of a competitive retail environment. These include an efficient wholesale market and a lack of vertical integration between the distribution and retail businesses. When these conditions are not met, additional difficulties arise in retail market deregulation.

9.3 Barriers to Retail Market Development

In the electric power industry, experience shows that the development of the generation side of the market, in systems that have opted for restructuring, is now fairly complete, even if rather imperfectly so in some instances. However, full liberalisation of the retail business has usually been postponed to the later stages of the deregulation process. International experience is characterised by delays in the stipulated processes, the persistence of cross-subsidies that leave no niche for retailers in certain customer segments and inconsistent consumer protection policies.

Many reasons can be given for this slow pace of retail business development. Two of the most prominent are: interference from the regulated tariff (typically due to political constraints and a lack of trust in market outcomes, usually a result of market power concerns), and the incomplete unbundling of activities that constitute a natural monopoly from businesses that can theoretically be market-oriented (in the context of this chapter, mainly the separation of distribution from retail, respectively). These issues are briefly addressed below.

9.3.1 The Interference of Regulated Tariffs

Governments have typically been reluctant to eliminate regulated tariffs. According to ERGEG [13], as of 1 July 2008, regulated end-user prices for electricity in the European Union still existed in 17 countries for households and in 14 for non-household consumers, while regulated end-user gas prices persisted in 16 countries for residential users and in 13 for other consumers. End-user price regulation has disappeared completely only in a few systems, namely the UK, The Netherlands and the Nordic countries, although, in the latter two, many retailers are publicly owned or have significant governmental influence, particularly at the

local level. Such companies tend to preserve the traditional concern for the protection of their customers, limiting competition from alternative suppliers.

This situation (i.e., the persistence of regulated tariffs) undoubtedly represents a significant obstacle to the development of the retail market. Among other effects, it discourages consumers from searching for alternative suppliers and typically prevents their exposure to more elaborate price signals.

The reasons for this reluctance on the part of regulators are often numerous, but political motivations are not lacking. Electricity tariffs have traditionally been a tool widely used by governments to achieve complementary and often scantly transparent objectives (protection of electricity-intensive local industries, subsidies for domestic fuels, territorial policy or inflation control, to name a few).

The early stages of the liberalisation process (the 1990s and the early years of the twenty-first century) concurred with a period of declining fuel prices and capital costs. That situation has recently reversed, which partly explains why many regulators are averse to eliminating regulated tariffs (by way of example, in 2007 the French government decided to backtrack, approving the "Tartam" tariff, which basically means allowing all customers who have signed a contract with a supplier to terminate it and revert to the regulated rate).

The major problem is not the existence or preservation of regulated tariffs, however, but the fact that these tariffs have often been ill designed and calculated, establishing values deliberately below the minimum levels needed to cover the cost of energy. When this happens, the unfair competition posed by the regulated tariff is a barrier that has rendered the retailing activity practically impossible.

It is nonetheless understandable that regulators are usually reluctant to leave prices for all consumers entirely to the market. The determining factor has often been their lack of confidence in wholesale markets, because of an inadequate market structure and the fear of market power. In such cases, regulators have preferred to postpone any necessary changes in market structure or the application of transitory measures to mitigate market power (e.g., virtual power plant auctions) and have continued to intervene in the retail market by maintaining regulated tariffs. Spain is a good example of the difficulties that an inadequate power industry structure may pose for the establishment of a deregulated retail market.

9.3.2 Structural Barriers

This section ignores obstacles to retailing that may result from inadequately designed regulated tariffs. Rather, an appropriate tariff design, including a shopping credit to help overcome consumer inertia, is assumed. The focus is on barriers of another type: insufficient unbundling of distribution and retail activities, inadequate switching procedures and improper commercial practices.

9.3.2.1 Insufficient Unbundling

Between distribution and retail activities

As stated by the EU Directive 2009/72/EC, non-discriminatory access to the distribution network makes possible retail competition.[25] The scope for discrimination, as regards third-party access, however, is less significant at the distribution level than at the transmission level, where congestion and the influence of generation or supply interests are generally greater. This situation might change in the near future, if new business models at distribution level become relevant, as it seems they will. The rules on legal and functional unbundling currently in place in the EU might lead to effective separation, provided they are more clearly defined, properly implemented and closely monitored. To create a level retail playing field, distribution system operators' activities should be monitored to prevent them from taking advantage of the competitive edge afforded by vertical integration of the retailing and distribution activities, in particular as regard household and small non-household customers. However, the experience so far has not been satisfactory, as described below.

In principle, as any basic textbook on the regulation of energy systems allows, the basis for a successful restructuring process is the separation of activities. The determination of exactly where unbundling should begin and end, however, has been one of the issues most heatedly debated in structural reform.

Insufficient unbundling is one of the most serious obstacles to retail competition. European Directive 96/92/EC laid down the initial requirements for unbundling network and commercial activities, and subsequent Directives 2003/54/EC (for electricity) and 2003/55/EC (for gas) provided for a distribution system operator or DSO independent (at least in legal and decision-making terms) from any other industry activity. Subsequent debate on the Third Package within the EU revolved around unbundling, ending with the publication of the EU Directive 2009/72/EC. The focus was always on the level of separation between generation and transmission activities, however, while the separation between distribution and retail has seldom been tabled.

The literature nonetheless offers a substantial corpus of papers that analyse the pros and cons of forced distribution/retail unbundling. In particular, the Dutch experience, the only case in which the regulator drastically opted for ownership unbundling, triggered a number of papers debating the suitability of this approach.[26]

The pros and cons of ownership unbundling and whether the pros as a whole outweigh the cons are not addressed in this chapter, which nonetheless acknowledges the consensus opinion that unbundling affects retail competition to one degree or another.

---

[25] It important to notice that this unbundling does not exist in the US (except for Texas) and in Latin America, despite the fact that in a significant number of US states there is retail competition. The situation in Latin America is even worse from the unbundling perspective, since generation and distribution are still bundled in the vast majority of cases.

[26] See Mulder et al. [34] for a good review of the various points of view in the discussion on the Dutch experience.

- Mulder et al. [34], for instance, assert that the impact of ownership unbundling on retail competition appears to be quantitatively insignificant. Fetz and Filippini [17] contend that "economies of vertical integration between electricity production and distribution result from reduced transaction costs, better coordination of highly specific and interdependent investments and less financial risk". Boschek [3] claims that the benefits of ownership unbundling are "neither supported by the best available economic evidence nor can any assessment built on it be effects-based".
- Davies and Waddams Price [7], by contrast, analysing the UK market, find "clear evidence that those UK incumbent electricity suppliers which remained vertically integrated with their local distributor have retained a higher market share than those where these functions have been undertaken by separately owned companies". Also, according to Ofgem [39], "until very recently, the five former incumbent electricity suppliers charged electricity customers in their former monopoly areas an average of over 10 % higher prices than comparable 'out-of-area' customers".

Distribution and retail activities are often still integrated in other jurisdictions, forming a framework that favors certain irregular practices that prevent the retail market from successfully developing. Examples of the many known patterns of behavior in this category (some of which are rather rare but no less reprehensible) are listed below.

- When a consumer tries to change to a supplier that is not part of the same business group as the incumbent distributor (which is in charge of metering consumption), the distribution company has been reported to require an increase in the contracted capacity (the maximum accommodated by the meter). This undesired capacity increase (not required when the customer remains with the retailer that belongs to the same group as the distribution company) may offset the savings that the consumer expected to obtain by switching suppliers.
- Excessive rates are often charged for the metering equipment and the associated service.
- Access to commercial information is uneven, giving the retailer belonging to the same group as the distributor an advantage.
- The retailer's commercial advertising includes references to the distributor's services, such as advantages in technical service or quality of supply.

Between generation and retail activities

In principle, unlike the vertical integration between retail and distribution activities, such integration between generation and retailing is not usually identified as a problem for the healthy operation of energy markets. Both are competitive activities, and it is difficult to distinguish where the former ends and the latter begins. Generators often sell their future output under long-term contracts, sometimes to retailers and at others directly to large-scale consumers.

Integrating generation and retail in the same company should pose no particular problem in a well-developed quasi-perfect market. But it has proven to be an issue

in today's electricity markets, particularly where (in near all cases, actually) they are characterised by a rather small number of generating companies.

Electricity suppliers (meaning in this case generators who are also retailers) try to vertically integrate, matching their generating capacity and market share in the wholesale market to their retail portfolio. The alleged justification for this strategy is to balance their buying and selling positions, in an attempt to hedge risks. But such a situation is an obvious indication of market dysfunction whose symptoms are insufficient liquidity and high transaction costs. Holding an open position in the market, whether long or short, is penalised. A mere glance at liberalised electricity markets the world over shows that this is always the case, i.e., that in a market in which large incumbents are somehow vertically integrated (in this case we talk about generation and retail), competition for just retailing companies is almost impossible. Indisputable proof lies in the fact that no significant independent retailer operating on any energy market does not have a significant (sufficient) generation portfolio, in other words, it does not have a very short position (i.e., much less generation than demand). While plenty of such actors have entered electricity markets everywhere, none has ever managed to establish a significant foothold.

This is a matter of growing concern for regulators, but unfortunately no clearly workable solution has yet been forthcoming. The issue is extensively addressed in Ofgem's [40] "Retail market review", for instance. The Office highlights the problem in its report on the survey,[27] and provides some evidence of its existence, but whether or not the solutions suggested would be effective is far from clear. Ofgem proposes, for example, "a new licence condition that would require the Big 6 to make available between 10 and 20 % of their power generation into the market through a regular Mandatory Auction (MA)". While the measure is undoubtedly well intentioned, it poses a logistical problem: buyers for such large amounts of energy would not be readily found. And if they were, generators might well claim that the result would be a mono- or oligopsony.

This matter can nonetheless be viewed from another perspective. In an oligopoly, incumbents that have reached a balance between supply and demand have no significant incentive to increase its market share in the retail market at a faster pace than in the generation side. Attracting more customers would result in a short position that entails market risks. Rather, their observed behavior has been to consolidate their position and to self-manage their portfolios, removing liquidity from (especially short-term) energy markets. This lack of liquidity becomes a crucial entry barrier for newcomers.

[27] "A lack of wholesale products and wholesale market transparency combine to frustrate the trading activities of non-vertically integrated suppliers and may protect any advantaged position of the Big 6; and there is further evidence of companies pursuing similar pricing strategies" [40].

9.3.2.2 Switching Procedures

ERGEG [11] defines supplier switching as "the action through which a customer changes supplier; for instance, a switch is essentially seen as the freedom (by choice) to change supplier for a specific supply or metering point and the volume of energy associated with it".

The absence of appropriate mechanisms for making this change is the second type of barrier. A number of obstacles have been identified, some of which are listed below.

- The procedures for the exchange of information between the distributor and retailers are insufficiently developed. The switching process involves numerous tedious steps that amount to a high "transaction cost" for switching customers.
- Deadlines are not precisely specified, resulting in unacceptable delays.

Undesirable situations involving improper commercial behavior have been also reported, such as:

- Invoicing scams, in which consumers are sometimes charged amounts unrelated to their actual consumption.
- Abusive commercial practices by retailers in their attempt to attract new customers, for instance, "moving" consumers to the liberalised market (obviously with the retailer belonging to the same group as the distribution company) without their explicit consent, or using misleading advertising.

ERGEG [12] "has identified two strategic priorities for the supplier switching process. These are (1) promote easy, cost efficient and standardised switching and activating/deactivating procedure and (2) ensure customer confidence and sound monitoring systems." In addition, ERGEG has set out some basic guidelines, which are shown below.

- The customer's right to switch supplier should be statutory.
- The process of switching supplier has to be easy from the customer's point of view and the customer shall not pay any direct fees for changing supplier.
- The process of data exchange has to be cost efficient and standardised for the suppliers and the distribution system operators.
- Clear roles and responsibilities among actors are of vital importance throughout the entire procedure.
- The switching period should be as short as possible. There should not be any unnecessary obstacles for switching from the customer's point of view.
- The customer should only need to be in direct contact with one party, preferably the new supplier, when initiating the switch.
- There should be easy access to relevant and correct information for the customer prior to switching. The regulator or some other competent body should ensure the availability of a list of alternative suppliers.

- Regulators and/or other authorities should ensure sound market monitoring. Information about various market indicators should be available in order to make national analyses and comparisons between countries and markets. Harmonisation on the definitions of switching and the statistics needed should be sought across markets and countries.

9.3.3 Brief Conclusion: Measures to Encourage Retail Competition

The primary conditions necessary to establish a functional retail market include a reliable and transparent wholesale market, a regulated and properly designed default tariff (as explained above), adequate metering equipment and data processing, institutions that effectively protect consumer rights and a stable regulatory framework. In addition to these basic conditions, certain extra measures and procedures may mitigate the problems caused by the aforementioned barriers.

Unbundling is the fundamental step in power system restructuring and liberalisation. The separation of monopolistic and liberalised activities is essential. As indicated above, the current European Electricity Directive stipulates that distribution grid operation must be unbundled (legally, at least; separate ownership is not mandatory) from any retailing activity. Nevertheless, as the examples above show, only complete (ownership) separation between the retailer (for all customer categories, i.e., default regulated tariffs and market purchases) and the distribution company will remove all additional obstacles to retailing.

Ownership unbundling would effectively eliminate any incentive for the inappropriate practices described here. However, in many European countries, this measure is difficult to apply so late in the restructuring and liberalisation process. Regulated retailing should at least be legally unbundled from the liberalised business, to prevent insider access to information by the latter that would constitute a commercial advantage. In addition, explicit minimum quality standards and metering responsibilities, usually assigned to the distribution company, must be established.

Switching procedures should include clear mechanisms for accessing commercial information. An appropriate data management procedure should guarantee the availability of information for all interested retailers, to the extent allowed under data protection legislation. Any such procedure should consist of at least a common but decentralised scheme, with standard data management and switching procedures. If effective, this option would avoid the extra cost of creating a new centralised body. However, if experience shows that this does not provide sufficient guarantees, a centralised switching agency should be created. The deadlines and criteria for revising the measures adopted must be duly specified.

Other consumer protection measures should be considered, such as limiting the maximum duration of contracts or providing for cancellation. Additional procedures might include preventing users from choosing the supplier pertaining to the

same group as their distribution company, or not allowing regulated retailers to operate in areas supplied by their group distribution company. Finally, strict supervision by regulatory agencies is necessary to prevent potential irregular practices and furnish advice on the appropriate package of measures to be finally adopted.

That notwithstanding, a number of innovative ideas could be considered in retail market design:

- Allow for full retail eligibility, while establishing a simple and standard regulated tariff with a short number of alternatives (such as a quarterly TOU tariff with no more than two rates, peak and off-peak, calculated on the basis of a default service auction). This tariff would be in place indefinitely, as a default tariff, a fallback alternative for consumers (and a regulated price to beat).
- Unbundle retailing and distribution by banning retailers from operating in areas that are supplied by the distribution company in the same business group. (When set out in a White Paper prepared for the Spanish Government in 2005, this recommendation met with the discontent of the incumbent utilities.) Even the regulated tariff should be offered by retailers in a different group.
- Include an additional but crucial item. To counter retailers' incentive to "sell more to earn more", the retail charge calculated in the regulated tariff should be a flat amount, irrespective of the amount of energy consumed by their customers. It could be likened to a shopping credit but, contrary to standard practice (formerly the price to beat in Texas, for instance[28]), it should be defined as a flat amount, not a percentage over the sum of power billed.
- This design would enable retailers to make a profit, mainly through innovative tariff design and demand response tools. They are probably unable to add significant value because they purchase energy on the wholesale market (where no substantial earnings can be made), but they can customise tariffs suited to each customer's particular consumption pattern (where significantly different from the standard pattern used to establish the regulated tariff) and also to provide consumers with incentives to time their power use more efficiently.

9.4 Metering

Metering was not an element of concern in the regulatory discussion until rather recently. The unbundling of retail activity that followed in the wake of industry restructuring in some power systems (see next section) and the recent development of metering hardware technology have triggered a debate on metering in which the industry has yet to reach a consensus.

[28] In Texas, the SB7 bill defined a transitory phase during which incumbent electricity companies could not offer their customers a lower price than the one defined by the regulator, in order to prevent predatory practices and to allow new market entrants to become established.

In traditional systems, metering was one of the duties performed by the vertically integrated utility, and in the case of distribution companies, part of their duties as regulated retailers. At the same time, consumer demand was only measured on an aggregate basis, using monthly (or sometimes quarterly or even annual) values as the indicator for charging customers for the electricity consumed. These principles continue to be in place in many systems that have opened up their wholesale market to competition but maintain oversimplified regulated rates.

In this new context, the two main issues can be roughly summarised as set out below.

- Should metering be unbundled from the distribution activity? And if so, should it continue to be regarded as a regulated activity or should competition be allowed at this level? Who should own meters: the distribution company, the consumer, the retailer, none of these? Who should be responsible for their management: the distribution company, the retailer or another organisation altogether?
- How should regulation deal with these new advanced meters? Should metering equipment renewal be compulsory? If so, what minimum standards should be established?

These two regulatory issues are discussed below.

9.4.1 Liberalisation of the Metering Activity

Metering has been traditionally considered to be a part of network operation. Electro-mechanical metering equipment was regarded as the obvious technology for small consumers and, since no improvement was foreseen, metering was regulated as a monopoly business. However, power system restructuring, allowing explicit nearly real-time prices for consumers (each hour or each quarter of an hour), and technological developments have opened the door to competition.

Two of the very few (at this writing) examples of power systems where metering has been liberalised are the UK and Germany. In the latter country metering was deregulated in 2008, the distribution company remains as the "default" metering operator, subject to a regulated tariff, and consumers can choose their operator and, in particular, the type of meter desired. In the UK, price controls have been restricted to the equipment installed by distribution network operators prior to 2007 [40].

Liberalisation of the metering activity raises a number of significant questions related to the potential impact on retail market development. Renewing metering equipment involves incurring significant initial costs. When the retailer owns the meter, there is a certain risk that it might turn into a barrier for the consumers to switch providers, since changing the equipment might not be straightforward nor cheap. At the same time, certain clauses (such as minimum-term requirements, i.e. a permanency contract) allowing the retailer to hedge against the risk of incurring

a stranded cost if the consumer decides to switch soon after the new equipment is installed might also turn into a barrier of this type. This is why metering in many systems is still considered to be a regulated activity unbundled from retailing, at least as far as installation, maintenance and reading are concerned. Data management beyond the calculation of billing items will necessarily be incumbent upon the retailer, although data protection and data unbundling are also issues that need to be carefully addressed.

If this activity is ultimately liberalised, a minimum level of standardisation is required to prevent the erection of barriers to changing providers.

9.4.2 Advanced Meters and Smart Demand Management

Because the price of electricity varies with the time of day, the amount of electricity used each month should not be the only criterion for charging for this service. Two customers might well consume the same total volume of electricity, but one primarily at off-peak hours when wholesale market prices are low and the other mostly at peak hours when prices are high. It seems unfair not to distinguish between these two types of consumers.

The obvious solution to this problem is to install meters that measure consumption every hour (or every quarter of an hour, or even less), thereby providing for a more accurate computation of the cost of electricity. The dilemma is whether the expected improvements, because of the enhanced demand response, are worth the effort and expense of installing advanced meters for all consumers. This is a complex decision for utilities and regulators, as many issues are involved: costs, efficiency gains, other potential uses (e.g., adding emergency demand control), potential obsolescence of the meters given the speed with which technology advances and the efficacy of these meters in educating consumers as part of current efforts to achieve a meaningful demand response to price signals.

Advanced meters are the key component in most demand response management systems. More specifically, they are indispensable when implementing time-based pricing tools.

Most countries have allowed or required large-scale consumers to install interval meters to manage their power consumption more efficiently, resulting in savings for both the consumer and the system as a whole. The next and much more ambitious step is to extend this feature to residential consumers.[29] Any progress in this regard obviously entails replacing the old electro-mechanical meters.

Electro-mechanical metering equipment has barely evolved in the past 50 years. In principle, it offers a substantial competitive advantage: it is already in place. These meters have significant drawbacks, however, some of which are listed below.

[29] In the EU, Italy and Sweden were the first two countries to opt for a compulsory roll-out.

- Reading costs are high, as all meters must be read in situ and accessibility to such meters varies widely from one location or country to another. In some the meters are located inside or outside each home, while in others they are installed in specific meter rooms, typically one per building. When meters cannot be read, consumption must be estimated from historical records.
- Temporary price signals cannot be sent to the end consumer; as residential electro-mechanical meters only record cumulative power consumption, the power demand at different time intervals cannot be billed separately.[30]
- Information on each customer's consumption profile is lacking, and this hampers retailer planning and their ability to individually advise end consumers.

Electronic meters provide solutions to the above shortcomings and make new alternatives possible.

- Retailers would have the infrastructure required to broaden the variety of their offerings and customer services, which would contribute to increasing the efficiency of market signals and to driving market competition.
- In some countries (England, for instance), special conditions are in place for the fuel poor, i.e., households spending over 10 % of their income to keep their homes reasonably warm [8]. These conditions consist of pre-paid tariffs that simplify economic management in such cases and minimise the credit risk to which retailers would be exposed (such provisions go hand-in-hand with certain financing facilities[31]). The problem is that these tariffs call for the installation of a special meter whose maintenance significantly increases costs (and, therefore, the power bill). New and more advanced meters, more highly developed and initially less expensive, might palliate such problems [38].
- The use of advanced meters may encourage the residential sector to participate in distributed generation (with solar panels, small wind generators and so forth). Advanced meters are a necessary component in such facilities, for the power generated and consumed must be measured by the minute [38].
- In addition to remote management (which may entail substantial savings), distribution companies would be able to accurately measure the quality of supply at the connection point. Calculations of other parameters such as loss coefficients or reactive power would also be more accurate. These new meters simplify the detection of grid faults and reconnection after power outages.

In light of this new world of functionalities and expectations, why are residential customers still using electro-mechanical meters? Beside the fact that there

[30] Billing is possible, however, if a clock is installed along with several electro-mechanical meters, or with a single meter fitted with a number of integrators. This was the arrangement used for Spain's "night-time tariff", for instance. The problem is the high cost involved (not only of installing the new devices, but also of taking more complex on-site readings).

[31] See www.consumerfocus.org.uk.

is still no clear evidence that making this switch would be cost-effective for small consumers, other factors such as technological immaturity and uncertainty about actual demand elasticity have cast doubts on the wisdom of the change.

9.4.3 Unbundling "Smartness" and Metering

As briefly summarised below, electric power system restructuring, which has resulted in the implementation of real-time prices in wholesale markets, has also boosted the development of advanced electronic meters and smart demand management solutions based on this equipment, commonly referred to as "smart meters". From the regulatory standpoint, it would be essential to clearly distinguish and unbundle the two features: metering, i.e., the ability to measure consumption more accurately, and "smartness", i.e., the number of smart and active demand management actions that can be performed when quantifying the actual value of consumption at any given time.

The most suitable approach would therefore be to decouple the two functions, reducing and simplifying the capabilities of the metering device to a minimum (namely, limiting them to the ability to measure consumption in reduced time blocks, facilitating data storage[32]). Price signals, consumer strategies to respond to them and the appropriate management commands can and should be the domain of another device (sometimes called an energy box). Consumers could communicate with their retailer and home appliances could even be managed through wireless communication via the supplier's website. The endless possibilities and flexibility of this approach is likely to lead to its prevalence in the medium term and its generalised use in the long term.[33] This approach reduces the costs of rolling out advanced meters and leaves the development of innovative ideas to the market, minimising future barriers related to the need to install metering equipment.

9.5 Energy Conservation and Efficiency[34]

> Energy efficiency is the antithesis of what we have come to expect from government initiatives: it is long-term, dull and effective. The Financial Times, Friday, September 12, 2008.

[32] Automatic meter reading (AMR) devices can be read remotely and also provide time-of-day information. The Federal Energy Regulatory Commission's definition (2010) clearly follows along these lines: 'Advanced Meters: Meters that measure and record usage data at hourly intervals or more frequently, and provide usage data to both consumers and energy companies at least once daily. Data are used for billing and other purposes. Advanced meters include basic hourly interval meters, meters with one-way communication, and real-time meters with built-in two-way communication capable of recording and transmitting instantaneous data.'

[33] See, for instance, www.google.com/powermeter/about/.

[34] This section has been fully written by Ignacio Pérez-Arriaga. The author is grateful for the comments received from Pedro Linares and Carlos Batlle.

A business area that is closely related to retailing and that would require a chapter of its own is efficiency and conservation in the use of electricity. A brief review of the regulatory aspects involved will be presented here. Given the possibilities of utilisation of electricity for heating and transportation, and the potential of energy service companies and the incumbent utilities to provide broader services than strict electricity supply, it is unavoidable that regulatory measures for energy conservation and efficiency (ECE) in electricity go beyond the boundaries of the power sector, and conversely.

ECE is the largest and most effective approach to reducing the environmental impact of the energy sector—also of the power sector—. The Intergovernmental Panel on Climate Change, in its Fourth Assessment Report [24], estimates that 7–14 % of the global greenhouse gas emissions might be saved with negative cost measures, most of which include ECE. The IEA 450 Scenario [21],[35] which aims to limit greenhouse-gas emissions to 450 parts per million in the atmosphere, identifies energy efficiency policies and measures as the cheapest abatement option available and the most important source of abatement of CO_2 emissions. Efficiency is responsible for half of cumulative global abatement relative to the New Policies Scenario, or 73 Gt, between 2011 and 2035. These efficiency improvements will come from all consuming sectors, including buildings, appliances and equipment, lighting, transport and industry.

Energy efficiency is defined as the increase in the efficiency with which energy is used to provide a given amount of a product or a service, measured in units of output per energy unit. On the other hand, energy conservation is the absolute reduction in energy demand compared to a certain reference situation; it is measured in energy units and can be achieved through improvements in efficient energy use or a decrease in the demand for the products or services that use energy.

Utility is derived from the products and services that require energy to provide them, not from the energy itself. Therefore, it is possible to provide the same level of services and products with a lower consumption of energy. Note, however, that most policies have the objective of improving energy efficiency, which does not necessarily result in energy conservation—or not as much as expected—because

[35] The 450 Scenario is an outcome-driven scenario, illustrating a global energy pathway with a 50 % chance of limiting the increase in the average global temperature to 2 °C. This would require the long-term concentration of greenhouse gases in the atmosphere to be limited to around 450 parts per million of carbon-dioxide equivalent (ppm CO_2-eq). The New Policies Scenario, which takes account of both existing government policies and declared policy intentions, would result in a level of emissions that is consistent with a long-term average temperature increase of more than 3.5 °C. The outlook in the Current Policies Scenario, which assumes no change in government policies and measures beyond those that were enacted or adopted by mid-2011, is considerably worse, and is consistent with a long term temperature increase of 6 °C or more.

of the rebound effect.[36] Efficiency increase should not be the final objective of ECE policies, which must be targeted to obtaining actual savings in energy use.

The experience has shown so far that the successful deployment of ECE measures critically depends of well-designed regulation. A comprehensive regulatory approach may include the following policy instruments [27, 30]: (a) establishment of minimum performance standards; (b) removal of barriers to market-based ECE measures; (c) taxes; (d) incentives to improve efficiency and to save energy; (e) market-based mechanisms and (f) information policies. These regulatory instruments can be applied at electric utility level—more generally, at electricity services company level—or at end consumer level.

Minimum performance standards are meant to make sure that appliances, buildings, processes and equipment are compliant with some basic efficiency requirements that will become tougher with time, as the technologies evolve. Standards can also be targeted to achieve energy savings; this would be more effective, but it is not frequent. Standards are easy to implement and they are politically attractive, since their cost for the consumer is opaque. However, they cannot be adapted to the different characteristics of all consumers. Utilities might also be subject to mandatory requirements, such as to prove that ECE measures are not discriminated with respect to new investments in generation or networks. This enhanced planning approach was common in the US during the late 1980s and 1990s under the name of Integrated Resource Planning (IRP) [50].[37] IRP has been an accepted way in which utilities can create long-term resource plans. As the electric industry began to restructure in the mid-1990s, however, IRP rules have been often repealed or ignored. Tradable white certificates (TWCs) and energy

[36] Because of the rebound effect, an increase in energy efficiency in general does not result in a proportional reduction in energy demand. There are several reasons for this. An improvement in the energy efficiency of providing a product or service: (a) decreases its cost and thus also its consumption since demand is price elastic; (b) increases the available income to consume other products or services that also use energy; (c) under a macroeconomic perspective, relative reductions in energy efficiency and the associated prices of the different sectors may favour energy intensive sectors. Obviously the practical importance of the rebound effect is very context dependent, but it has to be taken into account when designing ECE policies, since it reduces their effectiveness and increases the cost of actual energy savings. Linares and Labandeira [27] discuss this topic in some detail.

[37] The US federal government defined IRP in the 1992 Energy Policy Act (text available at: http://www.ferc.gov/legal/maj-ord-reg/epa.pdf): "The term integrated resource planning means, in the case of an electric utility, a planning and selection process for new energy resources that evaluates the full range of alternatives, including new generating capacity, power purchases, energy conservation and efficiency, cogeneration and district heating and cooling applications, and renewable energy resources, in order to provide adequate and reliable service to its electric customers at the lowest system cost. The process shall take into account necessary features for system operation, such as diversity, reliability, dispatchability, and other factors of risk; shall take into account the ability to verify energy savings achieved through energy conservation and efficiency and the projected durability of such savings measured over time; and shall treat demand and supply resources on a consistent and integrated basis."

efficiency resource standards for utilities have a mandatory, standard-like component, but they have been classified under "market-based instruments" below.

Only by *removal of the key existing barriers* can market-based mechanisms be successful in ensuring a substantial deployment of ECE measures. From the end consumers' viewpoint, typical barriers for the acquisition of efficient appliances or for refurbishing inefficient buildings are the high upfront costs and lack of information. From the electric utilities' perspective, given that their remuneration frequently depends on the amount of supplied electricity, they will resist any ECE measures that have the potential of eroding their income.

Taxes may be used to account for externalities that are not included in the price of electricity (see Chap. 11). In theory, taxes will increase the electricity prices and will have a positive impact on ECE; however, if the price elasticity of electricity consumption is small the effect of taxes will be also small. Taxes are transparent and increase the price of electricity; therefore they are unpopular and politicians try to avoid them.

The adoption of ECE measures by both supply companies and end consumers can be accelerated by adequately designed *incentives and subsidies* (e.g., tax credits, direct payments or ear-marked innovation funds). On the one hand, these incentives must compensate the supply companies for the incurred costs and the loss of income that may follow the adoption of ECE measures by their customers, as well as to develop innovative schemes to improve their own efficiency. On the other hand, they must mobilise consumers to adopt these ECE measures. These measures are typically quite effective, very popular and they are widely used. When these measures reduce the price of electricity, they promote the rebound effect.

Market-based mechanisms. In the US the Energy Efficiency Resource Standards (EERS) establish long-term energy savings targets—a percentage reduction in energy sales—that utilities must meet by means of customer energy efficiency activities. EERSs are meant to achieve sustained investments in energy efficiency. The long-term goals of EERSs send a clear signal to market actors, creating a level of certainty that encourages large-scale investment in cost-effective energy efficiency.[38] In the EU the TWCs are a comparable instrument to the US EERS. The TWCs also set an absolute reduction target for energy demand, and then allow trading this obligation among the concerned agents by means of tradable certificates. TWCs have been implemented in Italy, France and the UK. As with other incentive schemes involving an economic compensation to demand response (e.g., capacity mechanisms), in both EERSs and TWCs it is cumbersome to define the reference or baseline situation with which the actual demand response has to be compared.

[38] As of September 2012, twenty-four US states have fully-funded policies in place that establish specific energy savings targets that utilities or non-utility program administrators must meet through customer energy efficiency programmes. See http://aceee.org/topics/eers for updated information and [16] for a thorough review of demand response experiences.

Information policies try to reduce or eliminate some barriers that may significantly reduce the effective response of consumers, for instance through energy labelling and expectations about price changes and savings.

ICER [20], Michaels and Donnelly [33], EnerNOC [10] and IEA [22] are references of interest to search for additional information on this topic.

9.5.1 ECE Instruments Addressed to the End Users

Despite the improvement in energy efficiency and dematerialization of the economy, the demand for energy services and products keeps growing. This is obviously due to the increase in population and living standards, but there are other less apparent factors at play. One is the rebound effect, which has been discussed previously in this section. Here we shall focus on why clearly beneficial ECE measures have not been widely implemented, or adopted by the end consumers, despite their apparently large socio-economic benefits. Why the low hanging fruit, even the fruit that is lying on the ground, is not picked-up?

The low investment in ECE is likely to be related to market failures, but also to a lack of consideration of behavioral aspects [45]. Linares and Labandeira [27] present a summary of the main findings regarding this issue and provide a list of the major reasons that may explain an investment in ECE lower than expected: (a) low energy prices, typically because they do not include all external costs; (b) investment costs that are higher than theoretical studies had anticipated; (c) uncertainty (because of the future price of energy) and irreversibility of ECE-related investments; (d) failures in the information to consumers; (e) lack of economic rationality ("bounded rationality") on the part of the consumers, further induced by the fact that the implications of adopting ECE measures on the consumer's budget are generally not significant, and this leads to give more importance to upfront costs; (f) slowness of the technology diffusion process of new ECE technologies; (g) the "principal-agent problem", i.e., the fact that frequently the one paying for the investment does not coincide with the one benefiting from it; (h) difficulty in accessing adequate financing; (i) the heterogeneity of consumers prevents that a single product or measure suits all of them and (j) the divergence between social and private discount rates, i.e. consumers want to see a shorter recovery period of their investment than what is socially optimal. This is the justification for the regulatory support of the ECE measures that have been described previously.[39]

Quoting Sioshansi [45]: "We cannot rely entirely on changes in the supply-side of the equation to reduce the industry's carbon footprint. Changes in the demand-

[39] The European Commission estimates that EU consumers could save up to €13 billion per year if they switched to the cheapest tariff available. This potential is currently untapped, as many are still not fully aware or able to make full use of the opportunities created by the market, as only one among three EU consumers compares offers.

side as well as changes in energy consumption habits and, perhaps more profoundly, lifestyles changes may ultimately be needed to address the carbon problem."

9.5.2 ECE Instruments for Utility Regulation

Electric utilities are the usual vehicle that regulators have used to implement ECE measures that look after improvements in efficiency and energy savings in both the utilities themselves and in the end consumers. However, it is not clear that this should be the case, since electric utilities may have perverse incentives to cooperate with ECE programmes. Here it is necessary to distinguish between—on one extreme—vertically integrated utilities that are remunerated under the traditional cost-of-service framework, and—on the other extreme—power sectors that have transitioned towards an unbundled structure and a regulatory regime that has opened generation and retail to competition.

Traditionally-regulated vertically-integrated utilities are responsible for the centralised planning of all activities that are necessary for electricity supply. In this context, it makes sense to mandate that conventional planning should become IRP, so that ECE measures could compete on a level playing field with supply-side measures. Then, the customary cost-of-service procedures, with some adaptations, can be used to properly remunerate the incurred costs of ECE activities. Moreover, under this scheme, if the electricity rates are not reviewed frequently, the utility may decide to save costs by adopting ECE measures that avoid incurring into more expensive planned supply-side measures. This is a correct incentive for the utility, and the regulation should seek that a fraction of the savings can be shared with the consumers.

There are, however, some shortcomings in the traditional approach with respect to the adoption of ECE measures. In the first place, the remuneration of the vertically integrated utilities is sometimes determined with too basic procedures that are based on the total amount of electricity supplied, therefore creating a perverse incentive for the implementation of ECE activities. Second, electricity consumers are typically mostly subject to volumetric charges (i.e. €/kWh charges per unit of consumed energy) rather than capacity (i.e. €/kW charges per unit of contracted capacity or in proportion to peak demand) or fixed annual charges. Therefore, in the absence of ex post corrections in the next price control, any activities that reduce electricity consumption represent a loss of income for the utility.

The problem becomes more acute under regulation that is open to competition, because of unbundling and the absence of centralised planning. IRP is not possible anymore, since electricity production has been liberalized and neither the existence of an incentive of vertically integrated utilities to replace supply-side measures by less expensive ECE measures. Retailers want to sell more electricity and distribution networks are frequently remunerated in some proportion to the volume of

distributed energy, while in most power systems they collect their income from the consumers predominately through volumetric network charges (€/kWh of consumed electricity). Regulatory incentives for distributors and retailers to engage into ECE activities may try to reduce or overcome the natural trend of electricity suppliers to sell as much electricity as possible.[40]

In the US it has been developed and implemented a regulatory approach named *decoupling* that tries to neutralise the incentive of distributors-retailers[41] to increase their income by selling and distributing more electricity [36]. Decoupling can be defined generally as a method to separate revenues and profits from the volume of energy sold and, in theory, makes a distribution utility indifferent to sales fluctuations. In practice, it is rare that a mechanism fully decouples sales and revenues. Mechanically, decoupling trues-up revenues via a price adjustment when actual sales are different than the projected or test year levels. As the remuneration of the distribution company is not calculated based on the MWh usage of networks, the network company's revenues do not depend on the amount of electricity sold to customers. If the revenues collected from the customers deviate from the allowed revenues, the difference is collected from or returned to customers through periodic adjustments or reconciliation mechanisms. If a successful energy efficiency program reduces sales, there will not be any loss in revenue resulting from these energy efficiency programmes. If sales turn out to be higher than the projected, the excess revenue is returned to the ratepayer.

Decoupling significantly mitigates the perverse incentive problem, but it does not help in determining the right remuneration for distribution companies when they need to incur additional costs to enhance the automation of their networks, because of connection of distributed generation or electric vehicles, or implementation of active demand response, including ECE measures, as it was discussed in Chap. 5.

Ideally, regulation should encourage that electric utilities transition from being "electricity providers" to being "energy services providers", therefore adopting most of the ECE measures in a natural way. We have seen that this is not that simple, because of the existing procedures of remuneration of these companies and how the consumers pay for the electricity that they use. The pressure towards ECE, as well as the appearance of innovative regulatory schemes—such as the "white certificates" in Italy, UK or France, or the "energy efficiency resource standards" (EERSs) in many states in the US—, have opened the door to the so-called *energy service companies*, sometimes associated with the distributors and retailers themselves but also independent, which offer a host of measures to meet the demand reduction objectives. This is a promising area of activity where new business models will undoubtedly emerge.

[40] A parallel topic is the reduction of losses in distribution and transmission networks. This chapter only covers energy conservation and efficiency at the end consumption level. The regulation of network loss reduction can be found in Chaps. 5 and 6.

[41] In the US only in the state Texas it is mandated that the activities of distribution and retailing have to be unbundled.

9.6 Other Retail-Related Regulatory Issues

In addition to the major issues described in this chapter, a whole bundle of small but complicated regulatory decisions must be made when liberalising the retail business. A few are listed below, by way of illustration, and as a conclusion to this chapter.

- Just as consumers are protected against retailer failure, should consumers also be expected to provide security to enable retailers to hedge against defaults? If so, how should such security deposits be designed and managed, and under what circumstances could they be applied to unpaid bills?
- Further to the preceding item, who bears the cost of non-payment? Should the distribution company bear part of it in proportion to the network access tariff? Conversely, if a consumer's contract with a particular retailer expires and the consumer fails to enter into a contract with any other supplier, who bears the cost of default?
- When is consumer disconnection for non-payment justified? Since service interruption is not always immediate, who should bear the cost of consumption in the time lapsing from the date of the retailer's request for disconnection and the distributor's implementation of that request? What incentive would the distributor have to disconnect at short notice if delays involve no cost?

These are just a few examples of the many regulatory issues that may arise. As the retail liberalisation process unfolds, new challenges will no doubt ensue.

Annex A
Retail Processes

This Annex is based on the material prepared by Temboury et al. [46] and expands what is presented in Sect. 9.1.3 of this Chapter. We start from the same Fig. 9.1 in that section, now Figure A.1, which classifies the processes and relationships that constitute the business of retailing.

In the electricity retailing companies, as a general criterion, these sub-processes are handled by different business areas. Next, based on the different sub-processes, an overview of the scope and nature of these areas is provided.

A.1.1 Sales

The sales area is the marketing "face" of the company towards the customers. Figure A.2 illustrates the tasks that relate to the activity to be developed.

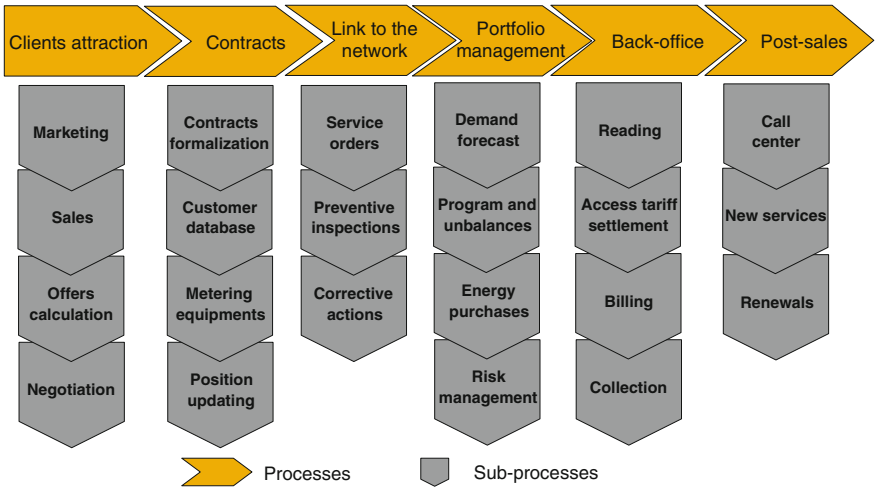


Fig. A.1 Retail activity processes

Positioning and commercial management

The commercial position of the company is based on the following pillars.

The brand

In every market research, the brand is listed as an essential element. It has to

- be close and credible, transmitting confidence, security, ability to provide good service to consumers,
- allow differentiation of a product (electricity) that is so difficult to differentiate and
- allow to charge a premium over competitors with a weaker brand.

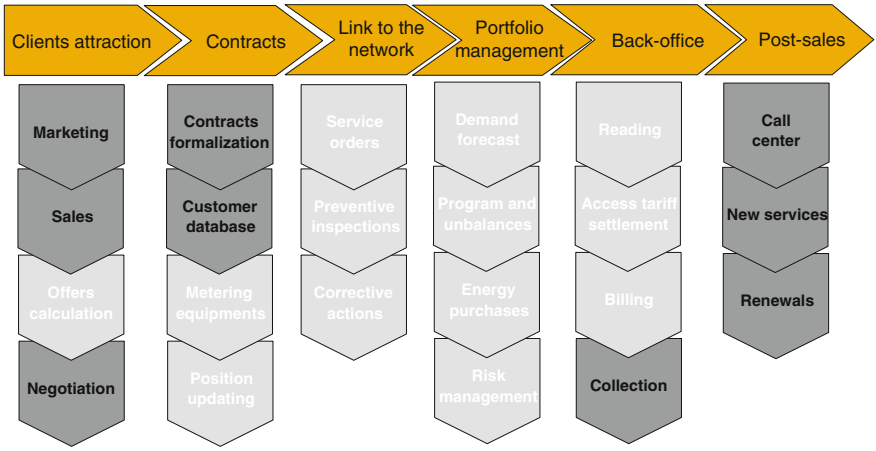


Fig. A.2 Sales sub-processes

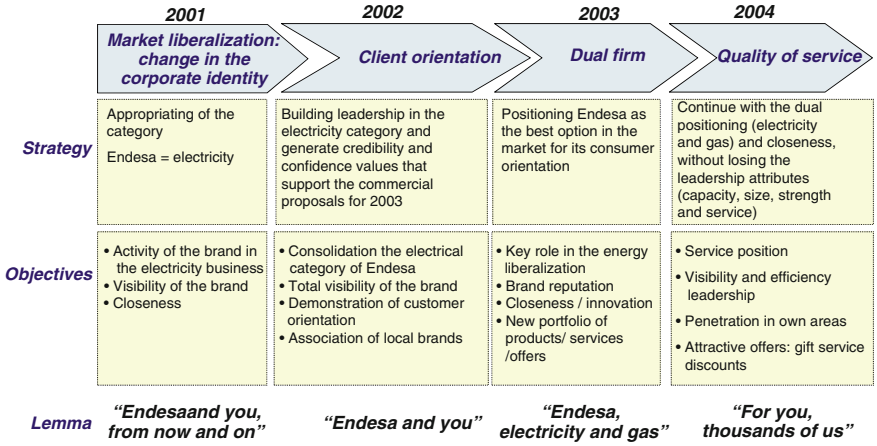


Fig. A.3 Example of brand positioning evolution

Figure A.3 shows an example of the evolution of the brand positioning of a retailer in an electricity market.

Customers' segmentation

The detailed process of customers' segmentation has the objective of revealing their values and needs. Not all customers value each service in the same way, which makes it important to categorise the priority of the different subprocesses depending on the considered client. Figure A.4 shows the findings of a study conducted in 2004 that illustrates the diversity of the sensitivities of the different types of customers.

All customers assign a high priority to the quality of supply, in particular to the continuity of supply. In case of a blackout, customers give great importance to a swift recovery, as well as to the availability of information about the causes and the current situation of the service restoration process.

Quality and operations

The care of the quality of service provision and hence the proper management of operations, is a key to boost loyalty and attract new customers.

An important point to keep in mind when investing in improving the quality of service is to know the sensitivity of the customers from any inconvenience that may arise. Figure A.5 illustrates the results of a customers' survey of an electricity and gas retailer in the Spanish market.

Loyalty

The ability of keep the customers' loyalty to the brand is a key aspect in the smooth running of the retail business since, in most cases, it is cheaper than attracting new customers. By means of specific offers, the company tries to differentiate itself from its competitors, while maintaining its flexibility.

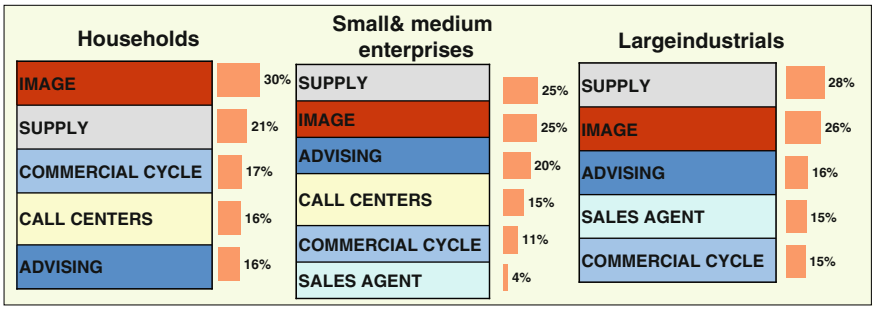


Fig. A.4 Valuation of the product "electricity supply" by customer type

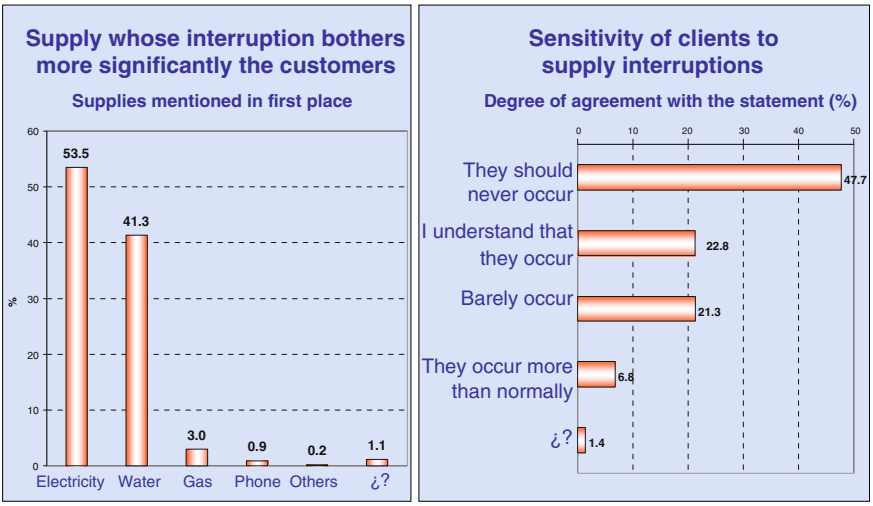


Fig. A.5 Sensitivity of customers to quality of service

In line with the discussion above, the first and obvious way of strengthening loyalty is the provision of good service. In addition, after-sales service plays a key role. Some further strategies are:

- Personalized attention to large customers. In these cases the ability of the sales agent is very important, when trying to create a relationship or identification of the customers with the brand.
- Provision of energy services and other added values.
- Loyalty programmes: loyalty cards for customers, customers' club, bonus programmes, etc. They have the additional value of easing the communication between the company and the customer, and of providing additional information on customers. In contrast, the fact that their cost is not negligible has to be taken into consideration.

Communication channels and customer service

The channels to contact with customers must be efficient and adapted to customer requirements. They differ depending on the country's culture. For example, in the UK "door to door" works for residential customers. In Spain, this alternative does not seem to provide the same outcome (especially considering its high cost). Common channels to reach residential consumers are:

- Advertising (TV and press).
- "Call Center" to conduct direct marketing (more effective if used together).
- Own offices or franchises.

Customer database and new service agreements

Concluding a new service agreement is the process in which consumers cease to pay the default rate, or pay a specific supplier, and start buying their electricity from another supplier. Administratively speaking, this process begins when the customer signs an energy purchase contract with the second supplier and entails a series of subsequent follow-through actions.

Given that one of the key components in this process is the list of all customers and their respective suppliers, there must be an unequivocal mechanism for identifying each point of supply. The first key decision to make when designing a customer database is whether it should be centralised, developed by an independent administrator, or administered by a local body, normally the individual distributor, under terms that ensure that all such lists are duly co-ordinated. In Australia, NEMMCO—the wholesale market administrator—is the centralised agency responsible for customer lists. In England and Wales, the settlement system was wholly centralised between 1990 and 1998, but since 1998, local suppliers have been responsible for compiling and entering customer information. This change was partly prompted by the high costs of the former arrangement. In Norway and PJM (Pennsylvania-New Jersey-Maryland in the USA), the responsibility for customer lists lies with local distribution system operators.

Another important database design issue has to do with its contents. If a database, which has been generated with the information provided by local suppliers, contains information on all the customers connected to the system, then, when a customer changes supplier, all that is added is the new supplier's name. By contrast, if the only information contained in the database is the data supplied by competitive suppliers, the first time that a change of supplier takes place will require a much longer and more complex process, because the customer's detailed data will have to be compiled and entered into the system first. This results in huge problems when droves of customers simultaneously try to change supplier for the first time, which has been usual in the early stages of liberalisation. However, the key advantage of this second approach is that it minimises the preliminary work required to start up a new market, which may be crucial for effective industry liberalisation.

When a customer concludes a service agreement with a new supplier, the service provider must notify the entity responsible for updating the database accordingly. In some systems customers are required to notify the database

administrator directly, primarily to prevent supplier malpractice, such as keeping a list of customers who have not concluded a service agreement with them, potentially causing inconveniences to these consumers. The initial stages of retail market liberalisation are especially susceptible to such abuse, in light of the particularly aggressive campaigns conducted at such times to attract new customers.

Such practices first appeared in the context of the US telecoms business: for the sake of operational efficiency, the regulatory authorities ruled that customer signatures would not be required to initiate the change to another telephone carrier. Therefore, there was no way of informing customers of the change a priori, and substantial abuse was committed in the form of unauthorised changes. Similarly questionable practices have been associated with door-to-door sales. In the United Kingdom, although contracts signed for an unsolicited service can be cancelled without penalty within 7 days of conclusion, such sales are still fraught with a fair amount of fraud. To avoid such problems, suppliers have been mandated to institute confirmation procedures (conducted by a different employee) to ensure that customers actually do want to switch suppliers. In any event, in nearly all systems where retailing has been liberalised, suppliers are held responsible for handling the administrative formalities, to save customers the inconvenience of having to do it themselves.

In some countries, Spain among them, the process of changing suppliers also involves the conclusion of a contract between the new supplier and the respective distributor. The terms and conditions of such agreements, which are regulated, specify that the supplier must pay the respective distribution charges (see billing below). The metering equipment is likewise checked on this occasion, to ensure it is regulation-compliant.

Meter readings may also be a compulsory part of certain supplier change processes. Depending on the meter reading cycle of the country in question, a special reading may, or may not, be required as part of the change process (if not, the change does not become effective until the scheduled meter-reading date).

Once the processes that relate to the customer have been discussed, we shall address the processes that correspond to the acquisition of the energy by the supplier: the estimation of the future demand of the customers; contracting and purchasing in the energy wholesale market—including portfolio risk management, which involves taking hedging positions the medium and long term, usually through financial derivatives—and determining the price to be charged to the client and the format of the corresponding contract (e.g., a constant price, direct pass-through of the wholesale price, etc.).

These processes are developed in the areas of *front*-, *middle*-and back-office. The front-office is responsible for the procurement of energy and for portfolio management in the short term (balancing deviations, i.e., the differences between the quantities previously acquired and the actual consumption of the customers' portfolio). The department of *middle-office* (which could be the area of risk management) manages the portfolio in the medium and long term, looking at the

overall position in the different periods, making the appropriate hedges and giving signals to the sales department on the range of offers that may pose to individual customers and to the front-office area to guide its processes, passing on the positions that remain open (i.e. the energy that has not been previously contracted in the long term). Finally, the *back-office* department (settlements) is responsible for closing all operations in different markets.

A.1.2 Front-Office Area

The *front-office* department buys in the market the energy that is necessary to meet the needs of customers (Fig. A.6).

First, the front office has to provide the estimation of the short-term consumption (i.e. 1 week) of the total portfolio. This task starts from the client-to-client estimations previously made by the *middle-office* department when evaluating the bids made to individual customers. The *front-office* area refines these forecasts to estimate the purchase requirements in the short term. Both departments must be familiar with the consumer characteristics, which could be very different in volumes and patterns. In the next section we illustrate this process from the viewpoint of the previous calculation to be performed by the middle-office area to calculate the right offers to provide prices to sales managers.

Once the demand for the company portfolio has been estimated, the main task of the *front-office* area is to manage the purchases in the various markets, preparing and submitting bids for buying and selling, with the purpose of minimising the cost of supply.

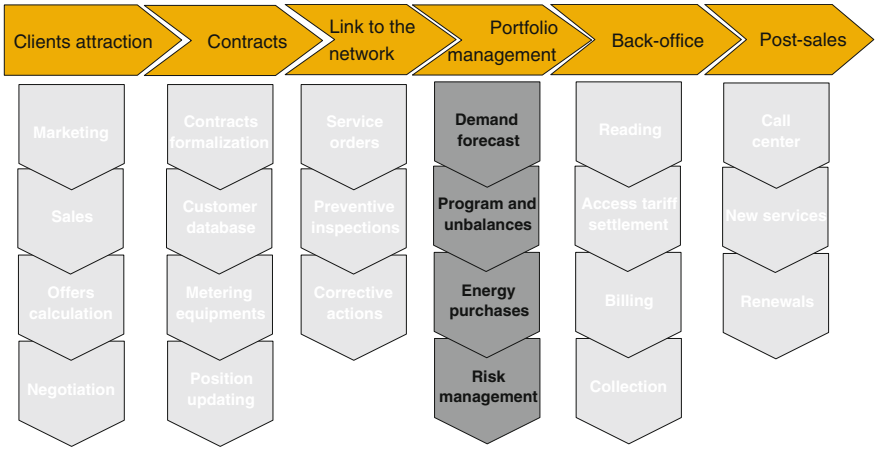


Fig. A.6 Front-office subprocesses

A.1.3 Middle-Office Area

As described previously, this area manages the company's position in all that relates to the quantity and cost of energy for the entire time range. Its goal is to make a first assessment of the costs of providing energy to customers (by calculating the prices and costs of any required processes) and then to optimize the management of the portfolio of energy, by minimising costs and optimising risk.

Load Analysis

Two key issues must be addressed, from the supply-side, when considering the possibility of concluding a service agreement with a specific customer and the terms to be included in the respective contract. The supplier must forecast the most likely price of energy on the wholesale market and must also analyse the characteristics of customer demand—total volume and distribution during the 24 h of the day.

Consumer analysis is generally based on the history of electric power use by the specific customer and other consumers in the same bracket. This analysis is used both to calculate the cost of supplying energy under the terms of the agreement and to analyse its associated risk and secure contracts or other hedging mechanisms to protect the company accordingly. Therefore, any mistakes in this process may occasion hedging errors and lead to undesired risks for the supplier.

The consumption profiles of the customers vary significantly depending on their nature. The Figs. A.7, A.8 and A.9, showing the hourly consumption of three different types of consumers during 25 consecutive days, illustrate this statement.

In conducting such processes, a choice must be made between two different conceptual approaches. The first one entails a thorough analysis of the customer's hourly load curves, followed by the application of exhaustive numerical analysis techniques to obtain reasonably accurate forecasts about future levels of consumption. This includes considering factors such as consumption seasonality,

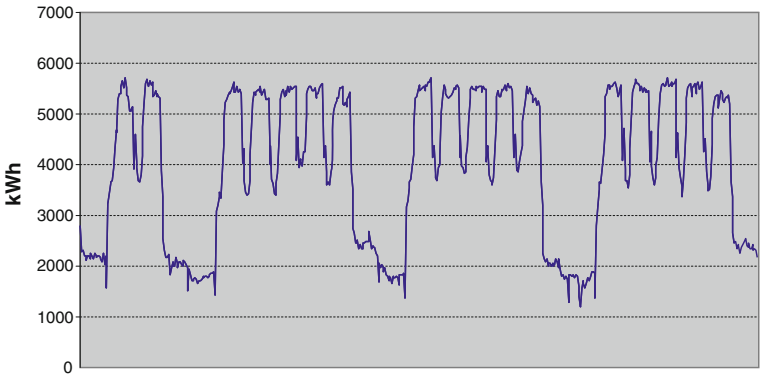


Fig. A.7 Load profile of a car manufacturer

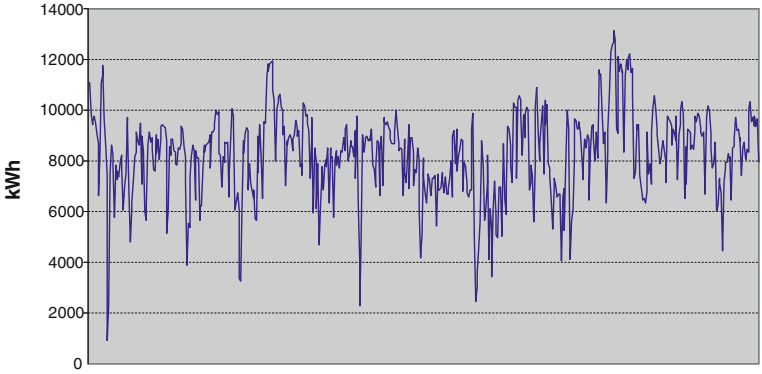


Fig. A.8 Load profile of paper industry

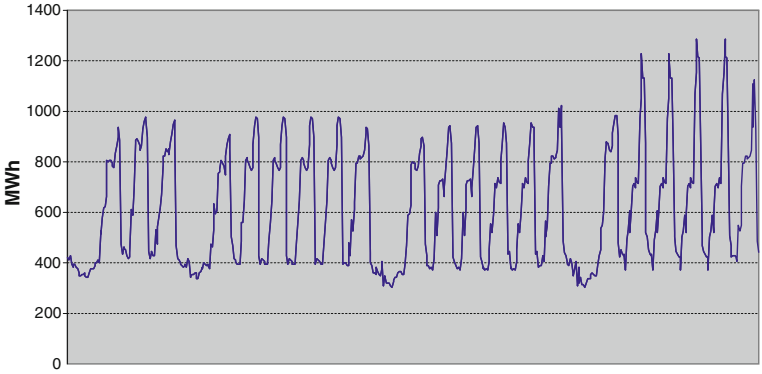


Fig. A.9 Load profile of a mall

changes in customer behavior during vacation periods and public holidays and, more importantly, the stability of the observed profiles. In other words, the analysis provides a measure of the customer's demand volatility with respect to some predefined patterns. An estimate of the risk of actual levels of consumption differing from the forecast can be deduced from such volatility and the economic implications associated with the existing contracts. In the second approach a more superficial analysis is conducted, which is based just on the customers' previous bills. A plot of their expected consumption curve is obtained by simply assuming that their future use will be in line with past patterns.

The choice between these two approaches must be based on an assessment of the required effort, measured in terms of analyst man-hours and computing resources, but also on the judgement about the justification of a very accurate forecast at this stage of the game. Miscalculating a large customer's demand can be very risky, so methods based on the use of detailed consumption curves are required in this case, whereas previous bills should suffice for smaller consumers.

Suppliers often predefine the demand of these smaller consumers by what is known as market segmentation. On the basis of a few basic values (such as the installed capacity and electricity consumed in the previous year) they estimate a given customer's consumption profile for the following year or decide just to simply offer a standard contract.

Cost Analysis

Once the total demand of the customers has been estimated, the supplier can project future energy purchase costs. The information used in this process depends largely on the maturity of the market in question.

The simplest case entails predicting the spot market price for the time periods of interest, on the assumption that all the energy will be bought on the spot market. The energy purchase cost curve can be immediately deduced from the price and consumption estimated curves.

Where efficient and sufficiently liquid futures markets exist, another possibility would be to use forward energy prices instead of spot market forecasts for these calculations. In other words, since such organised markets provide a forward energy price, the most obvious way of estimating energy purchase costs would be to determine the cost of forward purchases of blocks of energy matching the customer's demand profile on futures or over-the-counter markets.

Figure A.10 shows the global process to calculate the energy price/cost in general terms:

- Expected cost of energy: the energy cost of buying in the forward, day-ahead, intraday and secondary (reserves and balancing) markets to supply the expected consumption of the customer.
- Premiums for different risks: risk hedging (insurance) implies a number of additional costs to be internalized in some way when calculating the price of energy. This is not an easy task since, given that these costs are common to the whole portfolio, it is not evident how much to allocate to each type of customer.

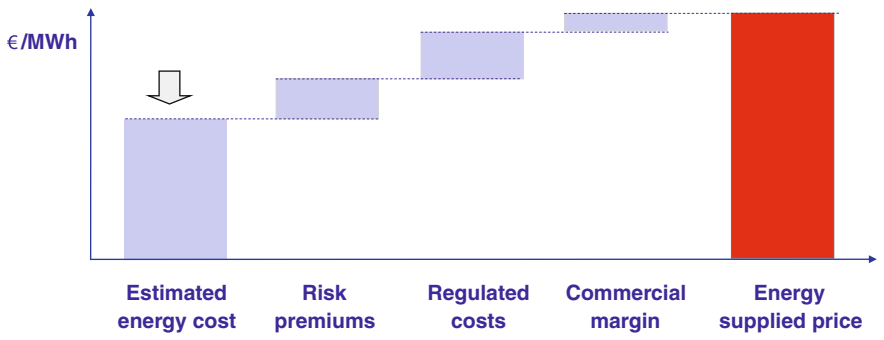


Fig. A.10 Estimation of electricity price/cost

- Regulated costs: the regulated costs that the retailer must internalise and pass-through to the customer, such as access charges, should be included in the estimated cost of the service.
- Commercial margin: this margin is intended to pay for the cost of the supplier commercial structure. The calculation of the value to be taken into consideration when analysing the bid for each type of client must be in accordance with the commercial strategy and business segment to which it belongs (there are customers more relevant and strategic for the company than others, what can lead to the convenience of adjusting the threshold).
- Price for the energy supplied: all this leads to a total price of energy. In the process this total price has been structured into the different commercial products.

Hedging

There are several risks involved in this process, which the supplier needs to hedge against.

Price risk

Price risk results from the uncertainty in the wholesale market prices, resulting in deviations with respect to the estimated energy purchase costs. For instance, assume that a supplier is purchasing the electricity from the spot market and has concluded a contract with a consumer at a fixed price that matches the supplier's best pool price forecast. If the pool price subsequently rises far more than anticipated, the supplier's revenue will remain tied to the contract price, while its energy purchase costs will increase, resulting in a net loss.

The way to hedge against this risk is to conclude agreements with generators that ensure that the energy purchase price remains constant for the estimated volume of demand. In the actual practice of energy markets suppliers are covered by a mix of standardised contracts negotiated on futures markets, bilateral contracts negotiated on OTC markets and generation itself, which may be owned by the company that also supplies the energy. In any event, the supplier must secure cover to reduce the risk of rises in the market price of energy.

Suppliers may also choose not to hedge such a risk and incorporate it instead into their operating costs. Under these arrangements, they charge customers a risk premium,—i.e., a higher price—to compensate for the risks taken. In practice, most of the suppliers' operations are covered by generation assets or contracts, and only a small proportion remains exposed to price risk.

Another alternative is to pass all or part of this price risk on to the customers. This would involve formalising agreements that stipulate that if wholesale market prices diverge with respect to the estimated amount, the consumer must pay for the deviation. Intermediate arrangements are possible, whereby only part of the actual price is passed through to the consumers. Different variations on this theme and different types of contracts are discussed later in this unit.

Quantity risk

Quantity risk refers to the possibility of substantial differences between actual customer demand loads and the demand forecast, either in volume or in the anticipated demand profile. Since wholesale market prices differ by the hour, changes with respect to the estimated load shape will have an economic impact.

This problem is often addressed by simply designing service agreements that obligate customers to absorb this risk. For example, if the rate structure is designed to distinguish between a specified number of time-of-day segments, each one with a different price, when a consumer's consumption profile changes, this will be conveniently taken care of. On the other hand, if the supplier decides to charge a flat price without differentiating between time-of-day segments, the contract should include a significant risk premium.

Imbalance risk

This is an additional economic risk that is also derived from deviations between the estimated and actual demand. Depending on the regulation of the specific wholesale market, suppliers may be required to present a balanced declaration of electricity purchased and sold at any moment of time during a prescribed time period (e.g., the next day). Any imbalances are typically penalized or they have to be cleared in short-term expensive ad hoc markets.

Some customers, such as distribution companies with significant amounts of distributed generators, pose higher risks. The most common way of dealing with this is to build a risk premium into the contract, because there is no simple way of handling such risks. Contracts can be designed with built-in incentives to ensure that actual operation conforms to programming, but this sort of agreements are complicated and seldom concluded.

Collection risk

Collection risk is the risk of the customer not paying, a problem that is common to any business and addressed in much the same way as in other markets. Broadly speaking, with large-scale customers the supplier must try to assess the likelihood of non-payment and, in the case of problematic consumers, call for bank bonds or some similar sort of security.

Regulatory risks

Finally, regulatory risk refers to the possibility of changes in the market rules by the regulatory authority. This problem is common to other markets, though in electricity markets—partly because the regulatory experience is still very recent—the risk may be higher. It is, in any event, a risk very difficult to hedge and simply forms a part of the business uncertainties assumed by suppliers when they decide to enter the market.

Types of contracts

Since the most significant of all the above risks is indisputably the price risk, much of a supplier's business involves optimising its energy purchases and the

agreements concluded with both the generators from which it buys energy on the wholesale market and the customers to whom it sells electric power, to protect itself against price risk.

From a consumer's viewpoint, one of the key services offered by a supplier is the possibility of hedging against price risk. Many consumers simply cannot buy their electricity directly on the wholesale market because they cannot afford to run the risk inherent in market price fluctuations.[42] To avoid such risks, they will try to negotiate a product with their supplier that allows them to reasonably forecast their energy costs.

Although there is a wide variety of commercial formulas in place, the most widespread (and simplest) models are as follows:

- Flat rate: This type of contract stipulates a flat rate per kWh used by customers. In this case, the supplier absorbs the full risk associated with uncertainty in the pool price. This system is particularly suited to small-scale agents or for large customers with a very constant level of consumption, which they are essentially unable to change, who want to avoid highly complex rate schemes. This is very common among smaller customers.
- Time-of-day: These contracts group the hours of the day into several different segments, each of which has a different price per kWh. Customers able to change their load profile in line with prices find this kind of contract attractive because they can benefit from its flexibility. In this case, the supplier shoulders price risk, but not quantity risk. The hours of the day must be grouped into segments, each comprising a time frame during which wholesale prices are likely to be similar.
- Pass-through: In this case, the supplier merely charges the customer the wholesale market price for each hour, multiplied by the amount consumed during that hour, plus the access charges stipulated by the regulator. In this case the customer shoulders all the risks, while the supplier merely offers an intermediary service (it represents the consumer on the spot market, deals with billing, handles relations with the distributor and so on). This type of contract may suit very large customers who are well acquainted with how the electricity market works and prefer to handle their energy purchases themselves.
- Contract for differences: This is a standard contract for differences between the end consumer and the retailer, for a predetermined load profile and a strike price, which uses the hourly spot market price as the reference price in the contract. Therefore, if the consumer strictly follows the contracted demand pattern, she is completely hedged and she pays the strike price for the demand. However, any deviations from the contracted load profile will be charged or credited at the current value of the spot market price. This type of contract has the double desirable property of hedging for the totality of the forecast consumer demand and sending the correct real time economic signal to the consumer.

[42] Besides, most consumers would find it difficult to deal with all the registration fees and other formalities involved in trading on the wholesale market.

There are, or course, many other types of agreements and much more sophisticated contracts. For example, sometimes the energy prices that consumers pay are indexed to the prices of another market (such as the oil market), so if prices on that market rise, the contract price rises as well. This places part of the risk with consumers. Customers capable of hedging against fluctuations on other markets (because it forms part of their normal business activities, for instance) may prefer to absorb this risk in return for a lower risk premium. Other examples of more elaborate contracts include financial options, pass-through with energy price ceilings and floors and so on.

A.1.4 Back-Office Area

The *Back-office* area completes the commercial process. Figure A.11 illustrates the sub-processes commonly related to this area.

Metering equipment

The regulatory issues related to metering have been already discussed in Sect. 9.4.

Billing

The billing process is a characteristic feature of supply, as it is in traditional systems. It consists of gathering information about the amount of electricity used by each of a given supplier's customers and the applicable rates. This information is then used to calculate how much each customer owes and issue a bill, notifying customers of what they owe with payment instructions and information on the amount of electricity used.

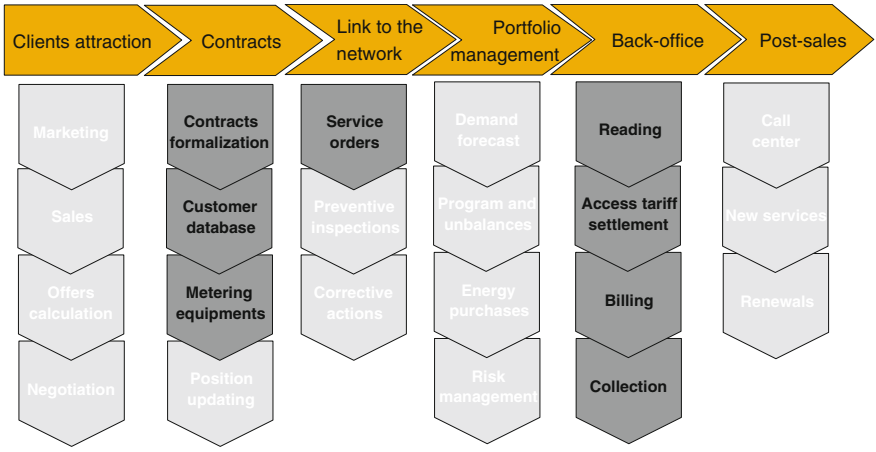


Fig. A.11 Back-office subprocesses

Opening up the retail market to competition and, more specifically, enabling all domestic consumers to change suppliers, can make this process far more complicated. For example, if the price of electric power equals the wholesale market price, a different rate must be set for each hour, which entails handling a much larger volume of information for billing purposes. More dramatically, the existence of hour meters would increase this volume of information almost exponentially.

Such considerations aside, the billing process is similar to the traditional scheme and there is nothing especially unique about it. In principle, the supplier and the supplier alone bills the customer and subsequently pays the respective access charges to the distributor, the transmission grid and so on. Some countries have discussed the possibility of a dual billing system, in which distributors would collect all the regulated costs, while suppliers would only charge for the deregulated services. However, for customer convenience, most systems have a single bill scheme, under which suppliers are responsible for subsequently paying both the distribution charges and any regulated charges.

The billing procedure is based on the measures and conditions of the contract and it is built upon the following stages:

- Arrival of the measure to the commercial system.
- Extraction of the data of the commercial contract.
- Disaggregation of the load profiles in block rates (if applicable).
- Block price application.
- Calculation of corresponding bonuses of surcharges.
- Metering equipment renting addition (if applicable).
- Taxes addition.
- Printing and sending the bill.

Collection
Collection management also works along the same lines as in a traditional system. It includes aspects such as different payment options (direct debiting, transfer, check and so on), bad debt processing and debt collection.

A.1.5 Other Services

Finally, another feature of supply is the existence of a range of supplementary products and services that are very commonly offered by competitive suppliers in addition to electric power per se. These services may involve the provision of some kind of energy-related advice on issues such as demand load factor improvements and recommendations about the most appropriate equipment for a specific customer, or energy savings-related consultant services.

Such offerings have been very common in liberalized power systems, where they are used by suppliers to persuade certain customers to move from regulated rates to competitive supply. In the industrial consumer segment, they have often

adopted the form of the installation of hourly meters or improvement of metering and control equipment.

Multi-utility offerings are still another approach adopted by some suppliers. For example, bundled electricity and gas products, or telephone service-related offerings can be regarded as additional services designed to add value to the primary product.

A business area that is closely related to retailing and that would require a chapter by itself is energy efficiency and conservation. Most countries have established programmes to reduce the electricity consumption or to make it more efficient. Distribution companies and retailers frequently take an active participation in these programmes, sometimes reluctantly (since in the end these programmes try to achieve a reduction in electricity sales or a shift of demand in time to avoid high prices) and some other times willingly (when the regulatory authority offers an attractive scheme of incentives). The current regulatory issues with the incentive mechanisms for these activities have been discussed in Sect. 9.5.

References

1. Assemblée Nationale et le Sénat (2010) LOI n° 2010-1488 du 7 décembre 2010 portant nouvelle organisation du marché de l'électricité. (In French, "New Organization of the Electricity Market Law"). Novemb 2010, www.legifrance.gouv.fr
2. Brennan TJ (2007) Consumer preference not to choose: methodological and policy implications. Energy Policy 35:1616–1627
3. Boschek R (2009) The EU's new competition policy standards. In search of effects-based, economically intuitive or efficient rules?. Intereconomics, Sept/Oct 2009. doi:10.1007/s10272-009-0306-y
4. Champsaur Commission (2009) Rapport de la commission sur l'organisation du marché de l'électricité, April 2009
5. CNE, Spanish National Energy Commission (2011) Boletín mensual de indicadores eléctricos y económicos. (In Spanish, "Monthly bulletin for electrical and economic indicators"). Sept 2011, www.cne.es
6. Conseil de la Concurrence (2005) Avis n° 05-A-23 du 5 décembre 2005 relatif à un dispositif envisagé pour permettre aux industries électro-intensives de bénéficier de conditions spécifiques de prix d'achat de l'électricité
7. Davies S, Waddams Price C (2007) Does ownership unbundling matter? Evidence from UK energy markets. Intereconomics 42(6):0
8. DBERR, Department for Business, Enterprise and Regulatory Reform (2007) Fuel poverty. www3.dti.gov.uk/energy/fuel-poverty/index.html
9. Defeuilley C (2009) Retail competition in electricity markets. Energy Policy 37:377–386
10. EnerNOC (2011) Utility incentives for demand response and energy efficiency. http://www.enernoc.com/our-resources/white-papers/
11. ERGEG (2008a) Obstacles to supplier switching in the electricity retail market. Guidel Good Practice Status Rev (Ref: E08-RMF-06-03), 10 April 2008
12. ERGEG, 2008b. Status Review Supplier Switching Process in Electricity and Gas Markets—Five case studies. Ref: E08-RMF-10-04. 19 September 2008
13. ERGEG (2009) Status review of the liberalisation and implementation of the energy regulatory framework (C09-URB-24-03), 10 Dec 2009

14. ERGEG (2009b) Status review of vulnerable customer, default supplier and supplier of last resort. European Regulators' Group for Electricity and Gas (Ref: E09-CEM-26-04), 16 July 2009
15. Executive Office of Energy and Environmental Affairs (2011) Default service overview. www.mass.gov
16. Fetz A, Filippini M (2010) Economies of vertical integration in the Swiss electricity sector. Energy Econ 32(6):1325–1330
17. Flaim T (2000) The big retail 'bust': what will it take to get true competition? Electr J 13:41–54
18. Frontier Eonomics (2007) Energy costs. Public report prepared for the Independent Pricing and Regulatory Tribunal, March 2007
19. ICER (2010) A description of current regulatory practices for the promotion of energy efficiency, International confederation of energy regulators
20. IEA (2011) World energy outlook 2011, International Energy Agency. http://www.iea.org/
21. IEA (2010) Energy efficiency governance, International Energy Agency. http://www.iea.org/
22. IPART, Independent Pricing and Regulatory Tribunal of New South Wales (2010) Review of regulated retail tariffs and charges for electricity, 2010–2013. www.ipart.nsw.gov.au
23. IPCC (2007) Fourth assessment report, The Intergovernmental Panel on Climate Change. http://www.ipcc.ch/
24. Joskow PL (2000) Why do we need electricity retailers? Or can you get it cheaper wholesale? Center for Energy and Environmental Policy Research, MIT, Revised discussion draft
25. Knops HPA (2004) Securing Dutch electricity supply: towards a supplier of last resort? In: Roggenkamp MM, Hammer U (eds) European energy law report I. Intersentia, Antwerp, pp 235–274
26. Linares P, Labandeira X (2010) Energy efficiency: economics and policy. J Econ Surv 24(3):573–592
27. Littlechild SL (2000) Why we need electricity retailers: a reply to Joskow on wholesale spot price pass-through. Judge Institute for Management Studies, University of Cambridge, Working Paper 21/2000
28. Littlechild SL (2009) Retail competition in electricity markets—expectations, outcomes and economics. Energy Policy 37:759–763
29. Long N, Bull P, Zigelbaum N (2011) Efficiency first: designing markets to save energy, and the planet. In: Sioshansi FP (ed) Energy, sustainability and the environment. Technology, incentives, behavior (Chapter 8). Elsevier Inc., Amsterdam
30. Lowell E (2006) Energy utility rate setting. Lulu. www.lulu.com
31. Loxley C, Salant D (2004) Default service auctions. J Regul Econ 26(2):201–229
32. Michaels H, Donnelly K (2010) Enabling innovation on the consumer side of the smart grid. MIT, Industrial Performance Center. http://www.mit.edu/ipc/
33. Mulder M, Shestalova V, Zwart G (2007) Vertical separation of the Dutch energy distribution industry: an economic assessment of the political debate. Intereconomics 42(6):0
34. NARUC (1973) Electric utility cost allocation manual. National Association of Regulatory Utility Commissioners, Washington
35. National Action Plan for Energy Efficiency (2007) Aligning utility incentives with investment in energy efficiency. Prepared by Val R. Jensen, ICF International. www.epa.gov/eeactionplan
36. Office of Energy, Government of Western Australia (2009) Electricity retail market review. Final recommendations report. Review of Electricity Tariff Arrangements. Report to the Minister for Energy, Jan 2009
37. Ofgem (2006) Domestic metering innovation. Consultation document. www.ofgem.gov.uk
38. Ofgem (2008) Energy supply probe—initial findings report. www.ofgem.gov.uk
39. Ofgem (2010) Review of current metering arrangements. www.ofgem.gov.uk
40. Pérez-Arriaga IJ, Batlle C, Vázquez C, Rivier M (2005) White paper for the reform of the regulatory scheme of the power generation in Spain (in Spanish) for the Ministry of Industry, Tourism and Trade of Spain

41. Pérez-Arriaga IJ (2006) Redesigning competitive electricity markets: the case of Spain. Chapter in: Complex electricity markets. In: Mielczarski W (ed) The European power supply industry series, Technical University of Lodz, Poland. ISBN 83 921636 7 2
42. PUCT, Public Utilities Commission of Texas (2011) Chapter 25: Substantive rules applicable to electric service providers. Subchapter B. Customer service and protection. §25.43. Provider of Last Resort (POLR)
43. Rose K, Meeusen K (2006) 2006 performance review of electric power markets. www.kenrose.us
44. Sioshansi FP (ed) (2011) Energy, sustainability and the environment. Technology, incentives, behavior. Elsevier Inc., Amsterdam
45. Sioshansi FP (ed) (2011) Energy, sustainability and the environment. Technology, incentives, behavior. Elsevier Inc., Amsterdam
46. Tierney S, Schatzki T (2009) Competitive procurement of retail electricity supply: recent trends in state policies and utility practices. Electr J 22(1). doi:/10.1016/j.tej.2008.12.005
47. Vázquez C, Batlle C, Rivier M, Pérez-Arriaga JI (2006) Regulated tariff design in a liberalized power system: the case of Spain. IIT Working Paper IIT-06-0XX, Instituto de Investigación Tecnológica, Universidad Pontificia Comillas, Madrid
48. Waddams Price C (2008) The future of retail energy markets. Energy J (special edition in honour of David Newbery):125–147
49. Wilson R, Peterson P (2011) A brief survey of state integrated resource planning rules and requirements. Prepared by Synapse Energy Economics Inc. for the American Clear Skies Foundation

Chapter 10
Regional Markets

Luis Olmos and Ignacio J. Pérez-Arriaga

As in most industries, the power sector has been witness to a relentless drive to convert local into regional markets. No one can deny the need for these markets, in light of the huge challenges that lay ahead on the road to a sustainable society where scarce resources are allocated, both fairly and efficiently, at a global level.

The last two decades have seen a general trend towards the creation of supranational or regional electricity markets, i.e. markets comprising several countries or states. Examples include the European Union's Internal Electricity Market (IEM); the market developing in South-East Europe that will join the IEM sooner or later; the NORDEL market in Scandinavia, MIBEL in the Iberian peninsula and the Single Electricity Market (SEM) in the island of Ireland; The Central American Electricity Market (Mercado Regional de Electricidad, MER), MERCOSUR and the Andean markets in South America, which will eventually merge as well; the US's standard market design (SMD) as an attempt to establish a common template for electricity markets across the country; the Australian National Electricity Market (NEM) or the market in the Mekong Delta region.

Even before the relatively recent move towards the creation of regional markets, national or local electric systems were already interconnected in some of the major regions of the world. However, the aim of such interconnections was to enhance not dispatching cost-efficiency but reliability. Power was exchanged when systems needed back-up energy to ensure their integrity. Now, however, electricity producers and consumers are generally thought to have much more to gain from local system interconnection and eventual integration into regional superstructures. In this chapter, a first introduction to regional markets is given in Sect. 10.1, where we discuss, among other things, the benefits of regional integration; the challenges faced and requirements to be met when creating these markets; the

L. Olmos (✉) · I. J. Pérez-Arriaga
Universidad Pontificia Comillas, Instituto de Investigación Tecnológica,
Alberto Aguilera 25, 28015 Madrid, Spain
e-mail: luis.olmos@iit.upcomillas.es

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_10,

principles to be followed in designing regional markets and stages in the integration of systems into a regional market.

Afterwards, Sects. 10.2–10.4 describe the main features of regional market operation and planning, as regards the common transmission grid where power suppliers and consumers meet. These include enlargement of the regional grid, regulation of agent access and cost allocation. Section 10.5 discusses the harmonisation needs in regional markets.

By way of illustration, Sect. 10.6 summarises the main current features of the Internal Electricity Market in the European Union. Lastly, the conclusions are set out in Sect. 10.7.

10.1 An Introduction to Regional Markets

Before we start, we need to fix the terminology in order to minimise confusion. In this chapter, it will be understood that a "regional" power system encompasses several "national", "state" or "local" power systems. These "national" (e.g. Italy, Costa Rica or Argentina), "state" (e.g. California, Ontario or New South Wales in Australia) or "local" (not corresponding exactly to the two preceding categories, as Northern Ireland, New England in the US or the power system run by the RWE transmission system operator in Germany, sometimes called a "control area") power systems are characterised by having a highly harmonised and coordinated organisation of the operation and capacity expansion functions, either under a traditional or a market-oriented regulatory framework. A "regional market" is the outcome of a more or less successful attempt at establishing a higher hierarchical level of organisation of several national, state or local systems, so that their original spontaneous interactions become stronger and subject to well defined commonly agreed rules. In this chapter the terms "national", "local" and "state" will be used without distinction when referring to the lower hierarchical level in general descriptions of regional markets.

This classification is not clear-cut, however. Some of the so-called "local" systems are also aggregations of other entities. For instance, the Regional Transmission Organizations (RTOs) or Independent System Operators (ISOs) in the US—such as New England ISO or ERCOT in Texas or PJM—consist of a large collection of member utilities. In the same way, many diverse firms operate within the Italian or the Norwegian system. However, the generation plants of these companies are centrally coordinated by a single system operator and they somehow participate in or have to provide information to a common trading platform that is also centrally coordinated (even if many of the trading decisions may be adopted bilaterally). And there are more than two levels in some cases. For instance, the national systems of Portugal and Spain have created a regional market, MIBEL, with a higher level of coordination than the larger market in which they participate: the European Internal Electricity Market. This is also the case of the SEM in Ireland and NORDEL in Scandinavia.

Therefore, in this chapter, Guatemala, Brazil, Colombia, California, PJM, New England, Ontario, Italy, Portugal, the Republic of Ireland or New South Wales will be at the national, state or local level. And the Central American Electricity Market (MER), MERCOSUR, the Easter Interconnection of the US, the European Internal Electricity Market, NORDEL, the Irish Single Electricity Market (SEM) or the Australian NEM are examples of regional markets. The electrical sizes are very different in some of these regional markets. The entire regional MER in Central America is smaller than most of the national, state or local systems in the US or Europe. The largest local systems in the US—MISO, PJM, ERCOT, New York ISO or ISO New England—are of the same order of electrical size as the largest national power systems in Europe—Germany, France, Italy, the UK or Spain–.

A roadmap to the contents of this section follows. The additional benefits of regional integration are discussed in Sect. 10.1.1. While economically sound, the creation of regional markets is now easy task. It entails significant technical and economic challenges, described in Sect. 10.1.2, that must be addressed.

The main principles underlying the integration of national systems are explained in Sect. 10.1.3. This is followed by an outline in Sect. 10.1.4 of the requirements to be met to rise to the existing challenges. The introductory part of this chapter concludes with an explanation, in Sects. 10.1.5, 10.1.6, respectively, of the typical stages in the process of creation of a regional market and certain recommendations on the features that should characterise its institutional backdrop.

10.1.1 Benefits of Regional Integration

The scale of technical problems involved in system interconnection has increased in the recently developed regional context. The number and volume of bilateral or multilateral agreements, normally among market agents from different areas but with access to the entire regional grid, have risen. At the same time, more and more regional electricity markets are being created, and the ones in place are growing, to enable participating control areas or countries to capitalise on the benefits of forming part of a larger market. The main reasons for building a regional market are briefly explained in the paragraphs below. Some of the benefits of regionally integrated markets are attainable through some level of limited coordination among interconnected power systems, while others require the institution of a truly operational supra-national market.

Market integration can lower system operation costs, because it affords the possibility of dispatching modern, more efficient generators in one area instead of older, less efficient units in another. Plants with very high production costs need not be operated whenever lower cost units in another country can supply the power needed. Thus, for example, large regional markets are better able to take advantage of the differences in the hydro inflow regimes across the region, in weather conditions or in the non-coincidence of load patterns or holidays.

In addition, a larger market can accommodate bigger and more efficient generators that would not otherwise be profitable. This is typical of regions where control areas and local loads are too small to warrant the construction of large generating plants (such as in Central America).

Competition among power producers in wholesale markets and power suppliers in retail ones is also necessarily keener in a larger regional market than in national markets, given the greater number of potential market agents. This enhanced competition should drive the global average of end user prices down (although they could increase locally in exporting areas), thereby sharing with demand at least part of the increase in the net social benefit resulting from greater efficiency.

The level of security of supply should rise when a regional market is created, thanks to the diversification of the sources of primary energy available. Resource sharing lowers the risks associated with the shortage of any given fuel or spiralling prices. This wider range of potential sources of primary energy also reduces the region's overall dependence on third countries.

Assuming a certain degree of coordination among the control areas or countries, regional markets should be more robust than national systems, for not only primary regulation, but secondary and tertiary reserves can be shared.[1] Moreover, the use of fewer and more efficient generation units to provide system reserves should also lead to substantial savings. The total reserves needed could be further reduced by fully exploiting the lack of simultaneity among peak load situations in different parts of a region.

Most importantly, the operational difficulties created by the uneven distribution of primary renewable energy sources and the non-simultaneity of peak load conditions across a region, when trying to meet the ambitious targets for the reduction of CO_2 and other pollutant gas emissions, are alleviated by renewable resource sharing within the region. In many situations, clean renewable generation that would otherwise remain idle or whose energy production would be wasted could be used to supply power for loads in far away areas of a region if the respective systems are sufficiently integrated. No one can deny the need for these markets in light of the huge challenges that lay ahead on the road to a sustainable energy system.

[1] Because of the need to maintain an instantaneous balance of supply and demand in a power system, as explained in Chap. 1 of this book, provision must consequently be made for operating reserves (i.e. generation whose output is able to adapt within seconds to changing system conditions). The installation of very large amounts of intermittent renewable generation (mainly wind, but also solar) is intensifying system variability, and with it the need for system reserves. Ensuring that such reserves are available at each local system calls for building many more permanently operating units than strictly necessary. With the integration of national or local electric systems, reserves could be shared under a scheme whereby each system could draw from the reserves in others whenever needed (automatically, in the case of primary reserves).

10.1.2 Challenges Associated with Market Integration

When planning the development of a regional market, the circumstances prevailing in each of the local systems in the region must be taken into account to ensure a successful integration. In other words, market design must take the baseline situation into consideration. The vast majority of national systems have been designed to be self sufficient, with relatively weak ties with other systems in the region. In most cases, these interconnections (when present) were originally intended solely to enhance system reliability and are clearly insufficient to host the power exchanges driven by economic and environmental needs. Generally speaking, each national market is regulated rather differently from all the others. National authorities seem reluctant to lose part of the control exercised over their respective national markets, a development that is nonetheless requisite to the efficient operation of a regional system. Institutional and regulatory challenges to regional market development must not therefore be underestimated. They are especially important in regions lacking any supra-national institution able to furnish certain legal guarantees respecting the transfer of sovereignty inherent in the establishment of a common market.

Perhaps the most challenging problem in the design and implementation of a regional market is to change the "national mentality" of the companies, consumers, institutions and regulators into a "regional mentality", where the prime objective is to maximise the global social welfare of the region, while making sure that the individual participant countries are also better off with the integration. For instance, in general there are important savings to be made when security of supply is considered at regional level, rather than separately for each individual country. However, this means that some countries have to trust that, in case of scarcity of supply, other countries will share whatever generation they have, according to prescribed regional rules, rather than giving priority to their local demands. This is not easy to accept; with the consequence that every country will install enough local generation capacity to meet its demand and the savings derived from regional integration will be minimal. Another typical feature in a regional market is that the prices of mostly exporting local areas go up with exports, therefore hurting the consumers (although benefiting the locally more abundant generators). Some regulators cannot accept this fact of life that is inherently linked to exports, and try to place all sorts of barriers to exports in their area, so that consumer prices stay low. Again, this is in detriment of benefits for the generators in the exporting local area and of global social welfare in the entire region.

As mentioned earlier, the technical problems involved in the interconnection of several power systems tend to be more intense in a regional context, where the rising number and volume of transactions between parties from different areas are normally the outcome of bilateral or multilateral agreements among market agents with access to the entire regional grid. Such technical problems include the need for inter-country cooperation to keep the frequency stable throughout the region; ways to ensure that control areas honour the power exchanges scheduled; and the

need for systems to come to one another's aid when local generation shortages arise or when stability problems threaten system integrity.

The integration of differing local markets poses new challenges that did not previously exist. The opening of national markets to international transactions renders transmission network management and development more complex. It may be readily concluded from the foregoing that the existing interconnection capacity is not presently able to cope in most cases with the flows stemming from all the economically and environmentally efficient transactions that should take place among agents in the region. Thus, a new regulatory paradigm for grid expansion must be implemented to yield optimal, or at least sufficient, regional grid development. Given the frequent lack of interconnection capacity between areas within a region, efficient, market-based mechanisms must be established to allocate any transmission capacity available for regional transactions. Furthermore, with the intensified use of local transmission grids for international transactions, the requests by national systems for compensation for such use, or for the benefit accruing to international agents as a result of such use, can be expected to intensify as well.

Last but not least, if a region is to build a fair and efficient market, it must ensure a level playing field where market agents compete on equal terms. The lack of regulatory harmonisation among the systems involved constitutes a sizeable challenge in this regard. One example is the structure of the network charges applied by each country in a region. In some countries, they are charged to generators only and in others only to consumers. The fact that some countries include regulated costs in these charges that are unrelated to the network is an additional source of complexity. All these factors are clearly obstacles to market integration. Surmounting such barriers call for substantial harmonisation. Other examples of regulatory instruments whose use must be harmonised are those that promote generation with renewable resources and the capacity instruments to enhance the security of supply in generation.

10.1.3 Basic Principles of Regional Integration

When faced with the aforementioned challenges, the authorities and actors in the region concerned must proceed according to a set of basic principles to successfully integrate their systems. These basic principles must be consistent with the baseline conditions but ambitious enough for agents to realise the potential benefits of complying with the requirements for creating a regional market, which are discussed below.

- Market development must be compatible with the preservation of state or country sovereignty and sovereign rights over natural resources.
- Countries in the region must commit to furthering the integration of their systems (and therefore markets) in the region. This involves, among other things:

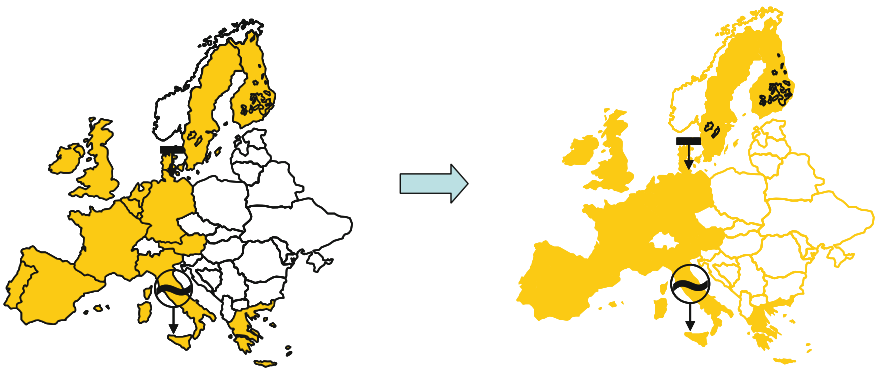


Fig. 10.1 Application of the single system paradigm to the IEM of the European Union

  - restructuring the region's institutional scenario, which must nevertheless respect specific regional and national characteristics,
  - applying decision-making criteria to manage system operation that are designed to defend collective rather than national or area interests, and
  - inter-actor coordination and information and technology sharing to improve system reliability.
- The market outcome must be the result of freely acting market forces, which means that:
  - competing transactions must have open access to the use of the transmission grid, and
  - barriers to the participation of external agents in local markets must be lifted.
- A level playing field must be ensured, meaning that the existing regulations must not discriminate among agents from different countries. This implies that harmonisation efforts must be made in a number of areas.

All these principles can be summarised in what is generally referred to as the "single system paradigm", whose guiding premise is that application of the necessary regulations should enable all the systems in a region to operate as if they were actually only one. This is illustrated in Fig. 10.1, where an outdated configuration of the European Union's Internal Electricity Market (IEM) has been taken as an example. The application of this paradigm to the design of the various aspects of regional market operation is explained in Sects. 10.2–10.5.

10.1.4 Requirements for the Creation of a Regional Market

Establishing a regional market entails a number of challenges, as discussed above, that can only be successfully addressed if the principles outlined in the preceding

section are followed. These principles can be translated into a number of more specific requirements to be legislated by the regulatory authorities in a region. The most prominent of these requirements are discussed briefly in the paragraphs below.

In the first place, a regional rather than a national or local approach must be adopted in the regulation of electricity transmission. The main objectives of such regulation are listed below, under the main items into which the approach may be divided: allocation of the transmission grid costs; agents' access to the network and grid expansion and cost recovery of grid investments.

- The objectives relating to the allocation of grid costs are:
  - guaranteed recovery of the cost of efficient network investments (at least of regulated investments granted regulatory approval), and
  - transmission of efficient economic signals to market agents—both in the short and in the long term–, regarding the cost to the system of installing their facilities in one part of the grid or another.
- The objectives in connection with the agents' access to transmission network capacity include:
  - assurance of open access to the entire regional grid by regional agents,
  - allocation of scarce transmission capacity in the regional grid through the use of efficient market mechanisms, and
  - mechanisms for allocating the transmission capacity that reduce, or at least do not increase, agents' incentives to exercise market power.
- The objective to be attained under regional grid expansion is:
  - institution of a grid expansion mechanism resulting in the construction of an optimal set of network reinforcements, or at least those investments required to integrate the national systems in the region [1].

In the second place, all unnecessary barriers to be freely conducted commercial transactions between agents from different countries should be removed. These include, among others, the pancaking of national tariffs when computing network tariffs applied to cross-border transactions (this is at the same time a discriminatory practice favouring local over regional trade); overly complex or discriminatory procedures for buying or selling energy in other markets as, for instance, overly stringent requirements to compensate imbalances of generation and demand in international bilateral contracts; the lack of transparency in cross-border trading arrangements; and the uncertainty surrounding possible changes in the rules affecting the inter-country energy trading.

In most countries, power exchanges have been established to facilitate trading. The successful implementation of a regional market requires some minimal level of harmonisation among them [1]. Ideally, some sort of common trading platform should be created but, at least, the compatibility of any regional arrangements with the existing national markets should be ensured. In other words, the prices and conditions to which generators and producers are subject in each system must not

interfere with efficient regional economic signals or the incentives for agents to participate in that market. The times of gate closures, the formats of bids and several other basic features of the national markets must be harmonised. Operational and safety procedures deeply rooted in national systems should be maintained as far as possible by making regional operational and safety procedures compatible with national arrangements, like the internal use of control areas, as it is customary in most European countries and many other systems in the world.

All provisions of national system regulations that may potentially interfere with the creation of a competitive regional market should be harmonised or removed to enable regional agents to compete on a level playing field. Issues meriting specific attention in this regard include rules requiring national self sufficiency or giving priority to local demand in cases of scarce supply, lack of harmonisation in the establishment of capacity instruments to enhance security of supply, the process for calculating transmission tariffs in each country, the inclusion of locally regulated charges in the tariffs paid by agents or the support schemes in place in each area for the deployment of renewable generation capacity. See Sect. 10.5 for details.

National transmission system operators and regulators should be independent from other national interests in the exercise of their duties. Moreover, the responsibilities for the various areas of regional market operation should be clearly assigned to ensure that due progress is made in the application of the rules and any necessary decisions can be made in a timely manner.

The pace at which the rules and changes required to create a regional market is applied must be consistent with the situation prevailing in the region.

10.1.5 Stages in the Integration Process

A functional regional market must efficiently integrate all the constituent national systems. Such a market cannot, then, arise spontaneously, but must be the outcome of a series of stages, as summarised below.

- Two or more national systems are physically interconnected, and generation companies engage in some voluntary bilateral trade or submit energy bids (including volume and price) to the neighbouring systems. This contributes to improving the economic conditions and the reliability of supply in all the systems involved. Technical coordination exists among them as well as some degree of harmonisation of security and reliability criteria.
- Interconnected national system operation is coordinated. National systems try to reduce their production costs by applying some common technical and economic rules. Wholesale energy markets are created where generation companies, large-scale consumers and suppliers or retailers can buy and sell power regionally. Some regulatory provisions on system operation are harmonised to provide fair market conditions for all agents and surmount any barriers to trade.

- Operation is fully integrated in the various national markets. All agents can participate in the regional market and freely choose their regional counterparty. Transmission network expansion is jointly planned at regional-wide level.

10.1.6 Institutional Setting and Framework in Regional Markets

As noted earlier, a number of regional markets are presently in place. The driving force or general philosophy behind regional markets may differ from one to another. Up to a point, this determines the market structure and therefore the stages involved in carrying integration through to completion, as designed in each region. Three examples will be mentioned here that show the relevance of the definition of the high level design of a regional market:

(a) The Internal Electricity Market (IEM) of the European Union (EU) is based on the rule that agents from anywhere within the 27 countries may access the entire regional grid by just paying a local G (generator) or L (consumer) transmission charge, which in general differ from one area (country) to another. Such agents are then free to negotiate region-wide transactions, either bilaterally or by participating in a power exchange. The EU has established 2014 as the year where a common target market model will be implemented EU-wide. This model includes a common trading platform (a sort of regional power exchange) for the day-ahead and intra-day markets. The regional bid-matching model is very simple and uses a very crude network representation, basically only including the interconnection capacities between countries.
(b) The MER (Mercado Eléctrico Regional, Regional Electricity Market) in Central America—has opted for a system in which a regional market overlies rather than replaces the existing local markets. Every day each one of the six participating countries (Panama, Costa Rica, Nicaragua, Honduras, El Salvador and Guatemala) clears its own day-ahead internal dispatch (all countries have internal local competitive electricity markets, except for Costa Rica) and then submit their offers to buy and sell at each one of their transmission nodes to the regional market. The regional market runs a region-wide optimal load flow for each hour of the next day, which provides the optimal regional dispatch of generation and demand and the nodal prices. In this way, each country maintains its autonomy while trading efficiently with the other countries in the region.
(c) The local electricity markets in the US that are run by Independent System Operators (ISOs)—like PJM, MISO, ISO New England, New York ISO, California or ERCOT in Texas—have similar advanced designs that follow a common template, centred on day-ahead markets with nodal prices in thousands of nodes, which are typically updated in the very short time, e.g., 5 min

before real time, see Chap. 7 or [5]. Now the challenge is to integrate these large complex wholesale markets—similar in electrical size to large European countries—into truly regional markets, of the dimension of the Eastern or the Western Interconnection. There are ongoing activities that are expected to end up coordinating these local markets bilaterally, and eventually also at regional level.

The market's institutional setting is also conditioned by its high level design. In a regional market, responsibilities must be clearly assigned to a series of institutions, some of which may be national or local, whereas others must be regional. The major system functions are listed below.

- Application and enforcement of regional regulation tend to be conducted by national governments and regulatory authorities, together with any existing regional political and regulatory institutions or associations of national institutions.
- Regional market operation may be the responsibility of a sole institution if a single regional market exists, or incumbent upon the various market operators if the regional market is based on the coordination of national markets.
- Regional system operation, analogously to market operation, may also be conducted under two alternative arrangements, depending on whether a full-fledged regional market (and system) exists or otherwise.
- Participants in the regional market (agents) who buy and sell energy at the regional level may create regional associations to defend their interests.

The development of some sort of regional regulation and the existence of region-wide regulatory institutions is key for the success of a regional market. The particular responsibilities incumbent upon national or regional regulatory bodies with respect to the operation of the regional market follow.

- They must lay down rules for regional market operation, system operation and transmission regulation at regional level. In particular:
  - They must promote the development of competition among market players through the application of regional regulations.
  - Regional and national regulatory bodies must coordinate their activity to create a level playing field for all the market agents.
- They must also enforce the legal and regulatory framework by monitoring regional market transparency, and functionality verifying that competition actually exists and preventing any violation of the market rules or applying penalisations when necessary.
- Regulatory authorities must also settle disputes or problems arising around the participation in the regional market.

10.2 Regional Grid Development

Given the prominence of the transmission grid in the construction of regional markets, the next four sections will be devoted to a more detailed examination of the relevant issues concerning the design and implementation of regional grid regulation: regional transmission network planning, access of international transactions to the grid, allocation of grid costs and harmonisation of rules. In this section we start with regional planning.

Since the transmission grid in place in a region must be able to accommodate socially efficient power exchanges, the region's regulatory authorities must lay down the rules governing decisions on the planning and construction of new regional transmission lines. The various approaches that may be adopted to regulating transmission grid enlargement are essentially the same for regional and national systems (see Chap. 6 for a comprehensive discussion of this subject, as well as Pérez-Arriaga et al. [29].

The proposals for building new lines should be made by a regional institution (which should logically at least include the national system operators) to maximise the social benefit of network investments. This centralised institution may also be competent to decide on line construction or this role may be left to regulatory authorities. Alternatively, the initiative to propose regional network reinforcements may be left to private parties, be they merchant investors willing to profit from line capacity management or associations of network users that would benefit from their construction.

International experience in this regard suggests that, except in the typically few cases where the beneficiaries of network investments can be clearly identified, and financing for these new lines is available, the rest of the network will have to be built under some sort of centralised planning, although competition for the allocation of the construction of the lines is possible and should be promoted. The investments that private parties are willing to finance and build tend to be less than socially optimal, because agents' revenues from the commercial exploitation of the capacity of efficiently dimensioned lines tend to be much smaller than the social value of these lines (see [27]). Centrally planned transmission investments can nevertheless coexist with others proposed by private parties as long as the latter are not socially detrimental.

Regional grid development should not, however, be the mere aggregation of the reinforcements envisaged in national transmission expansion plans, which tend to be laid with local interests in mind. Network investments made in the countries participating in a regional market must be coordinated, so that the lines actually built satisfy overall regional needs rather than only the needs of each country considered separately. Grid transmission planning must therefore be region rather than nation or state wide [2, 4].

The benefits of the new transmission lines required to integrate national systems are generally present across the entire region in question, i.e. they are perceived by agents in different countries. The local systems involved must agree on where lines

are to be built and how their cost is to be shared. Otherwise, no investments will be made. This is a particularly relevant issue in the current context, where pressure to decarbonise the economy is driving the construction of renewable electricity generation capacity. Since renewables tend not to be homogenously distributed throughout a region, in countries where they are available they should be used to produce electricity not only for local consumption, but for export to other systems. In fact, the installation of renewable generation capacity would not be socially efficient in many cases unless all or part of the electricity output can be exported to other areas or countries in the region. Given that the network reinforcements required for the integration of renewable energy sources benefit several of a region's systems, their construction is more difficult in the absence of coordination among local systems.

In addition to defining a regulatory scheme for the coordination of regional grid expansion planning, a set of rules and associated algorithms must be developed to assess the benefits accruing from each proposed regional network reinforcement. The local approaches followed to decide on the construction of local lines may not be wholly applicable on a regional scale, for the reasons set out below.

- Due to the difference in scale between national and regional transmission systems, transmission expansion needs are significantly larger at the regional than at the national level.
- The uncertainty arising around the assessment of the benefits attributable to regional lines is high due, among other reasons, to the need to factor into the equation the future development of renewable generation capacity across the entire system.
- The effect of new technologies (e.g. high voltage DC lines or new AC network overlays at higher voltages) is normally disregarded in the reinforcement of local networks.

Section 10.2.1 discusses the regulatory scheme governing the coordination of local system initiatives to build new transmission lines. Section 10.2.2 deals with the development of rules to assess the benefits of constructing new regional lines and identifying the main features required of a regional planning algorithm.

10.2.1 Decision-Making Framework for Regional Grid Development

Either of two models may be applied to regulate regional transmission grid expansion. In one, the initiative for proposing new reinforcements is left to the voluntary bilateral or multilateral coordination of local systems. The other relies on a central planning process to identify the required grid reinforcements. Regardless of which approach is adopted, some provision should normally be made for additional line construction by merchant developers under non-regulated risk arrangements. All these issues are discussed in the following paragraphs.

Proposals for building regulated regional lines put forward by local systems
In this approach, no single entity is responsible for developing the regional grid. Reinforcements to the existing lines or the construction of new cross-border lines are proposed by at least one national or local TSO or regulator. Decisions to upgrade the regional grid are not made unilaterally by any single country since, by definition, these upgrades affect more than one. Proposals to build new cross-border lines are jointly assessed by all the countries concerned.

In practice, a mechanism must be devised to decide whether the line is of a national or a regional scope. The criterion used may be line location (for cross-border lines) or any other if more efficient (an assessment of the benefits of the network reinforcement is certainly more efficient). By way of example, the allocation of line usage might be used as a measure of whether it is regional or local. If the conclusion reached is that other countries/systems will account for a significant portion of line usage, the new facility should be regarded as a reinforcement to the regional grid. Otherwise, it is defined to be a local reinforcement and the area, on whose land it lies, is free to decide whether or not to build it.

If the line is regional, some joint entity or committee representing the concerned systems is vested with the power to approve line construction.[2] This entity must assess the benefit the line will afford the system (i.e. whether it passes the so-called regulatory test, see Chap. 6 and Sect. 10.2.2). This committee may also decide on the allocation of line costs among the local systems pursuant to some pre-established mechanism. The committee's decision to approve line construction or otherwise may be binding or merely indicative. Such committees, which may request technical support to analyse regional investment projects, may be standing bodies or created ad hoc for each specific project.

In the US, FERC has published Order 1,000 in July 2011 [19] requiring improved coordination in transmission planning and cost allocation procedures at bilateral level between the existing Regional Transmission Organisations and also with other large single dispatched systems. This may be seen as a first step in the right direction towards a more comprehensive transmission planning at Interconnection-wide level.

Proposals for building regional lines as part of a centralised regional expansion planning process
Under this model, regional reinforcements are proposed by a regional institution with the necessary technical expertise (the regional association of network operators or a regional system operator, for instance). Local and regional investments must be coordinated in some fashion. As in the preceding approach, a mechanism is in place to decide whether a reinforcement is part of a local or the regional grid. The regional planner must normally study line construction proposals by market agents or other regional institutions (local TSOs or groups of generators and loads). If a proposal is not accepted, that decision must be formally justified.

[2] An alternative scheme would have a single committee for the approval of any regional line.

The final decision on the construction of new regional lines (the approval of the respective investments) is normally incumbent upon a regional regulatory institution. National (local) authorities and private parties may or may not be able to challenge the decision by the regional regulator to approve the construction of new regional regulated lines. In other words, the approval or otherwise of line construction by the regional regulatory authority may be binding or merely indicative. Regional regulated line construction is usually awarded under competitive tendering.

This is more or less the model that has been adopted at the Internal Electricity Market (IEM) of the European Union, as explained in detail in Sect. 10.6.2.

The role of private initiative: investments at risk

In either of the two regulatory approaches just discussed (centralised vs. decentralised investment proposals), private parties are normally able to propose, finance and build lines. Such investments may be put forward by network users (generators, producers) or private entrepreneurs willing to arbitrage prices between buyers and sellers (merchant investments). The specific characteristics of regional markets may encourage such projects (price differences between one and another area of a region may be larger and more stable than within each national system). Allowing private initiative to build new lines may be beneficial for the system [3] when:

- Public authorities or governments are unable to raise the funds needed to build a regional reinforcement.
- The regulated planning process for new line construction is very slow, retarding the materialisation of efficient investments.
- Potential network reinforcement is considered to be too risky an investment by authorities for it to merit approval as a regulated reinforcement, but a private party may be willing to undertake its construction.

A typical example of investment at risk is the construction by generation project developers of the lines connecting the generation sites to the rest of the system. Postponing line construction until funding becomes available in regions where access to financing is not straightforward could significantly delay the commissioning of new generation plants.

The foregoing notwithstanding, before a private party obtains regulatory approval to build a line, regional regulatory authorities must ensure that these investments are not detrimental to the system.

10.2.2 The Regional Regulatory Test

The "regulatory test" is the set of rules applied to determine whether the construction of a given network reinforcement, or series of reinforcements, is justified. This set of rules may vary depending on the identity of the party applying the test and the nature of the investment assessed. On those grounds, different kinds of regulatory tests may be defined, depending on the situation.

(1) The test may be conducted by a "passive" system operator (an operator not making the final decision on network reinforcements) when determining which network upgrades to propose. An "active" TSO (an operator who makes final decisions on network reinforcements) may also use it to show the regulator that the investments planned are economically justified and should therefore be considered when calculating the remuneration for the transmission business in the next control period.
(2) Regulators also use the test to determine whether a line proposed by some party should be built. The two alternatives in this case are described below.
    (a) The investment proposal may be put forward by a system operator responsible for central grid expansion planning. Such proposals come under the regulated investment category and include the investments proposed by both active and passive TSOs.
    (b) The proposal may be put forward by a private entrepreneur or an association of network users willing to finance the upgrade, in which case it would be classified as a merchant or venture capital investment.

The regulatory test applied in situations 1 and 2(a) must identify the most efficient of a series of possible investments. The regulatory test used in situation 2(b), by contrast, merely needs to verify that building the corresponding line is not detrimental to the system.

The following aspects of the planning methodology should be considered when deciding to build a regional line.

Regional scope

Further to the two basic regulatory approaches to grid expansion, both regional and national institutions may be responsible for proposing new reinforcements. In both cases, however, the proposals must be assessed from a regional perspective, i.e. attempting to maximise the social benefit for the region as a whole rather than for a specific area. The expected massive deployment of renewable capacity is obliging transmission planners or regional regulatory authorities to broaden their customary local focus. Given the uneven distribution of renewable resources throughout a region, some areas may have a surplus of renewable production when others have a shortage. The excess should be exported from the former to the latter across lines used primarily for regional power exchanges.

Dealing with uncertainty: scenarios

Given the large dimensionality and the considerable uncertainty with which the future transmission grid will be confronted, an approach to the planning problem as a full-fledged stochastic approach might possibly be too hard to solve. A pragmatic alternate approach is to define a reduced number of plausible uncertainty scenarios and a couple of time horizons (e.g. tactic at 10–15 tears and strategic at 20–30 years), see [10]. Any reinforcement undertaken should be cost-efficient in most if not all the scenarios considered. This approach could be revisited when more information about actual conditions becomes available.

Selection of planning criteria
Several criteria may be applied when assessing potential network reinforcements, depending on the aims sought: the improvement of system reliability while meeting some minimum requirement, the minimisation of supply costs, the maximisation of grid accessibility for renewable sources, the absence or otherwise of entry barriers and the degree of utilisation of local resources and demand management. Where all or part of these aims are considered simultaneously, as is preferable, planning could be addressed as a multi-criteria problem.

Considering the technological choices available
As noted earlier, many of the transmission lines that will be built in the years to come will be regional. Designing and developing a regional grid involves the consideration of the various technological options available, which include: (a) building a super grid or set of higher (AC, DC or both) voltage overlays over the existing network; (b) reinforcing the existing network with specific facilities as needed; (c) combining higher voltage overlays and reinforcements to the existing transmission network. The technology used may differ from one part of the grid to another. Offshore wind or solar generation capacity in northern Africa should probably be connected to the main European system, for instance, via large HVDC lines or corridors.

The large dimensionality of the planning problem
The larger scale of the regional as compared to any local or national grid may call for the application of advanced optimisation techniques, such as decomposition into subproblems.

10.3 Grid Access

Significant inter-area congestion initially exists in most regional markets, because the present transmission network in most areas was designed to supply local load with power produced in local generation plants rather than to provide for large scale, regional power exchanges. Thus, the mechanism used to allocate the interconnection capacity to regional market agents critically conditions international energy trading.

Access to the transmission grid in a region depends primarily on which energy trading model is adopted. Agents' use of the grid can only be maximised if access to scarce transmission (interconnection) capacity is gradually managed in several time scales. The corresponding mechanisms should probably differ from one time scale to another. Thus, the allocation of transmission capacity in the medium to long term, short term (day ahead) and very short term (intraday and real time), is dealt with separately in Sects. 10.3.2–10.3.4 below, respectively. This discussion is preceded by that of the calculation of the interconnection capacity available between national or local systems in Sect. 10.3.1.

10.3.1 Calculation of Regional Transmission Capacity Available

In a regional context, the amount of available interconnection capacity between regions must be maximised. This entails coordinating the computation by SOs in the region of available interconnection capacity, which in turn requires that all the systems in the region use a common network model. This issue has been addressed by the association of European Transmission System Operators in a report spelling out the main criteria for allocating transmission capacity among market agents [17, 18]. Note that the use of a common nodal pricing model for the entire region would take implicit care of the problem of determination of the available transmission capacity amongst the several local systems. However, only the MER in Central America has so far been able to produce such a feat, albeit in a quite small regional market. Nodal prices have not been implemented in the European electricity markets and in the US there is ongoing work in coordinating the separate nodal pricing models of two neighbouring systems, such as PJM and NYISO or MISO.

In the absence of a common model of nodal prices, one of the reasons for using, at least, a common network model is that allowing each national or local system to use its own would in all likelihood result in as many estimates of transmission (interconnection) capacity in the regional grid as participants in the scheme. Using different network interconnection capacities is incompatible with the computation of region-wide dispatching. If the computation of the available interconnection capacity is not harmonised, the capacity allocated to market agents at each border would have to be computed as the lowest estimate provided by the TSOs involved. The most probable outcome would be under-use of the region's actual interconnection capacity.

Moreover, if the TSOs do not use a common network model, the resulting map of flows in the regional grid might well be unfeasible, requiring the re-dispatch of initially accepted transactions and more than likely leading to a lower and less efficient use of regional grid capacity.

10.3.2 Allocation of Capacity in the Medium to Long Term

As explained in Chap. 6, efficient congestion management systems yield energy prices that differ in space and time. Price differences from one part of the system to another are significantly larger in regional than in local markets due to the shortage of interconnection capacity among the national systems in each region. Market agents engaging in international transactions in a regional market therefore may perceive a risk associated with energy price volatility for which they seek protection. Risk hedging products were created to address this problem. Markets for these products have developed and have already reached their maturity in certain regions, such as several ISOs in the US—like PJM, for instance—or the Nordic countries in the EU.

As mentioned earlier, two main types of risk hedging products exist: contracts signed with a counterparty that fix the price an agent will pay or receive for its energy in the future, and capacity contracts or transmission rights, that fix the price paid to use some amount of transmission capacity between an injection and a withdrawal point. At least one of the parties to an energy contract in a regional market may find it advisable to conclude a capacity contract to know in advance the price it will have to pay to access the area where its counterparty is located. In the absence of nodal prices, energy contracts without locational differentiation may be negotiated bilaterally between agents or unilaterally between an agent and a central clearing house. The description and discussion of these energy contracts, such as contracts for differences, can be found in Chaps. 6 and 7 of this book.

All issued transmission rights must be simultaneously feasible. Hence, in meshed grids, the allocation of most types of transmission rights to regional agents must be coordinated. Rights in general need to be mediated through a centralised regional institution, which may be created by the region's TSOs or market operators. This central institution would manage a single regional auction, with harmonised rules, IT interfaces and products for medium and long-term capacity allocation. If sufficiently accurate SPAs (single price areas) can be defined, transmission rights could refer to these regional areas. Otherwise, rights should refer to nodes instead of areas, in which case the regional grid model used by the central auctioneer should be nodal as well.

As explained in Chap. 6, several types of congestion rights can be defined. Physical rights entitle the owner to actually use the transmission capacity, whereas financial rights only entitle their holders to the respective congestion rents. Rights may also be defined between any two points of the system (point-to-point rights) or relate to a particular line or corridor (flowgate rights). Finally, congestion rights may be characterised either as the option or the obligation to use the respective transmission capacity or receive the value of this capacity computed in the dispatch.

The availability of both regional transmission capacity products, defined as options and as obligations (physical or financial), is recommended. In the long term, when significant uncertainty exists about regional operating conditions, options should probably be issued. In the medium term, obligations and options could be issued, and option holders should have the opportunity to sell their rights or convert them into obligations. This would maximise the overall use of available capacity.

Financial transmission rights are preferred over physical transmission rights, because the former are less likely to condition the physical use of capacity in real time. However, physical rights may play a significant role in transmission grid expansion in regions where authorities are unable to raise sufficient funds to build all of the regional reinforcements required. In such cases, private developers may build new lines of regional interest if they can secure access to the respective physical capacity interconnecting national or local markets. These are the conditions prevailing in Central America's regional electricity market, for instance [11].

Lastly, point to point are preferred to flowgate rights, because they enable agents in regions with a significant number of congested corridors to secure access to all the transmission capacity they need to undertake commercial transactions. Using flowgate rights would require these agents separately booking capacity in all the regional corridors they may need to access and which are likely to be congested. This could deter these agents from engaging in regional (international) transactions despite the commercial opportunities afforded. Using point-to-point rights, however, requires coordinating the dispatch of capacity over the region, which can only be achieved by means of a single centralised auction. Hogan [24, 25] and Chao et al. [22] discuss the use of both types of rights.

10.3.3 Allocation of Capacity in the Short Term

In the day-ahead time frame, all the scarce capacity that has not been previously committed physically should be allocated to agents who value it the highest. Several alternative methods exist to allocate this capacity. First, it may be allocated together with energy in so-called implicit auctions. Gilbert et al. [23] show that implicit auctions maximise the use of transmission capacity. Implicit auctions in a region may yield a different price for every node on the grid, i.e. a system of regional nodal electricity prices. The computation of pan-regional nodal electricity prices may be deemed the most efficient option in a competitive market. Nonetheless, a system of nodal prices may be more exposed than others to the exercise of Market Power (MP) by agents in importing nodes that are rather precariously connected to the rest of the system. The implementation of a system of region-wide nodal prices would require regionally centralised day-ahead energy and capacity dispatching.

Alternatively, if sufficiently accurate single price areas (SPAs) can be defined, zonal electricity prices may be used. In this case, the capacity allocated would be that connecting SPAs. Implicit zonal auctions in regions where the grid is normally meshed need to be computed by a central entity. Decentralised implicit auctions would only seem to be technically feasible in radial transmission systems, although their implementation in meshed grids has been attempted. The latter option would involve coupling the region's national or local markets through iteration [9].

Region-wide implicit auctions that must in all likelihood be centrally managed may encounter institutional opposition where local trading entities (countries, states or ISOs) are large, have already their own stable trading schemes and are not favourable to give up their identity to be managed at a higher regional level. In such cases, scarce regional transmission capacity (which normally is the interconnection capacity between local systems) must be allocated independently of energy in explicit capacity auctions. Explicit auctions would also have to be centrally managed (coordinated explicit auctions) if the regional network is meshed. In radial regional transmission networks, bilateral agreements between

each pair of neighbouring systems in the region could be used to separately allocate capacity at each border. Even in this case, however, some level of coordination among bilateral capacity auctions would be needed to ensure efficient region-wide capacity allocation. Explicit capacity auctions may be followed by either decentralised energy-only auctions or centralised joint energy and network capacity auctions. The use of coordinated explicit auctions combined with energy only local auctions is discussed in Pérez-Arriaga et al. [28].

10.3.4 Allocation of Capacity in the Very Short Term

This item discusses capacity trading arrangements between when day-ahead markets close and real time. As explained in Chap. 7, two types of such arrangements can be defined, depending on the time scale. Intra-day markets are open in the time frame between closure of the day-ahead market and the start of the real-time balancing window. Arrangements concluded after the closure of intra-day markets and through real time take place in balancing markets. Both types of markets are designed to enable agents to balance their positions to offset deviations between their scheduled and presumed power injections or withdrawals. Participants in both types of markets, as well as the system, would clearly benefit from access to region-wide bids and offers.

Intra-day markets may be of different types. The interconnection capacity between areas in a region may be allocated explicitly (separately from energy) or implicitly (together with it). Given the time constraints in these markets for agents balancing their position, transmission capacity products that entail the obligation to use this capacity, rather than the option to do so, are advisable under explicit allocation arrangements for capacity netting purposes. Netting is automatic when capacity allocation is implicit.

In both explicit and implicit mechanisms, trading may be either continuous or through sequential auctions held at predefined times after closure of the day-ahead market. Continuous trading allocates capacity (either explicitly or implicitly) on a first come, first served, basis. It provides agents with more time-wise flexibility in booking available capacity but, as it does not establish market prices for this capacity, it may breed inefficiency. Predefined prices can be used for the capacity booked in explicit continuous trading, which calls for the centralised management of all the bids and offers submitted by the region's agents.

Capacity allocation through intra-day auctions would involve organising a sequence of auctions where scarce regional transmission capacity is allocated, either explicitly or implicitly, to the highest bidder. Agents accessing this capacity must pay the auction price. Conceptually speaking, capacity trading in these auctions may be centralised or decentralised, although the latter would in any event require a considerable degree of coordination between the region's TSOs.

Once intra-day markets are closed, the system's balancing needs must be met in close to real-time markets. Balancing and/or regulation markets exist in most,

if not all, power systems to ensure safe operation. Since these markets are held very shortly in advance of real-time system operation, TSOs are normally responsible for buying or selling energy in keeping with system needs, based on the regulating or balancing bids and offers submitted by market agents.

As explained earlier, a regional system would benefit significantly from the integration of all its balancing or regulating markets. First, total system reserves could be reduced thanks to regional pooling. Hence, only the most efficient reserve offers would be considered, lowering purchase costs. Second, access to bids and offers located in any of the region's systems would render the overall operation safer and more reliable. Integrated balancing markets should preferably consist of centrally coordinated implicit auctions or continuous trading processes, since any coordination failure in these markets could seriously threaten system integrity (the system operator would have no time to react).

This is an open topic, under revision in most existing or proposed regional markets, [5 or 6–8, 16].

10.4 Allocation of Grid Costs

Chapter 6 contends that efficient short-term electricity prices with time and space differentiation (nodal prices in their purest form) are unable to recover more than a small fraction of total transmission grid costs. Extra charges must therefore be added to guarantee the recovery of grid costs. Traditionally, these charges have been applied separately by each national or local system in the region and have been designed to recover local transmission system costs.

As a result of the growth of inter-system power exchanges in most regions; however, the use that local agents make of others' grids and the concomitant benefits accruing to the former have also increased substantially. Therefore, allocating the full cost of each national or state network to local agents no longer seems fair. According to the rationale presented in Chap. 6, agents should be charged for the cost of the lines they benefit from (or the lines they use, as a proxy), which are not necessarily located only in their local system.

Thus, the use made or benefit obtained by agents of the regional grid as a whole should be considered when allocating the cost of regional transmission lines. Note that the method applied to allocate the cost of regional lines among agents indirectly determines the fraction of the cost of each of these lines to be charged to each country (or its agents). If the method used to allocate the cost of regional lines is perceived as unfair by a region's national systems (if the fraction of the cost of these lines that they pay is not proportional to the benefits obtained), they may object to the construction of new regional lines that are nonetheless needed for system integration. Consequently, the fair and efficient allocation of regional line costs is of vital importance in achieving regional integration. The remainder of this section defines and discusses the main principles that should guide regional line cost allocation and the main methods that have been used for this purpose.

10.4.1 Main Principles of Regional Grid Cost Allocation

As explained in Chap. 6, transmission tariffs, which are aimed at allocating the cost of transmission lines, should depend not on the commercial transactions taking place but on the location of agents in the network and the volume and timing of their power injections or withdrawals. This principle applies at a regional level as well. Economic benefits, or perhaps, as a proxy, a flow-based network utilisation method should be used to allocate regional transmission costs.

In regional systems specifically, which by definition encompass several smaller systems, transmission charges must not be computed by pancaking (accumulating) existing local charges. Instead, as explained above, the use of the entire regional grid by all the agents in the region should be the basis for computing regional transmission charges. This is in accordance with the single system paradigm discussed in Sect. 10.1.3, which should inspire regional regulation.

The paragraphs that follow address the main options for calculating regional transmission charges.

10.4.2 Main Options for Computing Regional Transmission Charges

Ideally, region-wide transmission charges should be computed centrally according to the use that each agent is expected to make of the regional network (or the benefits it is expected to obtain from each line). Thus, charges should be geographically differentiated, giving rise to pan-regional transmission charges computed as if only one power system were in place in the region (in compliance with the single system paradigm). Pan-regional transmission charges would be compatible with the creation of a level playing field. Any differences between charges paid by agents would depend solely on the network costs attributable to them, not on the local system to which these agents belong. This would send agents efficient geographically based economic signals, and therefore enhance the efficiency of expansion planning and system operation. The regional electricity market in Central America (abbreviated in Spanish as MER) has adopted a system of region-wide tariffs for allocating the cost of regional grid lines. The method adopted for calculating these tariffs and the general approach followed is described in PHB-HAGLER and SYNEX [21] and KEMA Consulting and ISA [20].[3]

Implementing a system of pan-regional transmission tariffs, however, requires some harmonisation of the transmission tariff-setting process across the region, at least as regards the recovery of the cost of regional lines. Furthermore, applying such a tariff scheme normally involves empowering a single institution to compute

[3] The complete rules of the Central American Regional Market can be found at http://www.crie.org.gt/files/rmer.pdf, in the website of the Regional Electricity Regulatory Commission, CRIE.

regional transmission charges, which would take the place of local regulatory authorities in this respect. In other words, the most conceptually sound approach to allocating regional grid costs calls for a fair amount of institutional and regulatory centralisation and harmonisation across the region. This may not be feasible in regions with deeply rooted local entities or regulatory practices (such as countries with well-established procedures to determine transmission network regulated costs and charges), where other less efficient, albeit more practical, approaches may need to be adopted. These are discussed in the following paragraphs.

An alternative to the reference model introduced above has been applied in the European Union's Internal Electricity Market: The Inter-TSO Compensation scheme (ITC). It involves computing compensation among national systems for the use that each system's agents make of other systems' grids. The global transmission charge to be paid by agents in each country is the result of deducting the net compensation to be received by the country from the cost of the national transmission grid. This net compensation is computed as the difference between the compensation to be received by the country for the use of its network by external agents and the sums charged to the country for its agents' use of grids belonging to other countries.

The global transmission charge to be paid by each country is allocated to local network users (in the country) according to the method devised by national authorities. Thus, the subsidiarity principle applies to the computation of local transmission charges within each country, subject to the sole condition that these local charges must add up to the global transmission charge owed by the country. As specified above, the global transmission charge to be paid by each country should equal the cost of the fraction of the regional transmission network used by the country's agents. Payment of the local transmission tariff by agents in each country grants them access to the entire EU (regional) grid. The approach followed in the EU's IEM for allocating the cost of the regional grid, together with some of the methods proposed for computing inter-country compensation payments for the use of national grids by external agents, are discussed by the authors in Olmos et al. [26].

Since under this approach local systems calculate the transmission charges to be paid by end users, transmission pricing practices need not be harmonised across systems to implement it. Moreover, local authorities play a major role in rate-making. This model is consequently well suited to regions where local entities and regulatory practices are deeply rooted and therefore difficult to change. Such a situation may prevail in regions where the regional transmission grid is an outcome of the interconnection of highly developed local (national/estate) networks. A system of inter-country compensation payments might be viewed as an intermediate step in a process finally leading to the application of pan-regional tariffs. Unless pricing practices in place in the systems comprising the region are harmonised and the transmission charges applied in them are comparable and fair; however, agents in some of the participating systems may be subject to undue discrimination, and significant inefficiencies are likely to arise.

One last option involves ignoring the use which agents in each local system make of others' grids. Agents would simply pay the local transmission charge resulting from the allocation to local network users of their local grid costs. This option can only be regarded as acceptable in a traditional framework where power exchanges among areas within a region are very rare and mainly related to the provision of emergency support for one of the systems in the region. Under these circumstances, ignoring the cost of the use of local grids by external agents would be justified by the small economic value involved.

10.5 Harmonisation of Rules

Harmonisation of a number of features of the operation and regulation of local systems in a region is requisite to the establishment of a functional regional market. Harmonisation measures should ensure the creation of a level playing field, which is vital to the integration of local systems. This should result in supplying the regional load by the most efficient generators at any given time. Thus, the value placed by market agents on scarce system resources would be maximised. At the same time, unfair discrimination among agents from different systems would be avoided. Other rules in need of harmonisation refer to the operational details involved in cross-border transactions. The paragraphs below discuss the features that must inexcusably be harmonised.

To facilitate cross-border transactions, the amount of interconnection capacity among systems should be computed using a network model that is common to all these systems. As explained in Sect. 10.3, this would maximise the amount of interconnection capacity available for trade. In addition, in regions where the regional dispatch is not computed centrally, local market opening and closing hours should also be harmonised. If the regional energy dispatch coexists with organised local markets, their opening and closing hours should be compatible and the interface between the two types of markets should be carefully designed to prevent significant energy dispatching inefficiencies. Obviously, many more features need to be harmonised if a centralised trading platform exists (format of bids, a common bid-matching algorithm, etc.).

The creation of a level playing field calls for harmonising a number of features. For instance, the network charges to be paid by agents in different systems should avoid introducing any market distortions (as explained in Chap. 6, this would occur if volumetric, €/MWh, transmission charges were used for generators). At least, a common ceiling should probably be established for the per unit network charges to be paid by both generators and industrial consumers to prevent significantly different transmission charges from distorting competition among these agents.

A region-wide decision should also be reached on the existence, and if this is the case the format and the need for harmonisation of any capacity instruments meant

to enhance generation adequacy and firmness (defined and examined in detail in Chap. 12) at local or regional level. The major questions to be examined are:

- Is the diversity of capacity instruments at local level something to be concerned about, because it may result in inefficiency or loss of security of supply in the functioning of the regional market?
- What is the scope of the impacts of these potential inefficiencies: only at investment level or could they also concern the efficiency of functioning of the short-term wholesale market?
- If potential inefficiencies do exist, is there any regulatory measure that could be applied to make sure that they do not materialize?

Support payments for each type of renewable technology should be harmonised to ensure that the construction and operation of a region's renewable generation facilities match the region-wide distribution of renewable resources. However, other local criteria—such as promotion of local industry and employment, development of depressed rural areas or energy independence objectives—may be in conflict with the strict efficiency objectives, therefore requiring some sort of compromise. This issue is dealt with at length in Chap. 11.

Other local regulatory charges may exist, which depend on specific features or legacies in each system. Some regional power systems may be part of a broader economic institutional arrangement—like the European Union—and transparency in the allocation of these costs may be required, so that hidden subsidies among consumers are avoided and electricity prices do not distort competition among energy intensive industries. Regulatory charges to generators have the potential to distort competition, in particular if volumetric charges (€/MWh) are applied. A particularly relevant case is the cap-and-trade scheme in place in the European Union, resulting in a common charge in €/ton of CO_2 emitted in the generation of electricity. This charge is common throughout the EU Internal Electricity Market and has no distortionary effects, but penalises those generation technologies with higher CO_2 emissions per MWh production.

10.6 The EU's Internal Electricity Market

By way of illustration, this last section of the chapter prior to the conclusions describes the current situation in one of the world's most prominent regional markets: the EU's Internal Electricity Market (IEM). Further details and updates—as this is an on-going process—can be found in the websites of the European Commission (IEM), ACER and ENTSO-E.[4]

[4] IEM at the EC: http://ec.europa.eu/energy/gas_electricity/index_en.htm, ACER: http://www.acer.europa.eu/Electricity/Pages/default.aspx, ENTSO-E: https://www.entsoe.eu/resources/network-codes/.

10.6.1 General Market Design

The IEM presently comprises 27 countries, 25 EU members plus Norway and Switzerland. The European legislation on the IEM consists of Directives and Regulations that have to be transposed into the national laws, plus guidelines and network codes providing directions for the implementation by the Member States. The 1996 original Directive on common rules for the internal market in electricity, marked the beginning of the formal process of market integration for electricity at the EU level. The electricity Directive (2003/54/EC) established the basic regulatory framework for this regional market [13]. In the light of the dysfunction in the internal market in electricity, the European Commission considered it necessary to redefine the rules and measures applying to that market in order to guarantee fair competition and appropriate consumer protection, resulting in what is called the third regulatory package [15].[5]

The institutions in charge of enacting the legislation governing the IEM are the European Parliament, the Council and the Commission. The Agency for the Cooperation of Energy Regulators (ACER) was created by the Third Energy Package to foster cooperation among European energy regulators, and to ensure that market integration and harmonisation of regulatory frameworks are achieved in respect of EU's energy policy objectives. Other relevant Europe-wide organisations include the European Network Transmission System Operators for Electricity (ENTSO-E), the Council of European Energy Regulators (CEER), the European Regulators Group for Electricity and Gas (ERGEG, the ensemble of all independent national regulatory authorities, which advises the European Commission on the consolidation of the Internal Market for Electricity and Gas) and the Association of European Energy Exchanges (EuroPEX). All the Internal Electricity Market stakeholders meet periodically in a series of fora known as the Florence Regulatory Fora, where they discuss the steps to be taken next to further the integration of national markets.

The driving force behind the IEM is the regional agents' ability to freely negotiate power transactions. Access to the transmission grid is open, contingent upon the payment of the local access charge. In systems with vertically integrated companies, the transmission grid must be operated by a legally separate entity (some exceptions may exist, but they will be subject to strong conditions). In addition, agents must have regulated third-party access to the transmission grid in all Member States. Electricity companies must keep separate accounts for all their distribution and transmission activities to avoid discrimination, cross-subsidies or distortion of competition. Retail competition is mandatory, so that all consumers

[5] The most current legislation comprises Directive (2009/72/EC) of July 2009, Regulation (EC) No 714/2009 of July 2009 on conditions for access to the network for cross-border exchanges in electricity, Regulation (EU) No 1227/2011 on wholesale energy market integrity and transparency (REMIT). There is also a Directive (2005/89/EC) of January 2006 concerning measures to safeguard security of electricity supply and infrastructure investment.

can switch suppliers for gas and electricity, and suppliers must provide clear explanations of terms and conditions. Main European pieces of legislation laying down these guiding principles include [6–8, 13–15].

Market operation within each country, however, is regulated by the competent national authorities or institutions and many important decisions, like the generation mix, are made nationally, rather than by the Union.[6] This principle is known as subsidiarity. In line with this principle, the region has no single system operator, since each country's system is operated by its respective SO (several SOs in some cases). Similarly, the region boasts several multinational wholesale markets comprising two or more neighbouring countries (e.g. NORDEL, MIBEL or SEM). The coordination between the existing wholesale markets, national SOs and regulatory authorities has been limited to date. Some progress in the integration of national systems has been made in the context of the existing regional initiatives. These are separate fora, one for each of the sub-regions into which the European market has been divided, where the major issues relating to sub-regional integration are discussed.

However, until now, the IEM has not been a truly integrated regional market, but a collection of national markets required to abide by some very basic common rules and guidelines, and agents—generators and loads—with the option to buy and sell anywhere, in theory at least. There are indeed still a lot of obstacles to overcome before a truly integrated market for electricity is achieved in the EU: the implementation of common models for transmission capacity allocation and congestion management across the EU, the efficient and secure integration of intermittent generation linked to renewables, the implementation of a stable regulatory framework for the development of new trans-European network infrastructures, the lack of ratemaking transparency, the high degree of vertical and horizontal integration in the industry in some sub-regions and the insufficient interconnection capacity between most countries, among other topics.

The Third Energy Package has given to process the new tools and a new boost. The EU aims to fully integrate national energy markets by 2014, and progress has been made in the creation of a power exchange common platform that is intended to be the basis for seamless transactions EU-wide in the long term, the day-ahead market horizon and shorter. Through this common bid matching platform, the bids (mostly simple bids, with some exceptions) made in a Member State will actually interact with bids from the other Member States, only subject to the limitation of the interconnection capacities. In principle, the underlying network model will be oversimplified, just a coarse zonal pricing model.

Work still to be done includes aligning national market and network operation rules for gas and electricity, as well as making cross-border investment in energy infrastructure easier. ACER and ENTSO-E work on a tight timeline to draft the

[6] Exceptions exist, as the mandatory target for penetration of renewables in energy consumption: 20 % by 2020, or the cap-and-trade scheme to reduce CO_2 emissions by at least 20 % on 2020 with respect to 1990 emissions; both measures have strong implications on the composition of the electricity generation mix.

Framework Guidelines and Network Codes that will have to be approved by the Member States by what is called "the Comitology procedure", so that there will be a fully integrated IEM by 2014.

Framework Guidelines and Network Codes aim at providing harmonised rules for cross-border exchanges of electricity. The areas in which the framework guidelines and network codes are to be adopted are: network security and reliability, including technical transmission reserve capacity for operational network security; network connection; third-party access; data exchange and settlement; interoperability; operational procedures in an emergency; capacity allocation and congestion-management; trading with regard to the technical and operational provision of network access services and system balancing; transparency; balancing, including network-related reserve power; harmonised transmission tariff structures including locational signals and inter-transmission system operator compensation, and energy efficiency regarding electricity networks.

In the absence of nodal pricing, bidding zones will be defined and, according to the Framework Guidelines on Capacity Allocation and Congestion Management for the operation of the IEM in the long term, day-ahead and intra-day timeframes, TSOs should adopt a flow-based method for the calculation of transmission capacity among TSOs in highly meshed networks, whereas a simpler method is allowed for less meshed networks. The process for definition of biding zones foresees a regular monitoring and possible redesign of the bidding zone configuration with the aim to increase the overall market efficiency. The capacity allocation method for the day-ahead market will be price market coupling and for the intra-day timeframe it will be continuous trading. For the long-term timeframe Physical or Financial Transmission Rights might be adopted. The design for the day-ahead and intra-day timeframe requires that capacities are physically firm, whereas for the long-term capacity financial firmness is required.

Integrated electricity balancing market is the last building block of the internal market for electricity. Integration of balancing markets is very challenging due to significant differences in the existing national balancing market arrangements, the very different balancing resources being available locally, and the fact that balancing has a strong impact on the security of supply and network operation. Here, the objective is to integrate the national balancing markets by fostering cross-border competition and more efficient balancing, while safeguarding the security of supply. Particular attention will be given to demand response and renewable energy sources with the aim to increase their participation in balancing markets.

The Third Energy Package tasks ACER with monitoring the internal markets for electricity and gas. The Agency's report will assess the internal markets for energy, and in particular concentrate on retail prices (including compliance with consumer rights), network access (including grid access for renewable energy sources) and on any barriers to the IEM.

The situation in the region with respect to a few major regulatory issues is described briefly below. The issues to be discussed include investment in new transmission infrastructure, grid access, transmission tariffs and security of supply.

10.6.2 Main Regulatory Features of the IEM

Investment in new transmission facilities

Generally speaking, the interconnections among national transmission systems in the IEM are weak. The resulting systematic congestions divide the system into several sub-regional load pockets. Differences among wholesale prices in these load pockets are significant. Given the conflicting conditions for EU-wide system planning, the current system of checks and balances between a centralised institution formally in charge and the preservation of the autonomy of the countries to decide what is built in their territories is quite satisfactory, although not everybody agrees.

The European Network of Transmission System Operators for Electricity (ENTSO-E) has the mandate to propose a non-binding EU-wide ten-year network development plan (TYNDP), including a European generation adequacy outlook, every 2 years, starting in 2010 [12].[7] The TYNDP of ENTSO-E builds on the national investment plans prepared by the transmission system operators (TSOs). ENTSO-E is presently also preparing a separate long-term plan for the electricity transmission network in 2050.

As indicated above, the expansion plan prepared by ENTSO-E is not mandatory. Only the EU Member States, with their national regulatory authorities and system operators, make final decisions on the transmission facilities to be built in their territories. But ACER is in charge of evaluating the plan prepared by ENTSO-E and of verifying that the national transmission expansion plans are consistent with the EU-wide plan by ENTSO-E. This might be enough to put some pressure on non-compliant Member States, so that the final result is satisfactory. If ACER identifies inconsistencies, it recommends amending the national plan or the EU-wide TYNDP as appropriate.

Once the TYNDP is positively evaluated and its consistency with national plans is assessed, ACER monitors its implementation, as well as the implementation of infrastructure projects that create new cross-border capacities. If ACER identifies inconsistencies between the EU-wide TYNDP and its implementation, it investigates the reasons and makes recommendations to TSOs, national regulatory authorities and other competent bodies, with a view to implementing the investments.

The European Commission has declared that some transmission projects have high priority, and is making sure that external barriers to their construction are removed. The goal is that all the administrative authorisations should be obtained in a single window and in a maximum prescribed time. This is important, since the construction of new lines is usually plagued with administrative difficulties, due to the wide variety of regulatory arrangements across the region and environmental concerns.

[7] https://www.entsoe.eu/system-development/tyndp/ [12].

Grid access

Grid connections must be provided wherever requested. If connection to the node in question is not possible, the agent submitting the request must be given an alternative. Grid congestion management guidelines specify that scarce transmission capacity must be allocated among market agents by efficient, coordinated, market-based methods. Significant progress has recently been made in the context of regional initiatives, as well as in the framework of collaboration between ENTSO-E and Euro-PEX in connection with the allocation of cross-border transmission capacity on most of the borders inside each sub-region. Market splitting and coupling are now being used, for instance, on the day-ahead time scale to coordinate capacity allocation on many of the borders within each sub-region. Moreover, ENTSO-E and Euro-PEX have put forward proposals for capacity allocation in both the long and in the very short term, along the lines of the methods described in Sect. 10.3. Capacity is nonetheless still being allocated on some borders using inefficient and uncoordinated methods such as pro rata division of the volume requested by each agent or distribution on a first come, first served basis. More importantly, substantial coordination- and harmonisation-related obstacles lie in the way of integrated congestion management in several sub-regions.

Initiatives have been taken to further the short-term creation of an integrated Europe-wide energy and capacity auction, as explained before. ENTSO-E and EuroPex, in collaboration, have worked to develop a regional (Europe-wide) centralised implicit auction. Any method to be finally adopted by the deadline of 2014 must be compliant with the Framework Guidelines and Network Codes.

Transmission tariffs

Instead of computing Europe-wide transmission tariffs, the countries and institutions involved have agreed to implement a system of inter-system or inter-TSO compensations (ITC) whereby each nation is credited for the use of its grid by others and charged for the use of other nations' grids. Once the final net amount to be charged to each country is determined, the national authorities can compute the local transmission tariffs for generators and consumers, which therefore account for the net compensations to be paid by the respective countries. Therefore, computing national transmission tariffs implicitly involves allocating the global net compensation to be received (difference between the compensation received and payments due for grid use) among local generators and consumers. EU Member States have agreed that some degree of harmonisation of national transmission tariffs is necessary to prevent regional market distortion, but the approval of the final rules is still pending. Nor has any final agreement been reached on a permanent method for computing inter-system compensation for the use of network infrastructure, while many countries in the region deem the present method to be inefficient and unfair.

The general hierarchical approach adopted by the IEM for allocating the cost of regional lines is sound and should remain. Inter-TSO compensation can efficiently

allocate the cost of regional lines, at least at the country level, with no need for a single system of tariffs across the entire region. Further progress must be made, however, in several areas.

- The principles on which the inter-TSO compensation method is based must be economically and technically sound to provide a reasonably accurate estimate of the use made of each regional line by each country. The inter-TSO compensation method presently in place, based on the computation of the impact of the transit through each country on its national grid, fails to comply with this requirement.
- Certain aspects of the methods applied by countries to allocate the regulated cost of the network to local users should be harmonised. Specifically, an agreement should be reached on the fraction of this cost to be paid by generators and large-scale consumers, as well as on the format (€/kWh, €/kW or annual lump sum) of the final network charge to generators and consumers. Presently, the average transmission charges paid by producers in each Member State must be within prescribed ranges, in order to limit the potential distortion of the temporary allocation method.

The present ITC mechanism also provides compensation for the costs of losses incurred by national transmission systems as a result of hosting cross-border flows of electricity.

Security of supply

The EU Directive (2005/89/EC) of January 2006 establishes measures aimed at safeguarding security of electricity supply so as to ensure the proper functioning of the EU internal market for electricity, an adequate level of interconnection capacity between Member States, an adequate level of generation capacity and balance between supply and demand.

Several countries in Europe have implemented, or are in the process to implement, a diversity of regulatory instruments that try to address their concerns regarding the adequacy and firmness dimensions. Examples are the capacity payments in place for a number of years in Spain and Ireland, or the capacity instrument in Greece, the reserved peaking capacities in Sweden, the safety net mechanism in The Netherlands, the future adoption of reliability options in Italy, the ongoing market reform in the UK or the current discussions on a capacity scheme in France. So far the European institutions have not adopted any measure or issued any guideline regarding whether it is convenient or not to harmonise these efforts. However, given the reasonable concerns about future generation investments in European countries—because of the financial difficulties associated to the economic crisis and the uncertainty regarding the regulatory-driven level of penetration of clean technologies and the future regime of functioning and prices in an environment of strong presence of renewable generation–, the issue of the convenience of some harmonisation measures of the adequacy and firmness schemes in the European electricity markets cannot be ignored any longer.

A true EU-wide approach to adequacy and firmness requires reliance of any country A in any generation capacity located in another country B that has been committed to provide guarantee of supply in A. This reliance is defeated if country B may call back the contract with country A of generators in their territory in case there is a supply crisis in B. Obviously, the problem with firmness only arises when both countries have a supply crisis, since only then the contracted generation in B cannot be replaced by anything else. What should prevail in this case: the contract or the potential request of the regulator in B to suspend any exports while demand in B cannot be totally met? It is clear that a true security of supply for electricity at EU level will only happen when import and export physical contracts have priority over any domestic demand without such contracts. This seems to be the direct interpretation of Article 4.3 in the Security of Supply Directive: "In taking the measures referred to in Article 24 of Directive 2003/54/EC *(it refers to measures to be adopted in emergency situations)* and in Article 6 of Regulation (EC) No 1228/2003, Member States shall not discriminate between cross-border contracts and national contracts". Unfortunately, most electricity laws of the Member States have explicit clauses maintaining that exports to other countries will be interrupted in case of a domestic emergency of supply. And these provisions have been applied whenever there has been the occasion for it.

Properly addressing security of supply at EU-wide level requires an answer to the three questions posed in Sect. 10.5:

- Is the diversity of capacity instruments at local level something to be concerned about, because it may result in inefficiency or loss of security of supply in the functioning of the regional market?
- What is the scope of the impacts of these potential inefficiencies: only at investment level or could they also concern the efficiency of functioning of the short-term wholesale market?
- If potential inefficiencies do exist, is there any regulatory measure that could be applied to make sure that they do not materialise?

Regarding the first question, one may think, in a first approximation, that the size of the large and medium size European countries and the relative uniformity in the generation mix—or at least the technologies setting the marginal market price—in most countries do not justify the need for an EU-wide capacity mechanism. Under these conditions, extending the size of the national markets to a European dimension would have a small beneficial impact of reducing the safe adequacy margins of installed generation capacity over the system peak demand.

However, this uniformity is quickly disappearing because of the strong levels of penetration of wind and solar generation in some countries. Major developments of renewable generation, such as the North Sea off-shore wind or Desertec in Northern Africa, may profoundly alter the generation patterns in some European countries and create local surpluses and deficits of installed generation capacity that could be used by other countries. In the medium and long term, the variability

of wind and solar may create local temporary deficits of electricity production in some countries that could only be met by very active demand response and recourse to generation in other countries. Well-established rules of generation support at EU-level would be very useful in this context.

Several issues have to be considered when answering the second question. Well-designed capacity instruments (i.e. regulatory measures meant to address firmness and adequacy issues) should not interfere with the functioning of short-term markets. However, disparity in the adoption of capacity instruments in different countries will result in loss of efficiency in the deployment of installed capacity. There are basically three types of capacity mechanisms that might be considered: (a) capacity payments, as in Spain or Ireland; (b) purchase or long-term contracting of peak capacity to provide security reserves, as in Sweden or The Netherlands and (c) market mechanisms to provide some regulated amount of generation capacity with or without firmness commitments, as in the UK, Italy or France. Extension of the national schemes to the participation of foreign market agents would probably have a filtering effect and the inferior approaches (a) and (b) would probably disappear. Still, the (c) type of approaches could come in many different flavours. If different countries choose different flavours (i.e. different adequacy and firmness levels, definitions of emergency conditions, economic terms, etc.) incompatibilities will appear that will prevent the use of the existing and future generation capacity to meet some ideal global EU requirements in the most efficient way.

Several levels of possible harmonisation requirements are offered next, in the format of market design principles, starting from the lighter ones, in increasing level of EU intervention:

(a) Mandate, or just encourage, that any capacity mechanism that a Member State (MS) decides to create is open to agents of other MSs. This principle seems to be consistent with the spirit of the abovementioned Article 4.3 of the EU Directive on Security of Supply.
(b) Make sure that the regulations and the coordination between MSs do not allow that any given adequacy or firmness value of a power plant is sold twice.
(c) MSs with capacity mechanisms should coordinate their implementation.
(d) Some level of harmonisation of the adopted capacity mechanisms would be desirable, in order to reduce inefficiencies.

A more ambitious design of a EU-wide capacity mechanism would go beyond principle (D) to establish a single harmonised instrument for the entire EU electricity market.[8]

[8] Note that principle (a) could be understood, particularly for countries at the European periphery, as just an extension of a multiplicity of individual national capacity instruments beyond their respective MSs borders, and therefore not the same thing as the—conceptually, at least—more efficient single EU-wide capacity mechanism.

10.7 Conclusions

The benefits of regionally integrating national or local electricity markets are likely to be far greater than the integration costs. For the integration to be successful; however, the regional approach must strike a balance between respecting the main features of the regulation of the local electricity markets and achieving the desired level of regional coordination.

The single system paradigm, which should maximise the net social benefit region wide, should be the benchmark for the most prominent features of regional market regulation. Such features—which require some level of harmonisation to be determined in each particular case—relate to regional grid expansion, the assignment of interconnection capacity amongst the region's areas, the allocation of the cost of the regional grid to its users, the computation of electricity prices, the application of support schemes for the furtherance of renewables, the design of incentives for the installation of new generation capacity, the type and structure of regulatory charges in place in each area and certain operational issues such as the management of system reserves or gate closure times in local markets.

The main design options for most of these features of regional market operation have been outlined in the foregoing and their strong and weak points discussed. Given the diversity of regional market situations and the lack of well-established experiences in this field, it is not possible to give detailed general-purpose recommendations. Nonetheless, from the single system paradigm and the contents of this chapter, a number of high level guidelines can be derived:

- Local markets rules have to reach a minimum level of harmonisation to allow efficient trading at regional level. The technical aspects are very dependent on the adopted regional integration approach: a common trading platform (as in the EU), seamless bilateral coordination between independently dispatched local systems (the trend in the US Eastern Interconnection) or a higher level regional market that manages the surpluses and deficits of the local markets (MER in Central America). In the very short term, coordinated intra-day, balancing and regulation markets would enhance system security and reduce operation costs.
- The design of short-term wholesale regional electricity markets must include a reasonably accurate representation of the effects of the transmission network. Nodal prices are the benchmark here; it has been implemented in the MER and it is a key feature of the US local markets to be integrated. This is a weak point in the IEM of the EU, only partly mitigated by the estimation of interconnection capacities by TSOs and the explicit and implicit network capacity auctions.
- The expansion of the transmission network should be planned at regional level. At least, a regional entity should be responsible for assessing the social value of investment proposals made by national systems or market agents. Most of the regional network reinforcements should be designed and built under regulated arrangements, without ruling out the possibility of merchant lines. Network costs—in particular for recent and new investments—should be allocated to the

beneficiaries of these investments. A hierarchical decomposition of this allocation process along the lines of the present approach in the IEM (first assign the costs of the regional network to the local systems; then respect whatever internal allocation method is chosen by the national or state regulator at local level in a second step) would facilitate transmission expansion considerably.

The foregoing discussion may be the basis for blueprinting regional market regulation suitably adapted to the situation in the respective region. One such blueprint has been described here by way of illustration for the design of the European Union's Internal Electricity Market.

References

1. ERGEG (2005) The creation of regional electricity markets. http://www.ergeg.org/portal/page/portal/ERGEG_HOME/ERGEG_DOCS/ERGEG_DOCUMENTS_NEW/ELECTRICITY_FOCUS_GROUP/ p 123
2. MIT (2011) Transmission Expansion, in MIT Study on the 'Future of the Electric Grid'. pp 268. http://web.mit.edu/mitei/research/studies/documents/electric-grid-2011/Electric_Grid_Full_Report.pdf
3. Olmos L (2006) Regulatory design of the transmission activity in regional electricity markets. PhD Thesis in Instituto de Investigación Tecnológica, Pontificia Comillas University, Madrid
4. Pérez-Arriaga IJ, Gómez T, Olmos L, Rivier M (2011) Transmission and distribution networks for a sustainable electricity supply. In: Galarraga Ibon, González-Eguino Mikel, Markandya Anil (eds) 'Handbook Of Sustainable Energy' Edward Edgar Publishing. United Kingdom, Cheltenham
5. Sioshansi FP (2006) Electricity market reform. An international perspective, Elsevier

References on the Functioning of Specific Regional (and National) Markets

6. Agency for the Cooperation of Energy Regulators (2011a) Framework guidelines on electricity grid connections. Agency for the Cooperation of Energy Regulators. pp 14. http://acernet.acer.europa.eu/portal/page/portal/ACER_HOME/Communication/News/110720_FGC_2011E001_FG_Elec_GrConn_FINAL.pdf Accessed 20 July 2011
7. Agency for the Cooperation of Energy Regulators (2011b) Framework guidelines on capacity allocation and congestion management for electricity. Agency for the cooperation of energy regulators. pp 15. http://acernet.acer.europa.eu/portal/page/portal/ACER_HOME/Communication/News/FG-2011-E-002%20(Final).pdf. Accessed 20 July 2011
8. Agency for the Cooperation of Energy Regulators (2011c) Framework guidelines on electricity system operation. Agency for the cooperation of energy regulators. pp 28. http://www.acer.europa.eu/Official_documents/Acts_of_the_Agency/Framework_Guidelines/FG%20on%20Electricity%20System%20Operation/FG-2011-E-003_02122011_Electricity%20System%20Operation.pdf. Accessed 2 Dec 2011
9. Belpex (2010) Market Coupling. http://www.belpex.be/index.php?id=4

10. De Dios R, Sanz S, Alonso JF, Soto F (2009). Long-term grid expansion: Spanish Plan 2030. In: CIGRE conference 2009. http://www.cigre.org
11. EOR (2005) Reglamento del Mercado Eléctrico Regional (RMER). Libro III del RMER de la Transmisión, Ente Operador Regional (EOR)
12. ENTSO-E (2010) Ten year network development plan, TYNDP. https://www.entsoe.eu/index.php?id=282
13. EP (2003) Directive 2003/54/EC of the European Parliament and of the Council concerning common rules for the internal market in electricity and repealing Directive 96/92/EC
14. EP (2003a) Regulation (EC) No 1228/2003 of the European Parliament and of the council of 26 June 2003 on conditions for access to the network for cross-borderexchanges in electricity. http://europa.eu.int/eurlex/pri/en/oj/dat/2003/l_176/l_17620030715en000100 10.pdf:10
15. EP (2009) Directive 2009/72/EC concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC
16. ETSO (2009) Development and implementation of a coordinated model for regional and inter-regional congestion management, a report by the association of European Transmission System Operators. http://www.entsoe.eu/fileadmin/user_upload/_library/publications/ etso/Congestion_Management/ETSO-EuroPEX%20report-25-02-09.pdf
17. ETSO (2001a) Co-ordinated Auctioning: A market-based method for transmission capacity allocation in meshed networks, a report by the Association of European Transmission System Operators (ETSO). http://www.etsonet.org/ upload/documents/Coordinated%20Auctioning.pdf: 22
18. ETSO (2001b) Key concepts and definitions for transmission access products, a report by the association of European Transmission System Operators, http://www.entsoe.eu/fileadmin/user_upload/_library/publications/etso/Congestion_Management/Key%20Concepts%20for%20Transmission%20Access%20Products.pdf
19. Federal Energy Regulatory Commission (2011) Transmission Planning and Cost Allocation by Transmission Owning and Operating Public Utilities, Order No. 1000. http://www.ferc.gov/industries/electric/indus-act/trans-plan/fr-notice.pdf. Accessed 21 July 2011
20. KEMA Consulting, ISA (2004) Informe de diseño de detalle del Mercado Eléctrico Regional. Versión Final pp 124. http://www.enteoperador.org/ (in Spanish)
21. PHB-HAGLER/SYNEX (2000) Diseño General del Mercado Eléctrico Regional. A report by Mercados Energéticos S.A., PHB-HAGLER BAILLY and SYNEX Ingenieros Consultores for the Unidad Ejecutora del Proyecto SIEPAC

References for In-depth Regulatory Study of Specific Topics

22. Chao H-P, Peck S (2000) Flow-based transmission rights and congestion management. Electr J 13(8):38–59
23. Gilbert R, Neuhoff K, Newbery D (2004) Allocating transmission to mitigate market power in electricity markets. RAND J Econ, RAND Corp 35(4):691–709
24. Hogan WW (1992) Contract networks for electric power transmission. J Regul Econ 4(3):211–242
25. Hogan WW (2002) Financial transmission right formulations. Working document. http://www.ksg.harvard.edu/hepg. Center for Business and Government, John F. Kennedy School of Government, Harvard University, Cambridge
26. Olmos L, Pérez-Arriaga IJ (2007) Evaluation of three methods proposed for the computation of inter-TSO payments in the Internal Electricity Market of the European Union. IEEE Trans Power Syst 22(4):1507–1522

27. Pérez-Arriaga IJ, Rubio-Odériz F, Puerta Gutiérrez JF, Arcéluz Ogando J, Marín J (1995) Marginal pricing of transmission services: an analysis of cost recovery. IEEE Trans Power Syst 10(1):65–72
28. Pérez-Arriaga IJ, Olmos L (2005) A plausible congestion management scheme for the internal electricity market of the European Union. Util Policy 13(2):117–134
29. Pérez-Arriaga IJ, Olmos L (2007a) Compatibility of investments in generation, transmission and distribution. In: François L (ed) Competitive electricity markets and sustainability. Edward Elgar pp 230–288

Chapter 11
Environmental Regulation

Pedro Linares, Carlos Batlle and Ignacio J. Pérez-Arriaga

Analysts may confuse things that are countable with the things which count.

John P. Holdren, 1980

In coincidence with the second wave of regulatory reforms of electricity markets (the reform of the reform), during the past decade the concern for the environmental impacts of electricity—clearly led by climate change—has also become widespread and will require a very demanding environmental and energy policy, reducing emissions and at the same time supporting the massive deployment of clean energy technologies (renewables, capture and storage of CO_2, nuclear or biofuels, plus measures of energy efficiency and savings). We have seen that, in particular, this will require an almost complete decarbonisation of the power sector, which, on the other hand, will have to feed with clean electricity much of the transportation and the heating sectors. The implications for electricity regulation are staggering, as environmental concerns have become as prominent as efficiency and security of supply: regulatory support of the several types of clean technologies, while they might need it; an in-depth review of the existing pricing and incentive instruments for electricity generation so that they are adapted to the new technology mixes that can be anticipated; demand response and how energy efficiency and conservation could be encouraged; rethink system operation and network planning at transmission and distribution levels; review of network remuneration schemes and design of instruments to promote innovation in new technologies. These challenges are reviewed in the final chapter of this book.

This chapter focuses on the environmental impacts of the activities involved in the supply and consumption of electricity and the regulatory measures that can be adopted to mitigate these impacts. Because of the relevance that electricity

P. Linares (✉) · C. Batlle · I. J. Pérez-Arriaga
Instituto de Investigación Tecnológica, Universidad Pontificia Comillas,
Alberto Aguilera 25, 28015 Madrid, Spain
e-mail: pedro.linares@upcomillas.es

C. Batlle
e-mail: carlos.batlle@iit.upcomillas.es

I. J. Pérez-Arriaga
e-mail: ipa@MIT.EDU

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_11,

production with renewable energy sources is expected to have in the future energy mix, a special section has been devoted to the regulatory aspects of these technologies. An increasing awareness of the operational challenges created by intermittent generation of electricity from renewable resources, such as wind and solar, has led to increased scrutiny of the public policies that promote their growth and the regulatory system that maintains operation of a reliable and economically efficient power system.

Environmental regulation constitutes a well-developed body of knowledge, and this chapter for the most part presents a regulatory analysis that applies to any generic industrial sector, and uses air pollution as a representative environmental undesirable effect that requires regulation. This material can be applied to the electric power sector and extended to other adverse environmental impacts.

11.1 The Need for Environmental Regulation

The production and use of electricity are associated with significant environmental impacts, the most serious of which is arguably global warming. The environment is affected in many other ways by the generation, transmission and use of electric power, however, which are listed in Table 11.1.

The mere existence of an impact does not necessarily justify its regulation. Why, then, should the environmental impact of electricity be regulated? Two different but related answers come to mind, based on the paradigms of economic efficiency and strong sustainability.[1]

The first answer refers to an issue discussed earlier in this book in Chap. 2: the fact that environmental impacts are externalities that affect public goods. In other words, in the absence of regulation, environmental impacts inflict costs on agents other than the responsible parties, who fail to take these costs into consideration. Since their origin is to be found in the public good nature of the environment, property rights cannot readily be assigned to these goods; therefore, the externality will not be corrected by the market.

The second has to do with the existence of physical limits to the use of the environment: pollution cannot go unchecked, because it may alter the ecological balance and, therefore, the conditions for life on Earth, or it may exceed some threshold beyond which catastrophic events are triggered. Electricity production has not historically been regarded as a significant cause of such a situation. The substantial rise in electricity consumption throughout the world, however, driven both by population growth and rising per capita electricity use, has led to major escalation in the environmental impact of power production and use, which may come to contribute significantly to the factors that are stretching the planet to its physical limits.

[1] According to the weak sustainability, man-made capital can replace natural capital. At the other extreme, the strong sustainability view sustains that natural resources are irreplaceable.

Table 11.1 Environmental impact of electricity, by fuel type

Fuel	Stage	Consequences
Coal and lignite	Fuel extraction and transport	Emissions: SO_2, NO_x, particulate matter, CO_2, radioactivity
		Liquid effluents: acid water
		Solid waste: mining waste
		Land use: subsidence, visual impact, habitat alteration
		Noise
	Generation	Emissions: SO_2, NO_x, particulate matter, CO_2, heavy metals
		Liquid effluents: chemical products, thermal pollution
		Solid waste: slag, ash
		Land use: visual impact
Oil	Fuel extraction and transport	Emissions: SO_2, NO_x, CO_2, H_2S, CH_4
		Liquid effluents: chemical products, fuel spills
		Land use: subsidence, visual impact, odour, habitat alteration
		Noise
	Generation	Emissions: SO_2, NO_x, CO_2
		Liquid effluents: chemical products, thermal pollution
		Land use: visual impact
Natural gas	Fuel extraction and transport	Emissions: SO_2, NO_x, CO_2, H_2S, CH_4
		Land use: visual impact, accident risk, habitat alteration
		Noise
	Generation	Emissions: NO_x, CO_2
		Liquid effluents: chemical products, thermal pollution
		Land use: visual impact
Nuclear	Fuel extraction, preparation and transport	Emissions: SO_2, NO_x, particulate matter, CO_2, radioactivity
		Liquid effluents: drainage water, radioactive emissions
		Solid waste: radioactive mining waste
		Land use: subsidence, visual impact, habitat alteration
		Noise
	Generation	Emissions: radioactivity
		Liquid effluents: chemical products, thermal pollution
		Solid waste: radioactive materials, spent fuel
		Land use: visual impact
	Waste management	Radioactive emissions
Hydro	Generation	Land use: hydrological changes, habitat alteration, accident risk, visual impact, microclimate alterations
Solar	Generation	Solid waste: heavy metals contained in components
		Land use: visual impact
Wind	Generation	Land use: habitat alteration, visual impact
		Noise

(continued)

Table 11.1 (continued)

Fuel	Stage	Consequences
Biomass	Collection and transport	Emissions: SO_2, NO_x, particulate matter, CO_2
		Liquid effluents: non-point pollution
		Land use: visual impact, erosion
	Generation	Emissions: NO_x, particulate matter
		Liquid effluents: chemical products, thermal pollution
		Solid waste: slag, ash
		Land use: visual impact
Electricity	Transmission and distribution	Emissions: electromagnetic fields
		Land use: habitat alteration, visual impact

Both considerations lead to the same conclusion: the need to regulate the environmental impact of the electricity industry, both to ensure economically efficient resource allocation and to keep human activity well within the Earth's physical limits. This issue, in turn, may be broken down into two separate problems: determining the optimal or safe level of environmental impact, and defining the policies required to achieve that level, ideally at the lowest possible cost. These will be the topics covered in this chapter. As ascertaining the safe level is a much broader issue better left to physical, chemical and biological analysis,[2] only the economically efficient level is discussed here. The second part, i.e., the analysis of the instruments necessary to reach the level defined, is common to both approaches.

In tackling these subjects sight should not be lost of a basic tenet of economic regulation: regulation only makes sense when the market failure to be corrected is larger than the possible regulatory failure, i.e. a mistaken target or implementation of an inadequate regulatory instrument that distorts the efficient outcome of the affected industrial sector, even when externalities are taken into account. The sheer amount of technical information required for environmental regulation and the uncertainty involved in many of the benefits and costs make regulatory failure in this area more likely. Environmental policy should consequently be designed with this in mind at all times.

Three major broad topics should be addressed in the analysis of regulatory measures to mitigate the environmental impact of the electric power sector: (a) climate change and the emission of greenhouse gases, CO_2 in particular; (b) energy conservation and savings; (c) support for development and deployment of clean technologies for electricity generation and other enabling technologies, such as transmission networks and smart grids. This chapter introduces the basic economic concepts that broadly apply to these three areas of activity, but details on the specific regulatory instruments are only given for the support of electricity generation using renewable energy sources, a subtopic of item (c) above.

[2] One such assessment can be found in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) [25], which analyses the safe level of CO_2 concentrations in the atmosphere.

The chapter begins by looking at how the optimal level of environmental impact is determined for a generic industrial sector, in Sect. 11.2. In Sect. 11.3 the basic elements and the advantages and disadvantages of the various instruments available to achieve the desired level of impact are reviewed, both from a theoretical point of view and also considering second-best settings, and still applied to a generic industrial sector. Section 11.4 critically reviews the technology policies that are used in the power sector to support generation with renewable energy sources and the challenges faced by power systems with a strong penetration of renewables. Lastly, some consequences of environmental regulation on electricity markets and the economy in general are examined. Throughout most of the chapter (with the exception of Sect. 11.4 and part of Sect. 11.5), and for the sake of clarity, air pollution is used as the typical environmental impact that needs to be regulated. However, all of the points discussed can readily be applied to other adverse effects.

11.2 Determining the Optimal Pollution Level

The first step for regulating pollution is to determine its appropriate level. Given that pollution is a public "bad", or rather, pollution reduction is a public good, this reduction, and thus the determination of its optimal level, is subject to the same shortcomings as identified for other public goods if left to the market (see Chap. 2). The optimal level must therefore be set by the regulator. How should this be done? As it was described in Fig. 2.16, the efficient level of pollution must be set at the point where the marginal social cost (damage) of pollution is equal to the private marginal benefit of the activity generating the pollution, as shown again in Fig. 11.1 in a highly idealised representation.

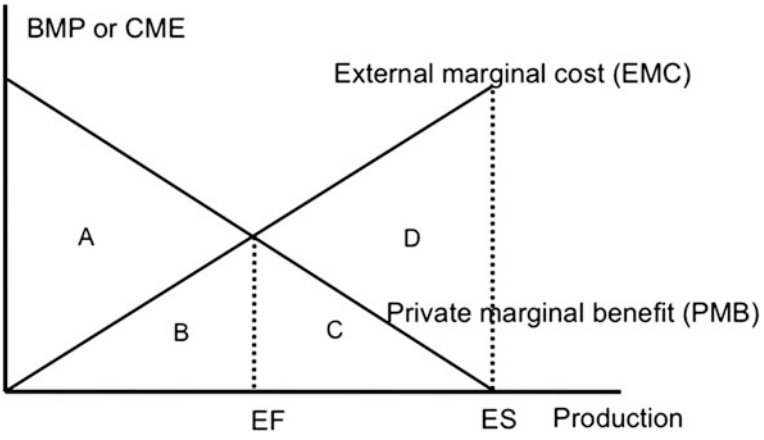


Fig. 11.1 Efficient pollution level

Along this chapter, for the sake of simplicity, we will represent marginal cost or benefit functions are straight lines (that is, we assume that cost or benefit functions are quadratic, and therefore the marginal functions are linear).

In a situation with no externalities, the industry's or the firm's operating output would be ES, determined by production factors and the market. This requires the following assumptions, made here for the sake of simplicity and without loss of generality:

- The firm (or all firms in the industry) operate(s) at the point ES where its marginal revenue is equal to its marginal cost: in other words, where its private marginal benefit (PMB) is zero. This assumption is valid either for competitive or oligopolistic/monopolistic markets, as long as the system is well adjusted. Under these circumstances, the total income earned by the firm is the area under the PMB curve (or A + B + C).
- This private marginal benefit results from the market equilibrium, and therefore includes changes in consumers' utility. That is, it is the private marginal benefit that maximises welfare in the market, not accounting for external costs.

This firm's activity may, however, be detrimental to other agents, and that harm may remain constant or increase, linearly or otherwise, with output. If it is assumed to rise quadratically, it can be represented by the external marginal cost curve, EMC. The total costs inflicted on other agents are equal to the area under EMC (or B + C + D).

For the production level ES, if the social benefits, SBs, are factored in as private income less external costs, they can be represented as area A less area D (A + B + C − B − C − D). This SB is not the maximum benefit possible, however, because if the firm's output (and, therefore, the external costs incurred) declines, social benefit rises. To what extent is this true? Social benefit may rise until the marginal social benefit equals zero, i.e., to the point where external marginal cost is equal to private marginal benefit. This is the point shown as EF (EF, for efficient) in Fig. 11.1, which identifies the maximum social efficiency sought. At this point, social benefit (area A) is maximal, while area B represents the optimal level of external cost or externality, A + B is the optimal private benefit and C + D the externality that is not socially desirable and must somehow be eliminated.

How is this point where marginal private benefits equal marginal social costs reached? First, marginal social costs need to be assessed (assuming that marginal private benefit can be determined by looking at the firm's books). However, assessing marginal social costs is no easy matter: it involves quantifying many impacts that are not readily monetised,[3] and it also entails aggregating social preferences for these impacts. Indeed, determination of preferences is a social choice problem and as such liable to be addressed by a number of mechanisms

[3] Many research projects have been devoted to the assessment of the externalities of electricity production. See [15] or [35] for a summary.

(voting, Kaldor-Hicks criterion or social welfare functions), none of which is flawless [30].

Once the marginal social cost curve is determined, social cost must be incorporated into the firm's decisions. This can be done in centralised or decentralised fashion. In the centralised approach, the regulator compiles all the information on the marginal social costs and marginal private benefits and finds their intersection, as shown in the figure, using the methods mentioned earlier. The decentralised approach consists of charging every polluter for the marginal damage produced by their pollution. As this affects their marginal private benefits, the end result is the production of an efficient level of pollution. In fact, the latter approach solves the two problems posed in the introduction: the market attains the efficient level of pollution by itself.

Although these two approaches may seem to be the "right" and easiest ways to decide on the efficient or optimal pollution level, they are not implemented in practice, for several reasons. The first is the sheer amount of information required, some of which is difficult for the regulator to obtain. As mentioned earlier, plotting the marginal social cost and marginal private benefit curves may be especially arduous. The decentralised approach is even more challenging, as the exact damage caused by each polluter must be identified.

The second reason is that economic efficiency is not the only factor driving environmental policy: other considerations, including uncertainty, distributional effects and administrative complexities, may lead the regulator to choose a level of pollution that is not necessarily the most efficient. The outcome may deviate even farther from strict economic efficiency when the optimal level is determined in international negotiations, such as in carbon dioxide emissions, or when it is imposed at a higher governmental or administrative level (one example is the environmental regulations mandated in European countries by the European Commission).

As a result, the level of environmental impact or pollution to be attained by an industrial sector is usually determined politically and, therefore, may not be the most economically efficient option. In turn, this also means that the instruments implemented to achieve this level of pollution do not ensure efficient resource allocation. Nonetheless, these instruments can still be required to be cost-efficient, i.e. to achieve the pollution level established at the lowest possible cost. This is analysed in the section below.

11.3 Instruments for Environmental Regulation

The next step after defining the pollution level is to select the instrument to be used to reach it. The instruments presently in place for this purpose can be divided into three major categories:

- Command and control instruments.
- Economic instruments.
- Other instruments.

Their basic characteristics, advantages and disadvantages are discussed below, now in the general context of the energy sector.

11.3.1 Command and Control Instruments

Traditionally, environmental costs in the energy industry have been internalised by setting standards, also called command and control methods. These consist of establishing and subsequently enforcing environmental restrictions for the operations conducted by each generating company or plant. Such restrictions are defined in terms of the quantity EF in Fig. 11.1, which is then applied to the individual polluters, generally through a simple method, such as using the same standard for all power plants with a certain technology. This simplification poses problems and, as discussed below, restrictions may be established in a variety of ways.

The most restrictive control mechanism is the *technology standard*, which imposes the use of a given technology (normally the best available—BAT—or the best available technology not entailing excessive costs—BATNEEC) and limits the environmental impact of this best technology to the established target. This is, for example, the approach used by the European Directive on Integrated Pollution Prevention and Control (96/61/EC), which makes the issue of an operating licence contingent upon installing BATNEEC.

Fuel quality standards are somewhat more flexible. They seek to control pollutant emissions by establishing the quality of the inputs used. For instance, Directive 2003/17/EC limits the sulphur content in petrol and diesel fuels for transport to control vehicle emissions in Europe.

An even more flexible mechanism consists of setting *emissions standards*, which impose limits on the quantity or concentration of pollutants generated by the source. These standards can be expressed in a number of ways in practice: maximum concentrations of pollutant emitted, total quantity of pollutant emitted in a given period of time, or minimum performance of cleaning or pollution reduction equipment. They can, moreover, be applied either individually to each pollution source or to a series of sources. Their flexibility stems from the fact that they allow polluters to use whatever means within their reach to limit emissions, rather than restricting action to a given type of technology or fuel. Examples of this sort of regulation are not difficult to find: many countries require power plants to limit their SO_2, NO_x or particulate matter emissions to a given value.

One drawback to these standards is that they restrict emissions but take no account of their impact, which may vary widely depending on site-specific characteristics. This leads, for instance, to setting the same emissions limits for power plants, while the emissions from one may cause greater environmental harm than

the other due to its location near a high density urban area or an ecosystem of great value. This situation is obviously not wholly rational.

To avoid such pitfalls, other control instruments have been proposed, such as *environmental quality standards* that take the more reasonable tack of limiting the impact on recipients. The drawback is that they are more difficult to establish and monitor, since they depend on the location of the polluting activity, the distribution of possible recipients and geographic or meteorological conditions. These standards are usually set at the local level, but regional regulations (e.g. relating to ground-level ozone) have also been enacted.

Lastly, *licences* may also be regarded as a command-and-control method. In fact, they may be the strictest standard, since they can prevent a power plant from being built or commissioned. Many types of licensing procedures have been devised, but the most widespread for environmental permits is the administrative procedure known as Environmental Impact Assessment (EIA), under which every plant or unit must receive a favourable judgement from an environmental agency or regulatory body. The licence may be awarded permanently or be subject to periodic reviews. Licences are also a way to group a series of requirements into a single procedure.

In general, all the command and control methods discussed are relatively simple to implement, even in complex environments, and also easy to monitor, with low administration costs.[4] In addition, since firms pay only the cost of conforming to the standard rather than the total cost of the damage caused by their pollution, they perceive these methods to have a lower cost and consequently find them more acceptable.

The command and control approach also entails several disadvantages, however. One is that it does not encourage the reduction of emissions below the established limits, for it creates no incentive to reduce pollution further, or to lower the costs of abatement technologies through technical innovation, for instance. A second is that, as it covers not all the damage caused but only the cost of control, it fails to send out the right price signals (the full price rise needed to internalise the externality), losing the effect that a higher price might have on reducing demand. The major drawback to command and control instruments, however, is their lack of equimarginality, explained in greater detail below.

Standards are usually set across the board for all power plants, regardless of their abatement costs, essentially because the amount of information required to tailor standards to the specific technical and economic characteristics of each polluter could be enormous, particularly if the number of affected facilities is large. Since the marginal cost for the different power plants of reducing pollution to the level required by the standard will differ in general (i.e., lack of equimarginality), then the cost of achieving a given pollution level is higher with this approach than with economic instruments, as discussed in the next section.

---

[4] Compared to other environmental regulation instruments. All environmental policies naturally entail certain administrative costs, which must be weighed against the environmental benefit they provide.

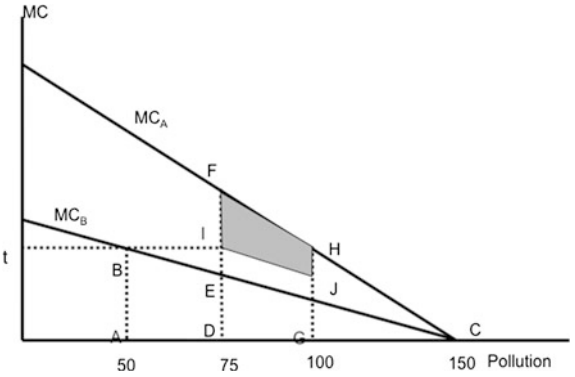


Fig. 11.2 Cost-effectiveness of economic instruments

11.3.2 Economic Instruments

Economic instruments are designed to address the shortcomings of command and control methods, particularly their lack of equimarginality (and the resulting lower cost-effectiveness). The example shown in Fig. 11.2 (with idealised curves, again) illustrates how this is achieved.

The example shows an industry with two polluting firms, A and B, each emitting 150 units of pollutants, whose marginal abatement costs (MAC) follow an upward pattern, different for each firm. To reduce total pollution from 300 to 150 units, the firms can be required to restrict their emissions to 75 units each (e.g. using an emissions standard). In this case, the total cost of reduction is the sum of the areas DFC and DEC.[5]

Assume, however, that these firms are allowed to trade their emissions quotas. As the figure shows, firm B's marginal cost to reduce an additional unit of emissions at point D (75) is lower than firm A's. In fact, firm A might pay firm B to reduce one unit of emissions so that A can increase its quota and still save money. Both firms would benefit from this arrangement. Trading continues until the marginal cost of reducing pollution reduction is the same for both firms, when the incentive to trade disappears. At this point, company A would emit 100 units and company B 50. The overall cost would be the sum of areas ABC and GHC. In short, the end result is the same, 150 units of pollutants, but with a savings in cost equal to area IFHJ. The outcome is that equimarginality is attained by allowing emission quota trading, and the desired level of pollution is reached at a lower cost than under command and control procedures.

This same result can be achieved by levying a tax (t). If a tax is applied, both firms will operate at a point where their marginal abatement cost is equal to the tax: they will be keen on reducing their emissions (and not paying the tax) until the

[5] Note that, by increasing pollution from 75 to 100 units, firm A has reduced its cost by the area DFHG, while by reducing pollution from 75 down to 50, firm B has increased its cost by the area ABED. Since ABED and DIJG have the same area, the net saving in cost is IFHG.

cost of reduction exceeds the amount of the tax. At that point, they would rather pay the tax than reduce their emissions.

Thus, the same objective can be achieved with two types of economic instruments: price instruments (taxes, generally) and quantity instruments (tradable quotas). Both have the same advantages over standards: they ensure equimarginality with fewer information requirements; they ensure that polluters pay for all their pollution and therefore, include its cost in the product price, with the added benefit of a reduction in demand induced by the higher price (usually termed as the output channel); and they create incentives for further pollution reduction (this is known as *dynamic efficiency*). However, taxes are less popular than standards because of their higher perceived cost and may also be more difficult to monitor and implement than the latter. It is also necessary to compute the right value of the tax.

Furthermore, the aforementioned advantages of price and quantity economic instruments over standards may not exist in all situations. First, if all firms are equal, standards yield the same outcome and maintain equimarginality. Second, if the impact of regulation on the price of the product is small, or if the product is price-inelastic, the output remains unaffected. Lastly, some situations may not require dynamic efficiency.[6] The second circumstance may be more or less common in the power industry, but the others are not: the power sector features several technologies and fuels, making emissions reduction costs highly variable, and the elasticity of demand is quite low. This is why economic instruments are becoming more popular in the environmental regulation of the power industry.

The following discussion identifies the differences between price and quantity instruments.

11.3.2.1 Price Instruments

The most common price instrument is a Pigouvian tax. This tax owes its name to A.C. Pigou, who first proposed it in 1920 [42]. The intention is to send polluting firms an economic signal to reduce output to a previously established level.[7] This signal takes the form of a tax, which is included in the firms' production function. Contrary to some popular interpretations, revenues from this Pigouvian or environmental tax need not be used for environmental purposes (see the discussion on the use of tax revenues in a later section).

---

[6] Dynamic efficiency refers to the existence of long-term incentives for further reducing emissions. Including incentives for innovation in technologies or changes in behaviour.

[7] In addition to this use for centralised internalisation, taxes may also be used to internalise directly a polluter's external cost as a decentralised way to achieve the economically efficient level of emissions, as seen in Sect. 11.2. However, as noted there, this is not the usual approach because of the vast amount of information required. Although determining the emissions level for each facility is achievable, computing the external costs specifically for each facility is very complex because it depends on the distribution of the receptors of the damages.

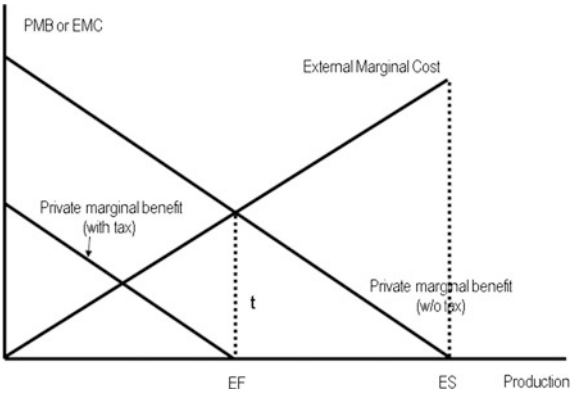


Fig. 11.3 Pigouvian tax

We start from the situation described in Fig. 11.1. Figure 11.3 shows that levying a tax of an adequate value t (assumed to be constant) reduces firms' private marginal income so that the private marginal benefit (including the tax t) of the firm becomes zero for a production level EF, which is now the optimal value sought by the firm. The tax should therefore be set as the external marginal cost at the point of maximum social efficiency.

Under this approach, the polluting agent is penalised twice, once in the form of the tax itself, and the other as profit not earned as a result of a decline in output.[8]

The final impact of the tax on the amount of pollution and the various agents depends on a number of factors, but the scheme fails to guarantee that the established pollution level will ultimately be met. Success in this regard depends on the uncertainty around the marginal cost and marginal benefit curves (dealt with in greater detail below), but also on other factors that may alter market operation. If for instance firms are able to cover the cost of the tax by raising the price of their products without affecting demand (due to small consumer elasticity, as is the case with electricity in the short term), they will have no incentive to reduce the level of their pollutant emissions.

Instruments of this type may vary: to avoid the "double penalty" problem referred to above, and also to change the impact of the tax on polluters, taxes may be levied only on outputs in excess of a certain limit rather than on total volume.[9] They may also be applied selectively: some sectors or polluters may be exempted from the tax. The basis for computing the tax may be actual pollution, or when this is too difficult, the associated inputs. This facilitates monitoring, although it may be less efficient, as it does not engage all available channels for lowering consumption.

Such variations usually exist to avoid increasing the tax burden globally or for certain sub-industries. It should be noted, however, that the tax burden need not

[8] This is not necessarily undesirable, as it allows the externality to be fully internalised.

[9] In which case they resemble a standard, which only imposes costs on excess pollution.

necessarily be any heavier if green tax reform is implemented, as proposed by some authors (see the section on second-best regulation).

Another price instrument is the *environmental subsidy*. In this case, instead of polluters being taxed, non-polluters are subsidised. This changes the relative prices of production technologies (just as a tax does), potentially raising the proportion of less polluting technologies in the technology mix, thus reducing pollution. In general, subsidies have the same characteristics as environmental instruments such as taxes, but with a very important difference: subsidies typically result in lower prices of whatever product is produced, and therefore lead to higher, rather than lower consumption and output. To reduce pollution to the same extent as taxes, subsidies need to be higher. In addition, subsidies require governments to raise additional funds, whereas taxes are an additional source of revenue. Subsidies are clearly much more politically acceptable, however.

11.3.2.2 Quantity Instruments

The other type of economic instrument is **tradable quotas**, first proposed by Crocker [12] and Dales [13]. Under this approach, the total allowable pollution (corresponding to EF in Fig. 11.1) is established and allocated to the agents[10] in the form of emissions permits or allowances. A market is created for such allowances, in which they may be traded by the agents in keeping with their respective interests. Firms with low pollutant abatement costs, for instance, will find it in their best interest to reduce emissions and sell their allowances to companies with higher costs. Trading may sometimes be restricted within certain geographic zones (also called "bubbles"), so that pollutant emissions are not driven to only one or a few specific areas where abatement costs are lowest, thereby preventing the generation of overly high local concentrations.[11]

Obviously, for this market to be effective, certain practical requirements must be met.

- Free-riding must not be allowed, and strong deterrents must be established in the form of high fines for companies producing emissions without allowances, for otherwise polluters would see no point in buying them. The amount of such fines also effectively caps the price allowances can command, affording governments a vehicle for controlling the allowance market price.
- The total allowable emissions to be distributed must be lower than the level of pollution emitted under normal conditions, for if it is greater, the price of the allowance will be zero, as nobody will need to buy one.

[10] Agents may be producers or consumers. Indeed, the decision about whether to impose the obligation on the former (upstream) or on the latter (downstream) may also have important consequences on the outcome, see e.g. [23].

[11] This is important for pollutants such as SO_2 or particulates, high concentrations of which may cause acute local or regional damage. It is irrelevant for global pollutants, such as carbon dioxide.

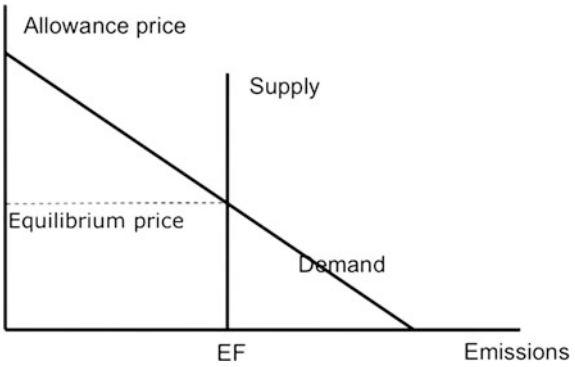


Fig. 11.4 Market for emissions allowances

The diagram in Fig. 11.4 explains how this market works. The supply of allowances, corresponding to quantity EF in the figure, is perfectly inelastic, since it corresponds to the emissions quota set by the regulator, whereas demand is represented by the marginal abatement cost. The amount a company would be willing to pay for an emissions allowance is any sum up to the cost of reducing its emissions by an equivalent volume. The point where the two lines intersect is the market price for allowances.

Assuming the absence of both transaction costs and the endowment effect,[12] this market price will be exactly the same regardless of how the permits are allocated, although the distribution of wealth will naturally change. Permits may essentially be allocated in two ways, as listed below.

- Allocation on the grounds of historic pollutant emissions rates (known as grandfathering) acknowledges companies' established rights, for which they are not penalised, but may also constitute a barrier to new entrants if this allocation is updated[13] (except in some markets where the cost of the allowance may be easily recovered from the market, such as the electricity market). It also represents a transfer of wealth from consumers to existing producers.
- Allocation by means of an initial auction affords all market agents the same opportunities and raises funds that can be subsequently reinjected into the industry or used to reduce other taxes (see the discussion below). The problem is the transfer of wealth from existing operators to the regulator, and the staunch resistance to this alternative that would be expected from the former.

Interestingly, the European Union's Emissions Trading System (EU ETS) only permitted a small fraction of allowances (5–10 %) to be auctioned in the early stages, with the remaining percentage being given away for free, probably to

[12] The endowment effect [27] refers to a situation in which firms or individuals place a higher value on goods given to them than on goods acquired through trade. It results in fewer transactions, as the seller values its goods (here the allowances) more than the buyer does.

[13] That is, if allocation is continuously revised based on previous emission levels, therefore converting a fixed payment into a variable income.

increase acceptability, but subsequently proposed unrestricted auctioning. However, full grandfathering may not be required to compensate firms: for example, Goulder [22] found that 85 % of allowances in a carbon trading scheme in the US could be auctioned without compromising cost recovery by existing firms. In electricity markets the rational behaviour of generators is to internalise the cost of allowances in their bids to the market, therefore raising the market price, regardless of the outcome of any prior allocation scheme. Therefore generators in electricity markets are able to recover all their costs without grandfathering.

Other complementary schemes, such as pooling and banking, may sometimes be implemented in conjunction with emissions trading systems.

In *pooling*, emissions producers form an association to jointly buy their required allowances. The device is used to facilitate administrative procedures but may also serve to distribute the cost of the allowance evenly among producers. This may distort the price signal, however, unless the cost is distributed on the grounds of each producer's need of allowances.

In *banking*, allowances acquired in a prior period (by purchasing more allowances than needed or reducing emissions more than required) can be used in subsequent periods (generally when the allowance price is expected to rise). Banking provides more flexibility by allowing emissions to be moved from one period to another, and therefore evens out allowance prices by levelling them over several periods (while this lowers compliance costs and also reduces the technology change signal, it also results in less volatile prices and more stability for investors). In addition, it helps to develop markets by promoting early period reductions that would not be attempted otherwise (when the allowance price is low). It also has implications for the possible exercise of market power in the allowance market, which must be addressed. In addition, banking may pose problems for flow-type pollutants (such as SO_2), since it may allow emissions to become highly concentrated in a very short time, possibly exceeding safety thresholds. This is a problem that in fact affects all tradable emission permits (or taxes), which are usually associated with lengthy use periods (at least a year) and allow the free movement of emissions within these periods. Banking simply strengthens this effect. When the risk of crossing a threshold exists, this mechanism should be made subject to, or replaced by, emission or air quality standards for the relevant time frame.

Tradable quotas are a very popular instrument for regulating the environmental impact of electricity generation. The Clean Air Act used them to reduce SO_2 emissions in the US, and many countries or areas are either currently considering or have already implemented (the European Union since 2005, several North East States in the US[14] and California in the future) this alternative to control CO_2 emissions.

[14] This is the Regional Greenhouse Gas Initiative, see http://www.rggi.org/.

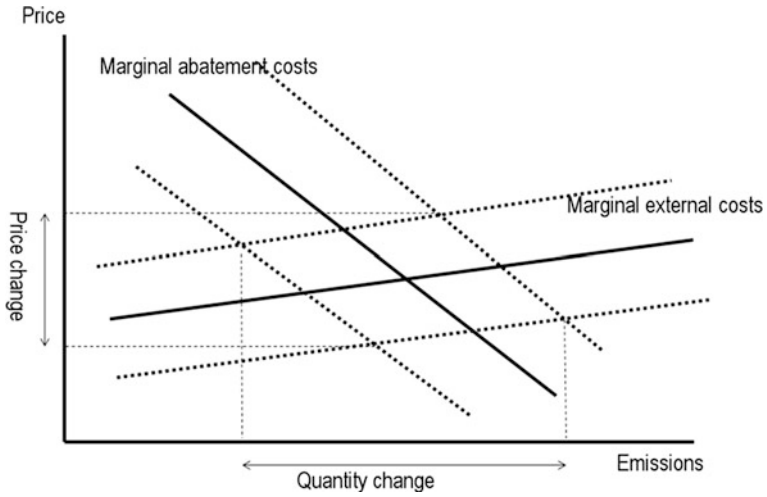


Fig. 11.5 Effects of uncertainty on the choice of instruments

11.3.2.3 Comparison of Price and Quantity Instruments

In theory, and statically speaking, the two mechanisms should be equivalent: the allowance price resulting from a quota should be equal to the optimum Pigouvian tax to achieve that quota. In practice, however, results may differ for several reasons.

The first is simplicity of implementation. Taxes may be the easier alternative, since they can use the existing fiscal system. Furthermore, no new trading allowance markets need to be created and the transaction costs are lower, particularly when many agents need to be regulated. However, new taxes are always difficult to accept by the public and, therefore, by the politicians that have to approve them.

The second is the need for information: when the goal is to reach a certain pollution level, the regulator may not have enough data about the polluters' private benefits to define the tax correctly. This also affects the speed of adjustment; even if the regulator is able to establish the appropriate tax, technological improvements may trigger a change. The ability to accommodate such changes is built into tradable quotas, which therefore require no adjustment.

Another difference has to do with distributional concerns: taxes can be earmarked, but with tradable quotas the impacts can be more precisely allocated, which may enhance their acceptability [16].

Lastly, a very important element is uncertainty. Weitzman [46] showed that when uncertainty exists (i.e. always), the difference between quantity- and price-based instruments generally depends on the relative slope of the marginal cost and marginal benefit curves. Figure 11.5 shows that when the marginal external cost curve is flatter (e.g. when the environmental benefit function is linear) and the

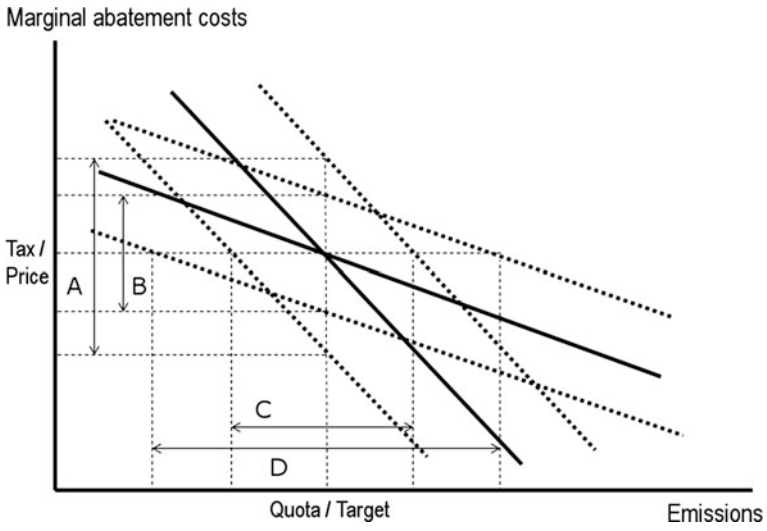


Fig. 11.6 Change in the amount of emissions for a given tax

marginal cost curve is steeper (as in the case of climate change, for example), price-based instruments are preferable, since it is easier to determine the right price than the right quantity (or viewed from another perspective, the margin for error due to uncertainty is larger for quantity instruments). In the figure, the solid line is the "expected" curve, while the dotted lines represent the shift in its position due to uncertainty.

Weitzman also contended that if uncertainty affects the slope of the curves, the use of quantity-based instruments is probably safer, since price-based methods may lead to disastrous consequences.

This analysis of uncertainty applies where the optimal pollution level is also sought. If that level has already been determined, the only relevant parameter is the absolute slope of the marginal cost curve. The flatter the curve, the more appropriate are quantity-based instruments, as even a small error in the calculation of the tax may result in a large variation in the amount of pollution emitted (compare C and D in Fig. 11.6). Conversely, if the curve is very steep, minor errors in quantity limits may generate great differences in allowance prices and, therefore, cost (compare A and B in Fig. 11.6).[15]

Another uncertainty-related issue is that price-based instruments limit the cost of the measures to be adopted, whereas quantity-based methods provide for the direct control of the amount of emissions.

Lastly, this analysis is based on a static situation. The results may vary in a dynamic scenario, since a price-based system may reduce risk for investors, for instance, and encourage higher investment in new technologies [37].

[15] This result may change if banking is allowed.

11.3.2.4 Hybrid Instruments

Hybrid instruments combine the features of price- and quantity-based instruments. They are usually proposed to reduce uncertainty for investors or consumers, and in some instances have proven to be more effective than "pure" price or quantity instruments [44]. The two most common hybrid instruments are the *safety valve* and the *indexed quota*.

The former is basically a tradable quota with a price ceiling. When the allowance price climbs "dangerously" high, the regulator may intervene, offering an unlimited amount of allowances at a fixed price (the price ceiling). In other words, safety valves are a tradable quota system until the price ceiling is reached, after which they revert to a tax.[16]

The indexed quota consists of replacing the fixed emissions quota with an emissions intensity target (usually linked to an economic indicator, such as GDP), so that in a growing economy, more allowances are distributed, lowering the allowance price.

A more detailed analysis of these instruments and how they work can be found in Newell and Pizer [38].

11.3.3 Other Instruments

In addition to "pure" command and control or economic instruments, others falling somewhere in between these two types may be used to regulate the environmental impact of the electricity industry. This section describes total cost dispatching, integrated planning, voluntary agreements and two-part instruments.

11.3.3.1 Total Cost Dispatching

This instrument is a sort of decentralised internalisation of externalities, so in some respects it might be included with the economic instruments. However, it differs in several ways: most importantly, polluters are not required to pay anything directly, unlike with taxes, and no explicit price is placed on pollution, since there is no trading. These features are more reminiscent of command and control than economic instruments.

Total cost dispatching consists of incorporating the external cost, if it can be calculated, or some type of "environmental adder" into the electricity dispatching algorithm. To decide which generating plants are to participate in supplying electricity at any given time, account is taken of this environmental adder or external cost, in addition to the economic costs of generation, and the dispatch order is established accordingly.

[16] Conversely, the regulator may also set a minimum price for the allowance, in order to provide more stability to the investors in clean technologies.

The major advantage of this mechanism for internalising externalities is speed, since it can be immediately applied to the existing generating plants, without waiting for new assets to come on stream. It is also simple to implement even in competitive markets, as electricity dispatching is usually centralised, which means that only the market operator's dispatching algorithm would have to be modified.

The chief drawback is that such speedy implementation may be traumatic, because altering the algorithm would mean that some generators would be dropped from the dispatching schedule and consequently unable to cover their fixed costs. Compensating for this *lucrum cessans* would add to system costs. Another shortcoming is that when the system has no overcapacity (i.e. if all power plants must be simultaneously operational), total cost of dispatching may change the merit order but not the final amount of pollution. Other problems have to do with the correct determination of adders (see [26], for example).

11.3.3.2 Integrated Planning

One alternative to solve some of the problems of total cost dispatching is to incorporate externalities or adders into the planning processes in the medium and long term. With this approach, decisions about future investments change, as not only economic costs but also environmental impacts are taken into consideration.

Known as *integrated resource planning*, this instrument has been in use for a long time in the US, although it has disappeared in most places where restructuring and liberalisation of the power industry has taken place. It gives the system time to adapt, and therefore minimises the economic impact on existing producers. However, it is almost impossible to use in liberalised markets, in which investment decisions are left entirely up to utilities, and such investment planning cannot be enforced. Moreover, the instrument shares some of the problems of total cost dispatching (e.g., calculating adders).

11.3.3.3 Voluntary Agreements

Voluntary agreements are either contractual or unilateral commitments by polluters to reduce their environmental impact. Such abatement may take place at the source, by decreasing pollutant emissions or lessening their impact (one recent example of the latter is the reforestation undertaken by some companies to compensate for their CO_2 emissions, so-called offsets).

The most common types of agreements are the integration of environmental concerns into business policy (usually through environmental auditing to ISO 14000 standards or similar), stricter reduction of pollutants than required by law, and environmental protection measures or measures to enhance efficiency. Green labelling systems or electricity disclosure requirements may also be regarded as voluntary agreements, since they may be freely implemented or otherwise by firms and consumers.

When these agreements are written up in contract form (usually with the public authorities), they are mandatory and infringements may be punished. When they are unilateral commitments taken on by the companies involved, they are voluntary, and therefore not subject to penalisation. In the latter case, they may be included in company environmental programmes.

Companies usually prefer these voluntary instruments, since they are less onerous for them and quicker to implement, given that they involve no specific law-making. Indeed, companies usually offer them unilaterally as a way to avoid more stringent regulation. It stands to reason that they would be expected to be less effective. However, even regulators may prefer them in some instances, such as when the cost of monitoring traditional regulation is very high, or when a given environmental impact is surrounded by considerable uncertainty. In these cases, the balance between the ability to regulate environmental impact and the desire for this regulation to be efficient and effective is tilted toward the voluntary approach.

In terms of efficiency, such instruments may be regarded to stand midway between emissions standards and economic instruments. They are also halfway between the two with respect to their advantages and drawbacks. Their positive features include flexibility, lower administration costs than other instruments, and straightforward enforceability, as they are self-imposed rather than forced onto the firms involved. They improve the corporate image of the companies signing them, which often enhances competitiveness. The disadvantages are that they may be viewed as a way of postponing measures that need to be taken, and they may arouse mistrust around compliance. They may also afford non-participating agents a competitive advantage (to prevent this, compensation may be established for participants). Moreover, like standards, they fail to encourage technological innovation, since they provide no incentive for polluters to reduce their emissions to levels lower than the statutory limits.

11.3.3.4 Two-Part Instruments

These instruments are combinations of command and control, economic and/or voluntary instruments. They are usually aimed at addressing the shortcomings associated with the "pure" instruments, particularly as regards administrative costs. Of the many examples found in the literature, Eskeland and Devarajan's [17] proposal may best illustrate this type of instrument. These authors acknowledged that the most effective way to address the externality generated by vehicle emissions would be to levy a tax on these emissions or to use tradable quotas. They nonetheless realised that such emissions, which depend on many factors, would be very difficult to track, and that installing real-time monitoring equipment on every vehicle would be much too costly. They therefore proposed replacing this optimal but highly impractical regulated solution with a two-part instrument: a tax on gasoline consumption and a fuel efficiency standard. The standard uses gasoline consumption as a close proxy for emissions, while the tax helps lower these emissions.

11.3.4 Second-Best Regulation: Environmental Regulation in Practice

Most of the foregoing on the relative advantages and disadvantages of environmental regulation instruments entails a series of assumptions that rarely hold, the most important of which are listed below.

- The economy has reached full efficiency.
- The only externalities to be corrected are environmental.
- Consumers and producers behave rationally.
- The regulator possesses all required information and pursues the public good.
- Distributional aspects can always be adjusted ex-post (according to the second theorem of welfare economics).

However, these five premises seldom concur. Economies are not usually efficient (e.g. due to preexisting distortionary taxes). Many other externalities are present (the most prominent ones for energy are energy security, knowledge spillovers, information asymmetries, network externalities and technology lock-in). Consumers' rationality is bounded, preventing them from making the most economically sound decision. Regulators have other incentives when performing their functions, and distributional concerns, which usually determine political acceptability, are critical and as important as efficiency. These problems are addressed briefly in the following discussion (for further details, see [31] for example).

Preexisting distortionary taxes change the way in which resources are apportioned in the economy, resulting in inefficient allocation. In general, environmental regulation exacerbates previous distortions, but it can also be used to correct them: revenues obtained from environmental regulation may be used to reduce distortionary taxes and therefore raise the efficiency of the economy as a whole. This is known as the double-dividend effect, which can be strong (where greater welfare is attained than without the environmental regulation alternative) or weak (when greater welfare is attained than without the environmental regulation involving no revenues that can be used to reduce other taxes).

The conditions requisite to a double dividend are set out below.

- Present distortions previously induced by the tax system must be substantial.
- The environmental tax burden should fall on economic agents not affected by any significant distortions.
- The environmental tax base must be large enough to avoid distortions.

The existence of this double dividend may influence the choice of instrument or its stringency. For example, many authors (e.g. [8]) have reported that in its presence, environmental taxes should be lower than when it does not exist. However, two elements must be considered when analysing the changes required with respect to the first-best result: first, the possibility of reallocating the revenue

(that is, using it to reduce other taxes), and second, the creation of scarcity rents through environmental regulation.

When environmental regulation creates scarcity rents (e.g. with a tax or a tradable quota system), such rents must be captured by the regulator to reduce other taxes; if they are not captured (when some sectors are exempt from taxes, or when emissions allowances are freely allocated), these rents may further distort the economy, in which case standards or subsidies, which do not create rents, may be better [20].

Another important issue in the energy industry is the existence of *externalities in technology markets* (knowledge spillovers, credibility problems and learning-by-doing), because of which environmental policies alone may not be enough to produce the required innovations in technology (dynamic efficiency). For example, if innovators cannot claim the benefit of their inventions, the incentive to innovate must be raised. This can be done either by modifying the environmental regulation (generally making it more stringent[17]) or, more usually, by addressing these externalities with other, more specific instruments, such as technology policies (see the following section). The usual conclusion, however, is that both environmental and technological policies may be required [18].

Information asymmetries and bounded rationality, which prevent economic signals from reaching their target with any efficiency, or which dull agents' awareness of these signals, also reduce consumer response to economic instruments. When these elements are present, standards may become much more effective and, therefore, sometimes more efficient than economic instruments. Another example of this is that downstream cap-and-trade systems may make allowance prices more salient to consumers than upstream ones.

Political "rules" and incentives also influence the theoretical choice of instruments. Governments prefer instruments that are easier to understand and implement, with hidden rather than explicit costs (so that the perceived cost of regulation is lower), and that afford them greater control over distribution. They also try to avoid taxes, because of their extreme unpopularity, which creates a strong bias in favour of standards. When they choose economic instruments, it is usually for new sources only [45]. Moreover, they prefer to use supplementary instruments such as technology policies to reduce the perceived cost of environmental policy. Regarding climate change, the absence of a broad agreement among the largest emitters of GHG makes it very difficult for a country or group of countries in isolation to establish a carbon tax or a cap-and-trade mechanism with the sufficient strength to be of any value in truly promoting clean technologies. The obvious reason is the immediate loss of economic competitiveness with respect to countries that have not implemented such measures, because of the substantial increment in the price of electricity and energy intensive commodities caused by these regulatory instruments.

[17] Or more predictable to investors, as with a minimum allowance price, for example.

Distributional concerns may also drive the choice of policy. The use of the instruments described may impact social or inter-industry equity. Whilst governments may have tools to correct such problems, it may at times be preferable to prevent them from arising. For example, instruments that entail raising electricity prices more than others (e.g., taxes versus standards) exacerbate the distributional impact of environmental policies (usually regressive, unless specifically designed otherwise) and therefore have lower preference.

Existing firms also have incentives to demand non-optimal regulation to raise additional barriers for new entrants or to generate extra rents [28].

Thus, the choice of the right policy instrument is a complex decision, which must take several criteria into consideration [21]: efficiency, equity, environmental effectiveness, political acceptability, administrative feasibility, even the possibility of fraud. A single instrument can hardly be expected to meet all these criteria while addressing the additional externalities and problems mentioned above.

One of the foremost conclusions, therefore, is that regulatory authorities need multiple instruments to deal with a variety of externalities and other second-best considerations. The right choice naturally depends on the environmental impact considered and on the institutional setting in place in the country or region.

For example, although some economists still insist on using a single instrument (usually a tax) to deal with climate change, an ideal climate policy in light of the foregoing should probably include a carbon price to be used as a benchmark (which should be established through an auctioned cap-and-trade system, with a safety valve to hedge against high prices), plus technology standards, technology policies, information and education policies and voluntary approaches. The intricacies of climate policy are analysed in depth in Aldy and Pizer [1] and Hanemann [23]. The same need to combine instruments applies to other environmental impacts; therefore, a coordinated approach to all of them is critical (see [33], for example).

In this setting, *technology policies* are acquiring an increasingly important role in reducing the environmental impact of the electricity industry: in addition to dealing with externalities in the innovation market, they also feature characteristics that make them appealing to policymakers. They consequently warrant a closer look, undertaken in the following section, with a focus on electricity generation from renewable resources. A similar case might be made for *energy efficiency* policies, but for lack of space they cannot be discussed in detail. Interested readers may wish to consult Linares and Labandeira [32].

11.4 Technology Policies for Renewable Electricity

Technology policies are known as such because they can be used to promote all types of technologies in need of support (e.g., wind, solar, carbon capture and sequestration, nuclear fission or fusion). A review of technology policies for low-carbon technologies can be found in Pérez-Arriaga, [39]. However, here the

analysis will focus on renewable electricity exclusively, which is the group of technologies currently receiving the greatest attention in the power sector and with a richer regulatory experience (e.g., [10]). A detailed review of the main instruments for renewable energy support can be found in Batlle et al. [6].

As described earlier, the basic rationale for technology policies is the existence of a diversity of market failures, with the consequence that, at the present time, a first-best environmental regulation, such as a tax on pollution or a cap-and-trade mechanism, may result in suboptimal investment in clean technologies, and also in suboptimal development of these technologies.

In the case of renewables this rationale is compounded by other policy objectives. For example, renewable energy technologies are typically based on local resources and thus contribute to energy security by reducing dependency. They may also represent an alternative for the industrial development of a country or region. In addition, technology policies make supply and demand curves for pollution reduction more elastic, thus lowering the perceived cost of environmental regulation. All this makes technology policy an effective vehicle for developing cleaner technologies, driving domestic technology and creating jobs. Finally, if renewables are considered as the major component in the long-term energy mix of a clean and competitive world economy [24] it seems wise to pay serious attention to them from the outset to facilitate the transition, instead of keeping the deployment of infrastructures that we know cannot take part in the long-term solution.

There are basically two types of technology policies: *technology-push* and *market-pull*. The choice of one or the other depends on the characteristics of the technology. If it is in the early stages of development, with substantial potential for improvement and cost reduction, the option should be to support R&D activities. When the technology is reaching a precommercial stage, improvements may only be made if the market for the technology grows significantly and allows for economies of scale. This can be achieved by market-pull policies, which are assessed below. Given the parallels between these policies and the economic instruments for environmental regulation described above, the former have been classified into *price*, *quantity*, and *voluntary instruments*.

Price, quantity and voluntary instruments are *direct methods*. These methods include investment supports, such as capital grants, tax exemptions or reductions on the purchase of goods, as well as operating support mechanisms, i.e. price subsidies, obligations, tenders and tax exemptions on production, see for instance Commission of the European Communities [11]. Here we shall focus on direct methods.

There are also *indirect methods*, i.e. implicit payments or discounts as well as institutional support tools that include: research and development funding, below-cost provision of infrastructure or services—costs of technical adaptations such as shadow connection charging, see Auer et al. [2], or costs of imbalances and ancillary services in general—, and positive discriminatory rules—such as regulations facilitating grid access, guaranteed purchase, dispatch priority, or advantageous network tariffs together with net metering, favourable building codes, etc.

11.4.1 Price Instruments

Price instruments provide economic incentives for electricity generation with renewable energy sources and are environmental subsidies, previously discussed in Sect. 11.3.2.1. These incentives for renewable energies can be awarded in the form of *investment subsidies* (generally used in the earliest stages of technology development, whether directly or indirectly through tax exemptions for imports, for example), or as an extra payment for the energy generated, usually called *feed-in tariffs (FIT)* or *premiums*. Another variant (in the US) is the *production tax credit (PTC)*, a tax exemption awarded for each kWh produced with renewable energy. The advantage of the latter two incentives for renewable production is that they truly encourage the generation of energy, while investment subsidies may result in inappropriate maintenance or operation or even in the abandonment of power plants once the subsidies have been cashed in.

The instrument works as follows: the regulator establishes the payment deemed necessary to attain the required output from renewable sources (by guaranteeing an acceptable rate of return for the investment) and then allows the market to operate freely. As discussed earlier with environmental subsidies, these instruments do not find as much political opposition as taxes, although they are not as opaque for the general public as other instruments (e.g. tradable green certificates) and may find some problems of social acceptability; also, they generally fail to send a complete economic signal to reduce consumption.

Payments for feed-in tariffs and premiums may be designed in several ways, as listed below.

- A *feed-in tariff* (FIT) consists of a fixed price per renewable energy generated, which includes the subsidy. Sometimes this fixed price is linked to the price paid by end customers, typically 80–90 % of this amount, as a sort of avoided cost of energy supply. The support is guaranteed for a long time, ranging from 10 up to 30 years, and the amount may depend on the technology (wind on-shore or off-shore, solar PV, concentrated solar power etc.), plant size or capacity factor. FIT for new facilities can be scheduled to decrease over time either at a pre-determined rate (so as to stimulate innovation) or according to the capacity that gets installed (to avoid over capacity). A thorough analysis of feed-in tariff design options can be found in Klein et al. [29].
- A *premium* (which may also be expressed as a fraction of the end customer tariff) is a payment (a fixed amount per kWh produced, in principle) to renewable generation on top of the market price for electricity, which changes with time. Similar to FIT, the premiums are valid for a prescribed long time period. Here, the remuneration is more uncertain than with a FIT, but there is an incentive to produce when the power system needs it most (strongly correlated with higher prices). In some countries this scheme is associated with the obligation to participate in the electricity market. As with the FIT, the amount of the

premium may depend on facility characteristics, and also on the electricity market price (in this case, expressed as a cap-and-floor or a contract for differences).

The advantages of price instruments for technology policies (compared to quantity instruments) are listed below.

- The investor security inherent in establishing revenues in advance lowers financing costs. This feature has made FIT and, to a point also premiums, both the most effective and, contrary to some expectations, the cheapest approach to furthering renewable energy [2]. These schemes are most suitable for technologies that have moved beyond the R&D phase but that have not reached market maturity and a strong presence in the system.
- Transaction and administrative costs are lower. Barriers to entry are minimised since investors do not have to find electricity buyers for their energy.
- Regulators will try to adapt the value of the subsidy to technology improvements for future investments and producers may take advantage of delays (planned or not) and asymmetry of information to profit from cost reductions obtained via innovation efforts and learning by doing.
- Information on the amount of government spending for this item may be provided in advance, but only if there is a cap on the total amount of power that can receive support.
- These instruments provide, fairly straightforwardly, for differential treatment of the various technologies involved or of the same technology at different sites. However, this can be also achieved with separate targets for different technologies with quantity measures.
- FIT, unlike premiums, do not contribute infra-marginal capacity for incumbent generators that operate both traditional generation assets in the electricity market as well as renewables, thus reducing their potential market power.
- Premiums, unlike FIT, do not suppress market price signals. Incentives to renewable generators to adjust their production according to day-ahead market prices makes sense in principle for technologies that are "fully dispatchable", like biomass or concentrated thermosolar with storage. In any case, participation of all renewable generators, even the intermittent or non-fully dispatchable ones, in the provision and charges for ancillary services and output forecasting should be encouraged.

There are, however, drawbacks to this approach as well.

- Incentives must be constantly updated to keep up with technological improvements and other cost factors. When such updating is not performed—a frequent occurrence—, substantial inefficiencies may ensue, with large disparities in the incentives for the different technologies. And it is very challenging to determine the right remuneration levels.
- This same need for updating may create a certain degree of regulatory risk for the renewable industry, however, since the regulator may over- or underestimate

the premium with every change. Updating should naturally never be retroactive, so that no risk exists for the investors in the existing facilities. Note, however, that usually FIT and premiums are not contracts—with well-defined legal guarantees—but just a regulatory instrument that is only backed by a regulatory commitment. Several instances of retroactive changes in price mechanisms have already taken place.

- Lastly, as in any price mechanism, the policy may not achieve the quantitative objectives pursued, as this depends on the free decisions of the investors and on other market factors beyond the regulator's control. Overachieving the target is also a possibility, giving the potential for quick deployment of some of these technologies, as it has been the case in several countries.

11.4.2 Quantity Instruments

The two basic types of quantity instruments for renewable energy promotion are *tradable green certificates* and *renewable energy auctions*.

Just as economic incentives can be likened to environmental subsidies, *tradable green certificates* (also known in the US as *renewable portfolio standards, RPS)* are based on the same approach as tradable quotas: a target quota is set by the regulator, to which agents have to comply by presenting certificates of purchase or production of renewable energy. These certificates can be traded on a secondary market. In this case, the certificates are usually awarded per unit of electricity produced with renewable sources (thus eliminating the complexity of the allocation of allowances). The quota can be technology-specific (this is called banding) or include several technologies. The banding provision is set to avoid that all investments happen for the least expensive renewable technology.

In the certificate market, the actors required to comply with the renewables quota (electricity generators, distributors or retailers, depending on the specific scheme devised) buy certificates from renewable energy producers. Since these producers still sell their energy on the electricity market, the price of the certificate tends to be the difference between the marginal long-term cost of the renewable technology (i.e. the total cost of production of the last unit of electricity needed to meet the quota) and the electricity market price. The certificate price can therefore be likened to the premium described earlier.

These systems are usually fitted with safety nets: a penalty is normally envisaged for non-compliance, so the maximum cost of the certificate, and thus the maximum total cost of the system, is known in advance. Sometimes a minimum price is also established to guarantee some degree of profitability for renewable facilities.

Here also, advantages and drawbacks can be identified. The major advantages are listed below.

- The amount of electricity to be produced with renewables is a known quantity (with obvious favourable implications for power system management).

- The market is entrusted with achieving efficiency by constantly incorporating technological change, so the target can be reached at a lower cost.

The drawbacks to the scheme are as follows.

- Price (depends on market conditions and can be volatile) and volume (as the quantity is established by the regulator) risks must be assumed by producers, a characteristic that may discourage investment. Such risks may certainly be significant, since a larger than expected renewable energy output, over which a small producer has no control, may drive the certificate price down to zero unless a minimum price is set. This has resulted in very large risk premiums in some countries, raising the cost of renewable generation with this method. This is the approach that has been used in several European countries, with scant success. However, the US version of this instrument, the renewable portfolio standards (RPS), in use in many states and under state-specific regulation, avoids the risk problem since the regulators accept that retailers sign long-term contracts with renewable generators at a mutually agreed price and then pass through the cost of these contracts to the regulated consumer tariffs.[18] This appears to be a superior version of this instrument and it becomes close to the renewable energy auctions method, to be discussed below.
- Certain practical difficulties and transaction costs are encountered: a market would have to be established for each technology, along with a suitable certification mechanism.
- Another drawback may be the possible appearance of market power when an electricity system has only a few large renewable energy producers, with the concomitant loss of efficiency. As in the case of premiums, ownership of renewable generation by any large incumbent utility will increase its market power under the tradable green certificates scheme.

Renewable energy auctions constitute another quantity instrument, but with some of the advantages of price mechanisms. These auctions are held at the initiative of the regulator, typically respecting uniform intervals, in which the regulator establishes a demand for a certain amount of renewable energy (usually classified by technology) and the bidders offer energy volumes and prices up to the quantity demanded. The regulator subsequently guarantees the price reached in the auction for the energy to be generated by the winner, usually by signing a long-term contract, and provided that the renewable power facilities are installed within the specified time period. The auctions could be restricted to plants of a certain size or technology.

[18] As described in previous chapters, retailing and distribution are bundled in the vast majority of US states and open retail access is available only in a few states.

This instrument may be viewed as halfway between a quantity and a price mechanism, since it introduces a competitive element in the allocation of economic incentives, while technological improvements can be factored into the equation automatically, and it provides financial security to the investor. It would therefore appear to offer the best of both worlds. Unfortunately, it also poses problems. The major drawbacks for this instrument are its complexity and its transaction costs, which hamper participation by small producers. This may also lead to market power issues (which may be addressed in the design of the auction).

Past experiences, when no renewable technologies were still mature, have revealed design flaws, leading to a general perception that auctions should not be used for promoting renewables. The major weaknesses identified have been the lack of credible non-compliance penalties when the winners of the auction fail to complete the project and the disconnection between the auction and land use planning. That notwithstanding, auctions have been successfully used for renewables procurement in several Latin American countries, where the initial flaws have been corrected by good design and very competitive prices have been revealed, see Batlle and Barroso [4]. Auctions appear to be the right heirs to successful FIT programmes for mature enough renewable technologies [14, 36].

The impact of uncertainty in the comparison among instruments

A final point in this comparison among support instruments for renewables concerns how uncertainty may affect the choice of instrument. This is a particular case of the previous discussion in Sect. 11.3.2.3. When the supply curve for renewable electricity (e.g. total production cost as a function of quantity) is flatter—as it is the case of wind—, uncertainty in the precise value of the cost of supply to be used in a price mechanism may result in significant errors in the quantity produced. Conversely, in a case where the supply curve is steep (e.g., under a tradable green certificate scheme with no banding, when all renewable technologies and consequently significant cost differences are combined), the use of a quantity instrument will result in major deviations in the price of the certificate (and, therefore, in the cost of the support) for different values of the requested quantity.

Another element to consider in the comparison of price and quantity instruments is the dynamic effect of the decline in technology costs. Generally, the cost of the technologies supported by these policies declines significantly along what is usually termed a learning curve: the more capacity that is manufactured and installed, the more is learnt about the technology, and the lower is its cost (not necessarily related to economies of scale). The shape of the learning curve also affects the choice of instrument. Steeper learning curves raise the preference for quantity instruments since, unless price mechanisms adapt very quickly (which is not usually the case), large sums may be awarded to producers instead of to consumers, who would not benefit from cost reductions under the price instrument approach, while they would under the quantity mechanism.

11.4.3 Voluntary Instruments: Green Electricity

This is the version of the voluntary agreements presented in Sect. 11.3.3.3 for renewable energy. Providers offer to sell "green" electricity (a variety of definitions exist for this term, but it is generally associated with small-scale renewable energy and cogeneration) at a certain cost premium (required because presently this energy is not competitive at market prices). Consumers are free to buy such power or not.

In this system, green labels or certificates (not to be mistaken with the tradable green certificate system) must first be established; these are generally issued by an independent body, which certifies that the electricity being consumed has indeed been produced at a renewable facility (virtually speaking, since it is not possible to identify the physical source of origin of the electric energy that is consumed). In its Directive 2001/77/EC for the furtherance of renewable electricity, the European Commission, for example, requires Member States to set up a system to guarantee the origin of electricity produced. Some private initiatives have also emerged, such as the RECS system, an initiative of certain European utilities.

In theory, this seems to be a very effective instrument because of the advantages enumerated below.

- It is fully adapted to the liberalised market, since it relies on consumer decisions.
- Its administrative costs are minimal.

However, problems associated with its implementation cannot be ignored.

- This is a typical case example of the presence of additional market failures: asymmetric information of buyer and seller and the fact that the outcome of this voluntary instrument is a public good (see Sect. 2.7). In consequence the volume of green electricity that is traded with this scheme is usually very small.
- It is difficult to combine with other support mechanisms, both in theory and in practice.
- It does not necessarily entail an increase in renewable energy: utilities may simply allocate all their "green" production to green electricity buyers, and leave the rest for "non-green" consumers. The premium paid by the consumers of this energy is only justified if it results in some *additional* green activity (e.g. installation of some renewable generators that are not profitable under the current regulation) by the selling utility, an action that otherwise would not have taken place.

For green electricity schemes to be effective, then, they must be adequately designed and probably need to be combined with other instruments [41].

11.4.4 Challenges for Large-Scale Renewable Energy Development

When renewable energy penetration reaches a significant level, promotion policies must become more refined to deal with other issues not previously addressed. In this section such issues are discussed only briefly. Interested readers may want to refer to Labandeira and Linares [34] and Pérez-Arriaga and Batlle [40].

- Since the cost of public support may grow into a significant sum, the source of funding merits attention. Renewables generation is typically supported by surcharges on the electricity tariff, although this is not always the case (e.g. production tax credits in the US come from the federal budget). There are solid arguments in favour of sharing this cost with other sectors or the national budget when the renewables target is only a part of a broader strategy also involving other industrial sectors, as for instance in a national climate change programme, see Batlle [7].
- Financing the progress along the learning curve, both of learning-by-doing and learning-by-research, is another major issue, because of the knowledge spillover effect. The issues are: how to share the benefits of the progress made and how to finance it.
- The most adequate support mechanism for any given technology should depend on its maturity. Probably the path to follow is to gradually move from price-based mechanisms (FIT and premiums), which are suited to less mature technologies, to auctions that are backed by long-term contracts, which have good properties with advanced players in more developed renewable markets. An issue that deserves serious investigation is the regulatory treatment of renewables once they reach grid parity, i.e. when they become competitive without subsidies in the electricity market.
- Most promising renewable technologies require substantial amounts of land. Targets for renewable generation and how to meet them need to be coordinated between the regional and local administrations (who are in charge of land use planning and who can try to extract rents from technology policies).
- Grid integration is another major issue in the presence of a significant level of renewables penetration and one that could be decisive in the success of failure of many wind and solar developments currently envisioned. The issues to be solved are those presented in Chap. 6. At transmission level: Who is in charge and how network planning is done; how to allocate the costs of network development; suitable business models for the transmission investors; how to address the problem of siting the network facilities. At distribution level: How the design, operation and losses of the distribution network, and the corresponding remuneration, change because of the presence of distributed renewable generation, of intermittent nature (wind and PV solar) for the most part; how to stimulate innovation in the distribution utilities to find the most efficient and reliable approaches to deal with this new situation.

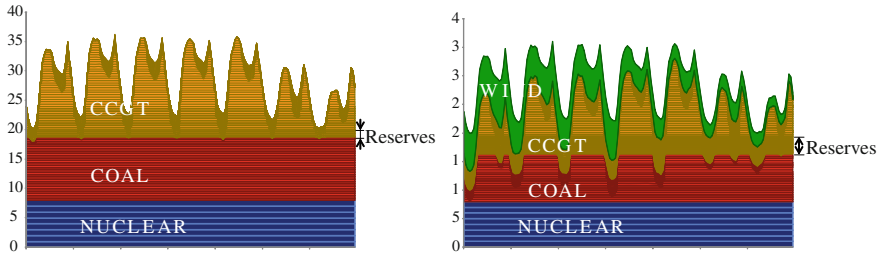


Fig. 11.7 Impact of wind generation on the operation of a power system

- Any ambitious renewable targets cannot ignore their impact on other aspects of the overall energy policy: objectives of other low carbon technologies, energy efficiency goals and reduction of GHG emissions. Although each one of these measures may have specific objectives, their mutual interaction, in particular to achieve carbon reduction targets efficiently, has to be jointly considered.
- A strong presence of renewable intermittent generation –such as is already the case in several countries– results in more frequent cycling (i.e. changes in the operation mode of the flexible generation plants, with shut-downs and start-ups caused by changes in wind or solar production) and losses of thermal efficiency, negative market prices or wind or solar curtailment. Strong intermittency penetration will reveal existing flaws in markets design and limitations of current analysis models that have been hidden until now because of their small quantitative relevance. For instance, since start-ups of many plants will become commonplace (see Fig. 11.7), the very nature of start-up costs and the way in which start-up costs are reflected in market will acquire much relevance. Increased penetration of wind and solar amplifies the differences in market prices resulting from different market rules. This will be another significant factor to consider when estimating the impact of renewable penetration on market prices.
- Market prices are naturally anticorrelated with wind or solar production (i.e. the more wind or solar production at zero variable cost there is at a given time, the lower will be the market price), therefore the per unit remuneration of wind and solar decreases with penetration.
- Renewable generation must contribute, within its possibilities, to the reliability and efficiency of system operation. Although wind and solar PV may not benefit much from being exposed to hourly day-ahead market prices, they should experience the cost of imbalances in the shorter term (i.e. after the day-ahead market) so that they may contribute to minimise the cost of reserves for the entire system by improving their estimation of their output and participating in the provision of reserves.

There is overwhelming evidence that system operation practices and power sector regulation will have to adopt innovative approaches to cope with the anticipated large presence of renewable generation. The main message is,

therefore, that system operators and regulators should act quickly to avoid that reality runs over current operation and regulatory practices, leading to inefficient and less reliable outcomes. System operation and electricity regulation must pave the way for a future power system where wind, solar, biomass and other renewables will play a major role.

11.5 Consequences of Environmental Regulation for Electricity Markets and the Economy

As with any other regulation, environmental or technology policies may have a significant impact on electricity markets or the rest of the economy. The effect sought, of course, should be the improvement of environmental performance or the increased penetration of a certain technology. These policies may give rise to other unexpected effects, however, which need to be taken into account in their design or evaluation.

The general consequences of the internalisation of environmental externalities are summarised below.

- Energy costs usually increase because lower impact but more expensive technologies are favoured.
- Windfall profits (or losses) may be created.
- Technology policies may distort market signals.
- Equity problems may arise, since the resulting measures or the increase in costs may not affect all consumers equally.
- Regional or industrial distortions may appear if the obligation to internalise externalities is not applied across all productive industries or geographic regions.

These topics are discussed in the following pages.

11.5.1 Impact on Prices

The internalisation of negative externalities results in higher prices for the good causing the externality (energy, in general, in this case). Assuming that the demand remains the same, when a negative externality is internalised, the equilibrium price of supply and demand becomes higher. This is illustrated in the simplified Fig. 11.8 (no differentiated time ranges, no differentiation of fixed and variable costs, etc.) where the initial supply curve S_o shifts to curve S_b (as a result of the higher cost of producing the good). Since it is assumed that the demand curve remains the same, the new price is higher (P_b), and the amount produced lower (Q_b). A similar analysis can be performed for positive externalities.

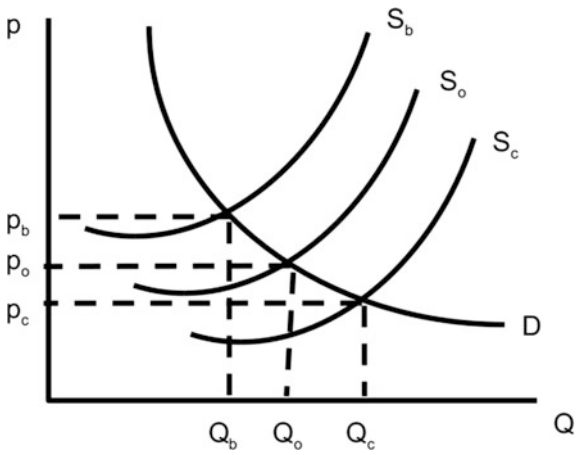


Fig. 11.8 Impact of internalising externalities on prices and quantities

This is not a problem, of course, since this is precisely the economic signal we were after: by increasing the price of electricity, resources are better allocated. Indeed, as previous sections have shown, instruments that do not fully incorporate the price signal (such as standards) are not wholly effective and efficient in internalising externalities. Increases in electricity prices are never popular, however. Indeed, as mentioned earlier, policymakers prefer to "disguise" them as far as possible by resorting to standards or similar instruments that fail to emit the full price signal.

In the next subsection we shall examine in more detail the diverse aspects that have to be considered when evaluating the impact of technology measures that support penetration of intermittent generation—we shall focus on wind—on the prices paid by consumers in an electricity market.

11.5.1.1 Impact of Technology Policies on Consumer Prices

The impact of technology policies on the prices that consumers must pay in a competitive electricity market will be analysed here. Since technology policies support the utilisation of more expensive technologies over less expensive ones, it is obvious that these measures always *increase the cost of electricity production,* if the cost of the support measure is accounted for. However, it is not clear whether they also cause an increment in the *price paid by consumers* for this electricity or not. This is a complex matter that requires careful analysis.

Limiting the discussion to renewables, in the operation time frame, wind and solar are generation technologies characterized by a variable cost of production that is basically zero. Therefore, at least in a first approximation, the expected global impact on the power system should be a reduction in the market prices—which basically depend on the variable cost of the marginal generation plant at a

given time—since other more expensive generation technologies have been displaced by the wind or solar production.

However, it remains the complex task of evaluating the several side effects. In mostly thermal systems (in which storage capabilities, as for instance hydro resources, are scarce), the presence of intermittent renewable generation implies a very significant change in the scheduling regime of the rest of the generating facilities. From the operation cost and market prices perspective, the presence of wind and solar also results in increments of the extra costs derived from more frequent cycling in the operation of the thermal units, which have also an effect on the marginal prices. Both impacts are separately examined next. Note that the discussion that follows is restricted to the short term impact of renewables' penetration. In the longer run the generation mix will evolve to adapt to the new situation of more volatile prices and a more "agitated" operation regime. Analysis of the long-term implications is a current topic of research.

The "merit order" effect

Technology policies generally act by subsidising specific technologies or by setting a certain quota for them. In whichever case, they reduce the need for conventional technologies (the effect is equivalent to a reduction in demand), and by doing so, they lower the electricity production costs.

However, while this effect is undoubtedly significant from a cost perspective, in a wholesale generation market the price reduction may be less important. This happens when the addition of wind or solar does not change the technology that sets the marginal price in most of the hours of the year. This is often the case in Europe and the US, with systems with a large component of combined cycle gas turbines (CCGT). To date, the newly installed CCGT plants worldwide are quite standard, with similar heat rates and also fuel costs within each power system.

The increasing cycling needs

The second impact is due to the lack of correlation of wind production with demand. Wind output alters the shape of the net load to be satisfied with conventional thermal generation, therefore changing the traditional way to schedule the thermal portfolio. Wind production may result in such a low value of net demand (mostly at night) that will force a large number of thermal units to shut down only to have to start up a few hours later.

Figure 11.9 illustrates this effect for a few days in 2010 for the Spanish power system. Despite the considerable amount of hydro flexibility available in this system, the large amount of wind generation in 2010 leads to significant changes in the economic dispatch that did not happen 5 years before. The thermal load can be significantly lower on a Tuesday than on a Sunday, and many CCGT plants (typically 400 MW in size) have to be started up daily when wind production is high at night (for instance, the CCGT output on Wednesday at 4:00 AM was limited to 2 GW while at the peak at 19:00 went over 11 and 0.9 GW on Sunday at 5:00 while more than 6 GW at 21:00).

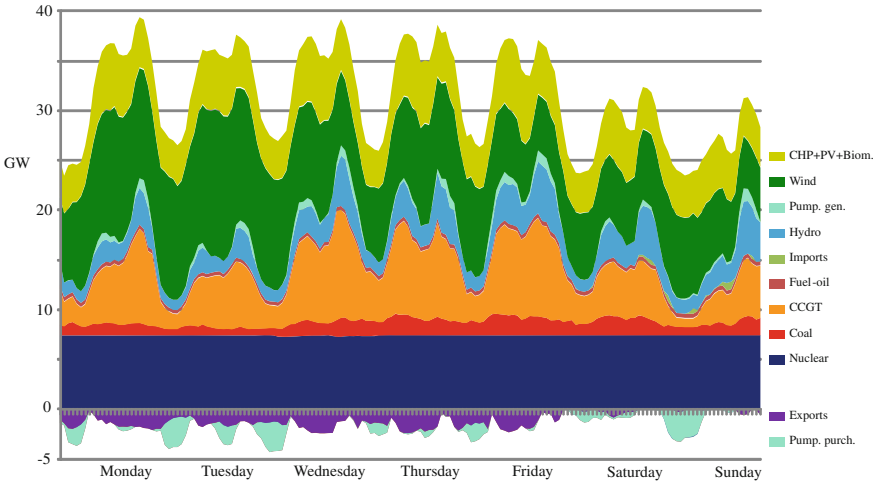


Fig. 11.9 Electricity supply in the Spanish system from Nov. 8th to 14th, 2010

The impacts of fossil power plant cycling operation are multiple: significant increase in equivalent forced outage rate, additional capital and maintenance expenditures and increase fatigue-related and creep-related wear and tear. These translates into a significant cost increase caused by operation, maintenance, and capital spending, replacement energy and capacity cost due to increased outage rate, cost of heat rate change due to low load and variable load operation, cost of start-up auxiliary power, fuel, and chemicals, cost of unit life shortening and general engineering and management cost.

Since the increasing penetration of wind and solar is unavoidably going to lead to a significant increase of these cycling-related costs, any sound economic analysis needs to properly take these expenses into consideration, particularly due to the fact that the actual and expected costs of cycling fossil units that were originally designed for base-load operation is greater than most utilities had estimated.

In this new context, a key factor that minimises the cost increase of the economic dispatch is the flexibility capability of the thermal plants and storage. The cost impact of a large introduction of wind and solar will therefore be inversely proportional to the amount of existing flexible generation and storage.

Large wind penetration amplifies the differences in the diverse pricing rules of the numerous electricity markets. As illustrated also by Batlle and Rodilla [5], the effect of the nonlinear characteristics of power plant operation (costs of starts, ramping limits, technical minima etc.) in the computation of the electricity market prices differs significantly in those markets in which a nonlinear pricing approach plus discriminatory side payments is implemented (in which marginal prices do not include the effect of non-convex costs, e.g. PJM or ISONE) from those in which a convex-hull or linear pricing approach is implemented (an hourly uplift perceived by all producers is added to the marginal price, e.g. Ireland) or just

simple bids are accepted (most European markets). In these latter cases, the market operator when calculating the market price or generators when designing simple bids, must internalise the nonlinear costs in shorter functioning periods, resulting in higher bids and, consequently, higher marginal prices for consumers.

The cost of the technology policies

Finally, the cost of any specific technology measure (e.g. a feed-in tariff) that has been applied to support the penetration of renewables must be considered when computing the price of electricity for the consumer. In most power systems this cost is passed through to the end consumers as a regulated charge that is added to the access charge that all consumers have to pay.

However, there are good reasons to believe that a fraction of this cost should be shared with other sectors that are also concerned in the global strategic energy and climate change policies, see Batlle [3] for a detailed discussion.

All these issues must be taken into consideration when evaluating the impact of policy measures that support intermittent renewable generation on electricity prices for end consumers.

11.5.2 Windfall Profits

Windfall profits arise when a new regulation creates rents that can be appropriated by existing producers. This is the case when, for instance, the electricity price in a power sector increases due to the introduction of an environmental tax (whether directly or through an emissions trading system). Indeed, the marginal price rises (because the marginal generating unit is more expensive to operate, for it needs to buy emission permits, pay emission taxes, or be replaced by a more expensive technology), and since this marginal price is paid to all producers, the total income received also rises. Regulation raises producers' costs as well, but not equally: some need not pay at all (because they do not emit pollutants, for example), while for others the rise in costs is smaller than the rise in revenues.

This effect occurs in liberalised markets, where the remuneration of generation is based on marginal prices. In markets subject to cost-of-service regulation, the tax creates an extra cost for the generators that must be paid by the consumers, therefore increasing the average cost of electricity for the consumers, but not resulting in windfall profits for the generators.[19]

Since the extra income is drawn from consumers, this might be considered wealth transfer rather than loss of efficiency. Indeed, the extra rents received by

[19] Note that in power systems under traditional cost-of-service regulation is possible to design and apply tariffs that are based on marginal costs. In this case the consumers would pay the same as under competitive market conditions. However, the generators would only receive their cost of service and there will be a surplus or a deficit, to be administered by the regulator as it sees convenient.

producers are a long-term signal for them to change production to cleaner technologies. Therefore, windfall profits do not necessarily have to be eliminated, except possibly in the two cases described below.

- When power plants were built under a regulated system and competition transition charges were not properly set considering this situation, these windfall profits should not necessarily accrue to producers.
- When windfall profits are received by "spent" technologies, i.e. technologies that cannot be built in the future (e.g. when there is a ban on nuclear power, or when there are no more hydro locations), the long-term signal is useless and, therefore, lowers efficiency.

In these cases, the regulator may need to intervene to reduce the amount of these profits (e.g., by setting a windfall tax).

11.5.3 Distributional Effects

The effects of environmental or technology policies are not felt equally by all consumers and producers. The impacts on producers are discussed above. Similarly, not all consumers are affected to the same extent by changes in electricity prices. Low-income consumers typically spend more of their budget on electricity (and other basic goods and services); therefore an increase in these prices has a greater impact on them. This is why environmental policies are generally considered to be regressive.

However, as mentioned earlier, this does not necessarily have to be the case. If the environmental policy creates rents that can be appropriated by the regulator, it affords an opportunity to redesign financial policies and reduce fiscal burdens for low-income consumers, thereby correcting the imbalance created. This is one of the premises put forward in support of emissions trading systems: by clearly separating efficiency and equity concerns, these instruments make the redistribution of income to offset the regressive effects of higher prices very transparent.

A second consideration relies on the elasticity of demand: electricity demand is usually fairly inelastic, so the burden of the environmental policy will fall mostly on consumers. But what if demand were made more elastic by implementing demand-response programmes, for example? Part of the burden could then be transferred to producers, and consumers might even be better off than before the policy was implemented.

A further discussion of the distributional effects of environmental policies can be found in Fullerton [19]. The pertinence of sharing the cost of renewable support measures between the power sector and other sectors is another relevant distributional topic.

11.5.4 Regional or Sectorial Distortions

Lastly, environmental policies may create regional or sectorial distortions. A good example is the European ETS scheme, which has created an emissions trading system exclusive to Europe for the industrial sector only. This creates an unfair advantage for non-European products (whose manufacturers do not have to pay for their emissions) and for Europe's non-industrial sectors (whose development is favoured over industry), fuelling the contention that environmental policies should be as comprehensive as possible.

Indeed, these sectorial or regional distortions may even be detrimental for environmental policy itself, in the event of leakage. By leakage is meant the rise in emissions outside the regulated area or sector (e.g., when non-European countries increase their production after becoming more competitive without a carbon price) that exceeds the emissions reduction in the regulated area (so that total emissions may even rise). The prevention of such developments depends on the ease with which production may be relocated outside the regulated area, the difference in emissions rates between the regulated and non-regulated areas, and other trade-related factors. Leakage detracts from the efficiency of environmental policy, since while policy costs do not budge, emissions decline at a slower pace. In this case, regulators may try to reduce leakage by implementing supplementary policies (border taxes, for example). Leakage in electricity markets is analysed in detail in Bushnell and Chen [9] and Quirion [43].

References

1. Aldy JE, Pizer WA (2009) Issues in designing US climate change policy. Energy J 30:179–210
2. Auer H, Resch G, Haas R, Held A, Ragwitz M (2009) Regulatory instruments to deliver the full potential of renewable energy sources efficiently. Eur Rev Energy Markets 3(2):125–158
3. Batlle C (2011) A method for allocating renewable energy source subsidies among final energy consumers. Energy Policy 35(5):2586–2595. doi:10.1016/j.enpol.2011.02.027
4. Batlle C, Barroso LA (2011) Support schemes for renewable energy sources in South America. MIT-CEEPR Working Paper 11-001
5. Batlle C, Rodilla P (2011) Generation technology mix, supply costs and prices in electricity markets with strong presence of renewable intermittent generation. IIT Working Paper IIT-11-020A. www.iit.upcomillas.es/batlle/publications
6. Batlle C, Pérez-Arriaga IJ, Zambrano-Barragán P (2012) Regulatory design for RES-E support mechanisms: learning curves, market structure and burden –sharing. Energy Policy 41:212–220
7. Batlle C (2012) A method for allocating renewable energy source subsidies among final energy consumers. Energy Policy (Forthcoming)
8. Bovenberg AL, Goulder LH (1996) Optimal environmental taxation in the presence of other taxes: general- equilibrium analyses. Am Econ Rev 86:985–1000
9. Bushnell JB, Chen Y (2009) Regulation, allocation, and leakage in cap-and-trade markets for CO. NBER Working Paper No. 15495

10. Canton J, Linden AJ (2010) Support schemes for renewable electricity in the EU. Economic Papers 408, April 2010
11. Commission of the European Communities (2008) The support of electricity from renewable energy sources. Commission staff working document. SEC(2008) 57. Brussels, 23.1.2008
12. Crocker T (1966) The structuring of atmospheric pollution control systems. In: Wolozin H (ed) The economics of air pollution, W.W. Norton, New York, pp 61–86
13. Dales JH (1968) Pollution, property and prices. University of Toronto Press, Toronto
14. Del Río P, Linares P (2012) Back to the future? Rethinking auctions for renewable electricity support. IIT Working Paper 12-038A
15. European Commission (2005) ExternE—Externalities of Energy. Methodology 2005 Update. European Commission, Luxembourg
16. Ellerman AD, Joskow PL (2008) The European Union's emission trading system in perspective. Pew Center on Global Climate Change
17. Eskeland GS, Devarajan S (1996) Taxing bads by taxing goods. Pollution control with presumptive charges. World Bank
18. Fischer C (2008) Emissions pricing, spillovers, and public investment in environmentally friendly technologies. Energy Econ 30:487–502
19. Fullerton D (2010) Six distributional effects of environmental policy. CESifo Working Paper No. 3299
20. Fullerton D, Metcalf G (2001) Environmental controls, scarcity rents, and pre-existing distortions. J Pub Econ 80:249–267
21. Goulder LH, Parry IWH (2008) Instrument choice in environmental policy. Rev Environ Econ Policy 2:152–174
22. Goulder L (2000) Confronting the adverse industry impacts of CO2 abatement policies: what does it cost? Climate Issues Brief 23. Resources for the Future, Washington
23. Hanemann M (2009) The role of emissions trading in domestic climate policy. Energy J 30 (Special Issue 2). Climate Change Policies After 2012
24. IEA (2011) Deploying renewables: best and future policy practice. Report from the International Energy Agency
25. IPCC (2007) Climate change 2007: synthesis report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, Pachauri, R.K and Reisinger, A. (eds.)], IPCC, Geneva, Switzerland, 2007
26. Joskow PL (1992) Weighing environmental externalities. Let's do it right. Electricity J 5:53–67
27. Kahneman D, Knetsch JL, Thaler RH (1991) Anomalies: the endowment effect, loss aversion, and status quo bias. J Econ Perspect 5:193–206
28. Keohane N, Revesz R, Stavins R (1998) Choice of regulatory instruments in environmental policy. Harv Environ Law Rev 22:313–367
29. Klein A, Pfluger B, Held A, Ragwitz M, Resch G, Faber T (2008) Evaluation of different feed-in tariff design options—best practice paper for the International Feed-In Cooperation. EEG and Fraunhofer Institute Systems and Innovation Research
30. Kolstad CD (2010) Environmental Economics, 2nd edn. Oxford University Press, New York
31. Labandeira X, Linares P (2011) Second-best instruments for energy and climate policy. In: Markandya A, Galarraga I, González M (eds) Handbook of sustainable use of energy, Edward Elgar. (to be published)
34. Labandeira X, Linares P (2012) A natural experiment of large-scale renewable energy policy: the case of Spain. Economics for Energy Working Paper 10/2012
32. Linares P, Labandeira X (2010) Energy efficiency: economics and policy. J Econ Surv 24:573–592
33. Linares P, Santos FJ, Ventosa M (2008) Coordination of carbon reduction and renewable energy support policies. Climate Policy 8:377–394
35. Markandya A, Bigano A, Porchia R (2010) The social cost of electricity: scenarios and policy implications. Edward Elgar Publishing, Cheltenham

36. Maurer LTA, Barroso LA (2011) Electricity auctions: an overview of efficient practices. A World Bank Study
37. Menanteau P, Finon D, Lamy M-L (2003) Prices versus quantities: choosing policies for promoting the development of renewable energy. Energy Policy 31:799–812
38. Newell RG, Pizer WA (2008) Indexed regulation. J Environ Econ Manage 56:221–233
39. Pérez-Arriaga IJ (2009) Regulatory instruments for deployment of clean energy technologies. Working Paper 09-009. Center for Energy and Environmental Policy Research (CEEPR), MIT. July 2009
40. Pérez-Arriaga IJ, Batlle C (2012) Impacts of intermittent renewables on electricity generation system operation. Econ Energy Environ Policy. Forthcoming
41. Pérez-Plaza M, Linares P (2008) Strategic decisions for green electricity marketing in Spain: learning from past experiences. In: Wang H-F (ed) Web-based green products life cycle management systems: reverse supply chain utilization. IGI Publishing, Hershey, pp 250–266
42. Pigou AC (1932) The economics of welfare. MacMillan, London
43. Quirion P (2010) Climate change policies, competitiveness and leakage, by Philippe Quirion. In: Cerdá E, Labandeira X (eds) Climate change policies: global challenges and future prospects. Edward Elgar, Cheltenham, p 2010
44. Roberts MJ, Spence M (1976) Effluent charges and licenses under uncertainty. J Pub Econ 5:193–208
45. Stavins RN (2001) Experience with market-based environmental policy instruments. In: Maler K-G, Vincent J (eds) The handbook of environmental economics. Elsevier Science, Amsterdam
46. Weitzman ML (1974) Prices vs. quantities. Rev Econ Stud 41:477–491

Chapter 12
Security of Generation Supply in Electricity Markets

Pablo Rodilla and Carlos Batlle

> *I had to abandon free market principles in order to save the free market system.*
>
> George W. Bush

As indicated in Chap. 3, calling "liberalization" and, particularly, "deregulation" to the regulatory reforms that have taken place worldwide during the 1990s and early 2000s is somewhat misleading. The main reasons are that only some of the activities involved in the supply of electricity have been subject to an in-depth reform, most governments still have a heavy hand on their power sectors and the volume and complexity of the new regulation is frequently similar, or even greater, than the so-called traditional regulatory framework, see [13] or [44]. This is why the term "restructuring" has been preferred in some electric power systems (particularly in the United States).

The subject matter of this chapter, i.e. the need for regulatory intervention to complement electricity markets in order to guarantee security of generation supply, is a good illustration of this point: a flurry of regulatory activity has been developed all over the world to make sure that competitive wholesale electricity markets provide a satisfactory level of security of supply.

As explained in Chap. 3, the objective of regulation is *to prevent (or conversely produce) inefficient (efficient) outcomes in different places and timescales that might (might not) otherwise occur*. In this chapter we show how, in restructured electricity markets, the intervention of the regulator is needed *to guarantee a*

This chapter draws heavily on [4], [5] and [42].

P. Rodilla (✉)
Instituto de Investigación Tecnológica (IIT), Universidad Pontificia Comillas,
Alberto Aguilera 23, 28015 Madrid, Spain
e-mail: Pablo.Rodilla@iit.upcomillas.es

P. Rodilla · C. Batlle
Institute for Research in Technology, Comillas Pontifical University, Madrid, Spain

C. Batlle
MIT Energy Initiative, MIT, Cambridge, MA, USA

C. Batlle
Florence School of Regulation, European University Institute, Florence, Italy

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_12,

minimum required level of security of supply in different places and timescales, since it has been largely demonstrated that *otherwise this level will not be reached*. In practice, this intervention amounts to an additional set of rules in the wholesale electricity market. As it is the case of many other regulatory topics, there is still no consensus on the nature of these rules, although much progress has been made and some implementations can be judged to be successful.

12.1 Introduction

12.1.1 The Security-of-Supply Problem

Modern society depends critically on the availability of electricity. The consequences of a lack of supply are known to affect regions and countries profoundly in their social, economic and political dimensions. Progress is undoubtedly linked to the availability of sufficient electricity; unsurprisingly, then, the avoidance of emergency situations and ensuring certain quality standards are among regulators' major concerns.

The actual physical supply of electricity to end consumers at any given time is the result of a complex and interlinked set of actions, some of which were performed many years in advance. As a result of these actions, the right technologies and infrastructures are developed and installed, contracts for the supply of fuel are concluded, hydro reservoirs are suitably managed, power plants are maintained in a timely and satisfactory manner, generators connect to the grid and start-up for operation when needed and operating reserve margins are established.

While this chapter focuses exclusively on the measures that must be implemented for the generation activity, other activities not addressed here are also relevant to the security-of-supply problem, like the adequate planning and secure operation of the distribution and transmission networks.

As with any other regulatory issue, breaking the security-of-supply question down into its major components facilitates understanding, discussion and the design of proper technical procedures and regulatory measures (if deemed necessary). Several classifications of these components are in use. Since, in this particular topic, the components of the problem usually make reference to the time periods involved, these components will be hereafter referred to as *time dimensions* or simply *dimensions*.

One of the most popular classifications of these dimensions, developed by [37], takes just two aspects into consideration: security (a short-term operation issue) and adequacy (a long-term capacity expansion issue). The dimensions listed below were first outlined in [7], and provide a more detailed breakdown of the involved time frames.

The four dimensions of the security of electricity supply issue

Next, we simply introduce the four dimensions of security of electricity supply in broad terms. Later, in Sect. 12.2.2, they are explained in more detail, with exclusive focus on generation.

- Security, a very short-term issue (in the realm of system operation, close to real time), is defined by the NERC as the "ability of the electrical system to support unexpected disturbances such as electrical short circuits or unexpected loss of components of the system or sudden disconnection" [37].
- Firmness, a short- to mid-term issue, can be defined as the ability of facilities already installed to respond to actual requirements to meet the existing demand efficiently.[1] This dimension is linked to both the technical characteristics of the generation and network facilities and their management in the medium-term (e.g. scheduling of maintenance of lines and generators or control of hydro reservoirs).
- Adequacy, a long-term issue, refers to the existence of enough available generation and network capacity, either installed or expected to be installed, to efficiently meet demand in the long term.
- Strategic expansion policy concerns the very long-term availability of energy resources and infrastructures. This dimension usually entails diversifying fuel supply and the generation technology mix, together with long-term network planning, and it is frequently associated to other aspects of energy policy, in particular environmental concerns.

These four dimensions are interrelated and cannot be decoupled from one other. For instance, when the regulator exclusively defines requirements for the shortest term dimension (e.g. the need for a certain amount of operating reserves to maintain security), longer term dimensions (such as which the new generation investments will be in the future) may be affected by this short-term requirement. This dependency works both ways; thus, when analysing the new investments needed to efficiently meet demand in years to come, a study must also be conducted of whether they will be able to meet future requirements for the shorter term (e.g. the operating reserve targets defined for the security dimension or the flexibility to mix well with large levels of penetration of intermittent renewable generation).

In the market-based context, market forces are ideally supposed to result in decisions that are equal to those that would be taken under an ideal (e.g. assuming perfect information) centralised operation and planning context. This applies to all

[1] Although the concern here is security, not economy, we shall add the qualification of efficiency systematically when referring to how the demand should be met. The reason is that price responsive demand is always met, when prices are high enough. But this may not be efficient under a maximisation of the social welfare viewpoint. We want to achieve levels of security of supply that are optimal under the joint point of view of supply and demand.

time ranges: short-term system balancing, medium-term operation management of existing facilities, when and where to build new facilities, or which generation technologies should be deployed under a long-term strategic perspective.

Security of supply is a very important concern in all time scales of decision-making in power systems. However, one cannot forget that it is just one among the three major objectives—the other two being the economic and environmental considerations—that comprise the maximisation of overall net social benefit. The energy policy of a particular country, or an aggregation of countries like the European Union, must take these three objectives jointly into consideration and issue policy guidelines that steer and establish limits, targets or other requirements to the decisions made concerning security, efficiency or environmental impact at lower levels. Examples of these energy policy guidelines are decarbonisation targets for the power sector as a whole, minimum production requirements with renewable resources, minimum interconnection capacities of the transmission network between countries or power systems, quality of supply targets or minimum requirements of diversification resources for electricity supply.

12.2 Security of Supply in Generation: Problem Diagnosis

12.2.1 Introduction

Ever since Chile restructured its power sector in 1982, the ability of electricity markets to provide enough generation to ensure security of supply has been called into question. Mistrust of whether the market, left to its own devices, can provide sufficient (and also efficient) generation when needed has gradually but inexorably led to the implementation of additional regulatory mechanisms. This issue has undoubtedly been gaining importance with the passing of time, and it is now a key item on energy regulators' agendas.

International experience has shown that regulatory intervention is necessary in the security and strategic expansion policy dimensions. In between these two dimensions, however, the debate on the need to intervene to ensure security of supply remains open and quite intense.

Therefore, in this section, after a brief description of the potential inefficiencies that may arise in each dimension of the problem (security, firmness, adequacy and strategic expansion policy), we analyse firmness and adequacy in greater detail. We show that the market, which ideally should be able to guarantee satisfactory resource management and investment incentives, fails in this regard because of a number of reasons. The premises under which the market would provide the optimal solution are unfortunately absent in practice.

12.2.2 The Four Dimensions of the Security of Generation Supply Problem

Next we analyse in more detail the main issues concerned in each of the four dimensions of the problem.

Security

The real-time operation of a power system must be centrally coordinated to ensure a continuous match between supply and demand. The System Operator (SO) is generally acknowledged to be responsible for such coordination. As pointed out in [47], between real time and the longer term "there are dividing lines that describe the System Operator's diminishing role in forward markets. Where to draw those lines is the central controversy of power-market design". Each system has traditionally used its own criteria to define the point at which the SO takes control to ensure security. After that point, usually known as *gate closure*, the security dimension comes into play.[2]

At gate closure, the scheduled generation is transferred to the System Operator to guarantee the continuity (avoid supply interruptions), quality (voltage and frequency are kept within acceptable margins), and efficiency (provision of electric power at the lowest possible cost) of supply. In a market environment, the general approach consists of having the SO purchase (supposedly through a transparent and competitive process) so-called ancillary services. These products are often classified into three categories: frequency control (primary, secondary and tertiary operating reserves); reactive power for voltage regulation (which, again, may be classified as primary, secondary and tertiary); and black-start capabilities (restoration of power).

Ad hoc markets for the quantities of operating reserves prescribed by the SO[3] are regarded as a good hybrid (market and regulation) alternative for ensuring security.[4] Even where market mechanisms are implemented, the regulatory intervention is evident. The establishment of the quantity of operating reserve requirements affects short-term prices and, consequently, long-term investment signals.

The intervention sometimes goes even further, for instance, market agents are nearly always compelled to submit bids for all their (mandatory) balancing reserves

---

[2] Gate closure has been typically defined by the deadline for the reception of bids for the day-ahead market. The existence of intra-day markets can move gate closure closer to real time, depending on the specific market design (see Chap. 7).

[3] Ideally, the agents of an electricity wholesale market would be able to deliver some operating reserves by their own initiative. In practice, regulators and system operators are unlikely to rely on just the market for security provision.

[4] A clear distinction must be drawn between these operating reserves and the reserves intended to produce electricity during times when demand threatens to be larger than available production capacity, i.e. the emergency reserves that are called upon to supply energy when generation capacity is scarce. While operating reserves are meant to provide security under normal operating conditions, 'emergency' reserves are closely related to the firmness and adequacy dimensions. Purchasing reserves to jointly cover both needs is clearly inefficient, since emergency reserves are more expensive than the normal operating reserves.

to both increase and decrease energy capacity for all the next day's time intervals. Moreover, in many power systems (in most European systems, for instance), the imbalance settlement is currently resolved through dual imbalance pricing, where a different price is applied to positive and negative imbalance volumes (with respect to the system's resulting net imbalance) for each hour. This dual imbalance pricing itself is supposed to provide incentives for market agents to try to avoid deviating from their scheduled programmes, but it is achieved at the cost of artificially modifying the very short-term system's marginal price signal. This specific design is meant to improve system security, but it also has some adverse effects on market performance and overall efficiency (see, for instance, [6]).[5]

Firmness

Even with an ample margin of installed generation over peak demand, if a significant part of this capacity is not readily available when needed for whatever reason (for example, lack of water in the reservoirs or of fuel in the tanks, or out-of-service plants for maintenance or because of a forced outage), demand may not be efficiently met.

From the firmness standpoint, regulators should evaluate whether just economic market signals can ensure efficient management of the existing generation resources, or if it would be appropriate to introduce an additional mechanism to ensure this result.

Regulators are very risk averse to experiencing situations with a deficit of generating resources in critical periods. This is why additional regulatory mechanisms frequently exist, as described in later sections, to provide incentives to enhance the availability of generation plants in such periods. Such incentives may include measures geared to minimising the likelihood of outages, adequately planning fuel supplies and maintenance programmes, or managing reservoirs more prudently.

A quantitative measure of "firmness" is needed, if this characteristic of generation plants is to be regulated. According to the definition of this dimension of security of supply, the best measure is "the amount of capacity that is available to generate when needed, during the required interval of time". For a given generation plant, this amount may be different in different occasions, and it will depend on the existing economic incentives, among other factors. The "firm supply" value for a given power plant is often estimated using historical data on availability, production or—better—availability or production when capacity is scarce. It may also be found with a mathematical model (fed with historical or estimated future system data). The determination of the "firm supply" of each generation plant is a controversial issue, since it has economic implications in the remuneration for this service. A purely market-based approach would consist of letting each plant commit to a self-defined value of firm supply (in MW) and then penalise the plant heavily if this product is not delivered when requested by the SO, or when some

[5] For instance, dual imbalance pricing and the fact that imbalances are jointly evaluated on a portfolio basis create a barrier for new entrants.

threshold level of some indicator of emergency conditions in the system is exceeded. Obviously, this latter approach would call for the definition of adequate guarantees, so as to avoid the risk of default of the plant in case of reiterative non-compliances.

Adequacy

If considered necessary, the regulator may decide to supplement the economic signals from the wholesale electricity market with additional incentives to attract new efficient generating units, so that there will be adequate installed capacity with the suitable characteristics to meet the estimated future demand efficiently.

As discussed in a later item, the instrument used to reach that objective basically involves providing new entrants with the extra source of income and/or the hedging instruments they require to proceed with efficient investments. As later analysed in more detail, the definition of time-related terms (such as lag period[6] and the duration of the incentive) and the volatility associated with this additional income are key factors. In response, the investors are expected to provide enough firm generation capacity to meet the expected future demand efficiently.

The "efficiency" qualifier requires that the regulators establish a mechanism to differentiate "good" from "fair" or even "poor" investments, so that it is avoided that inefficient or "junk" generation plants are attracted by the adequacy incentives. As with firmness, here it is also needed some measure of "capacity credit" or "firm capacity" to make sure that any adequacy target is met.

Strategic expansion policy

Adequacy mechanisms often consist of introducing incentives to directly or indirectly encourage the entrance of new capacity with a view to obtaining a reserve margin higher than would be naturally provided by the market if left to its own devices. This leads, in principle, to market agents' choosing from among the various technologies available.

However, it may be wise to introduce some very long-term criteria, in addition to the adequacy mechanisms, to reflect the regulator's very long-term view with respect to the different technologies. These strategic expansion criteria should be very much connected to any existing energy policy goals, as described previously.

For instance, if the long-term energy policy dictates a heavy reliance on renewable energy resources for electricity production, the corresponding strategic expansion policy may establish targets or economic subsidies to further the development of certain renewable technologies, in light of expectations that they will ultimately become efficient alternatives. Wind electricity production is a good example: after years of investing in support mechanisms for wind generation, cost convergence with traditional alternatives now appears to be imminent.

Security of supply is an ingredient of the broader goal of sustainability of supply. Sustainability entails supplying electricity to present end users while ensuring that future users will be provided for, generation after generation. This is

[6] Also known as lead time.

a tall order, given that the present model of electricity supply, and the entire energy model for that matter, are not sustainable.[7]

12.2.3 Is the Market Capable of Ensuring Generation Adequacy and Firmness that Meet Security of Supply Standards?

Changes in the regulation of the electric power industry worldwide have drastically modified traditional security-of-supply issues and approaches. In vertically integrated utilities, security of supply of generation under cost-of-service regulation was a major ingredient in centralised utility planning. Under the market-oriented paradigm, the regulation must ensure that appropriate economic incentives are in place so that security of supply is maintained at socially optimal levels. The main conceptual reason behind liberalising the generation activity was to promote efficiency at operation and capacity expansion levels. The remainder of this section examines whether some sort of regulatory intervention is necessary to ensure *firmness* and *adequacy* (in other words, enough installed firm generation capacity and appropriate medium-term management of this capacity, to efficiently meet demand within any required security of supply standards) in competitive wholesale electricity markets.

The ability of electricity markets to provide the system with the required security of supply was called into question from the outset. Some authors [47, 39, 26, 29] contended that the market fails in a number of different contexts and for a variety of reasons. A brief summary of these reasons is provided in Sect. 12.2.5.

This is arguably one of the most important issues still outstanding under the market-oriented regulatory framework. Although no international consensus has been reached in this regard, market failure is increasingly being acknowledged, leading to the conclusion that, without regulatory intervention, the market is unable to provide sufficient generation availability when needed.[8]

12.2.4 Ideally, the Market Solves the Problem

According to the conclusions in Sect. 2.3.3 of this book, obtained under the assumption that the considered wholesale electricity market is perfectly

[7] Sustainable development is defined in WCED [53] as development that "meets the needs of the present without compromising the ability of future generations to meet their own needs". A sustainable energy model must include certain essential features: lasting and dependable access to primary energy sources and adequate infrastructures to reliably produce and deliver the required amount of energy, equitable access to energy supply and acceptable environmental impact.

[8] In almost all electricity markets, the regulator has implemented rules that are meant to reinforce or restrain natural market trends, in an attempt to guarantee supply in the short, medium and long terms.

competitive, the producers and consumers in a market should reach a situation of equilibrium with a number of interesting properties. Among them, we should highlight here those that are relevant to the firmness and adequacy of a perfectly competitive power system:

- When consumers and producers negotiate in the short term on a perfectly competitive market who sells, who buys and the transaction price, the result is economically efficient and it is reached when the marginal utility of the aggregation of consumers equals the marginal production cost of the aggregation of suppliers, i.e. the system short-term marginal production cost for that level of demand.
- Due to the multiple levels of demand that the installed generation capacity has to cover, a mix of different technologies happens to be the optimal investment choice to maximise the total social welfare.
- The capacity expansion and operation results for the ideal centralised management model and the competitive market model are identical, and achieve the same optimal social welfare, if and only if marginal prices are used to remunerate the energy output; roughly speaking, this marginal price is equal to the variable cost of the last generation unit in the merit order (the marginal unit of electricity production).
- The energy market price is all that is needed to remunerate the generators if the regulator is not seeking any specific objectives, i.e. those that in the centralised management model have to be introduced as regulatory constraints, such as those on security, reliability or environmental concerns.
- When the regulator wants the market to have a reliability level that corresponds to the one achieved by centralised planning under some reliability constraint, the market price must include an additional component. As reviewed in Chap. 2, this component has the format of a "capacity payment", per MW of installed capacity, to any generator g of whatever technology. The "capacity payment" seeks to compensate the deficit in the remuneration of the generation plants under purely market conditions when it is desired to have an excess of installed capacity over the strict economic optimum.

A number of key premises were necessary to arrive at these ideal results (see Sect. 7.4.1):

- The market is perfectly competitive and all agents have perfect information. This requires that the ideal conditions for introducing generation "reform", as discussed in Chap. 3, be met. The textbook requisites for establishing a competitive market for generation include privatisation, vertical unbundling and an adequate horizontal structure [28].

- The short-term spot price always reflects demand-side marginal utility. This requires accounting for short-term demand elasticity explicitly.[9]
- Risk neutrality (i.e. no risk aversion) of all the system agents.
- The production cost function of the power system is convex. The main reason for not compliance with this premise is the existence of nonlinearities in the cost function of the thermal generators, due to the existence of start-up costs.
- Neither economies of scale nor lumpy investments are present.

Actual markets fail to comply with most of these premises (all but the last one, in most cases), therefore requiring ad hoc adjustments to arrive at the desired outcome.

12.2.5 Market Imperfections and Flawed Regulatory Rules

The paragraphs below examine the impact of not achieving some of the aforementioned premises:

- Risk aversion and inefficient risk allocation.
- Flawed regulatory rules that distort market signals by preventing the existence of high electricity market prices when generation is scarce.
- Lumpy generation investments or significant economies of scale.

12.2.5.1 Agent Risk Aversion and the Consequences of Inefficient Risk Allocation

For a wide variety of reasons, risk aversion is a particularly prominent characteristic of generator behaviour in power markets that significantly affects their long-term investment and medium-term resource management decisions.[10] Risk is high in power generation investment decision-making, and failures[11] are likely. Risk also plays a key role in the resource management decision-making process, although to a lesser degree.

---

[9] Non-elastic demand is assumed in the discussion that follows. Since it is essential, however, for prices to reflect the so-called value of energy not served when scarcity prevails, a value determined by the regulator that sets the price in the event of such paucity is consequently assumed to exist.

[10] The degree and effect of generator risk aversion depend on system structure and characteristics.

[11] "Investment failures" refer to investments in which net social benefit is lower than it would be with other available options or, equivalently, when market revenues are insufficient to cover the total generation costs. The uncertainty involved in power sector investment decision-making is the main factor responsible for these suboptimal (when evaluated ex-post) investments.

New facilities require very large investments. They have long lead times and involve a great deal of technological, price and regulatory uncertainty during their typically long economic lifespan. These issues make investment especially risky and generators appear to be more risk averse than investors in other types of markets. The major consequence is that, in pursuit of protection against low price scenarios, generators tend to install less capacity than if they were risk-neutral.

In actual power systems, suppliers have to make important decisions (generally in the medium term) to ensure that their existing generation facilities will be productive in the future. Therefore, they sign contracts to cover their future fuel requirements,[12] and decide when it will be more profitable to produce power with the limited hydro energy resources available (under the uncertainty of future inflows or the risk of spillage) and when to perform plant maintenance. All these decisions affect the future availability of electricity, and therefore the system's future performance. In the aforementioned pursuit of protection against risk (of low prices, losses arising from water spillage and fuel overcontracting), however, generators are usually conservative. They generally prefer to deploy limited water resources when prices are moderately high, for instance, rather than to wait for a possible (uncertain) scarcity of generating resources (when the peak price would be very high) in the future.

Demand is also risk averse

Consumers are also risk averse and seek protection against high prices. They therefore prefer a system with greater installed capacity and greater resource availability than if they were risk neutral.

The ideal market-based solution

For all these reasons, and from an ideal, theoretical perspective, both generation and demand have sufficient incentive to hedge and allocate risk efficiently (by signing long-term contracts). For generators, volatile prices may hinder project financing and lead to suboptimal resource management (resulting from the conservative use of resources). Demand also has clear incentives for accepting long-term contracts. Firstly, long-term commitments give demand the means to hedge against the aforementioned peak prices. Secondly, efficient management of generation-side risk produces certain benefits. Although this factor is usually overlooked, it may represent an even greater incentive than protecting against volatile prices.[13]

---

[12] Some contracts may involve rigid constraints, such as in "take or pay" or "use it or lose it" formulas.

[13] The required expected return on an investment (for any asset, but particularly for a generating unit) is widely accepted to depend critically on the degree of risk involved (the higher the risk, the higher the expected rate of return). Therefore, if demand plays a role in the long-term market and contributes to risk management by concluding long-term contracts, it lowers generators' risk exposure, and with it their required expected rate of return. Similar reasoning can be applied to medium-term resource management, where long-term contracts may also provide for greater efficiency. Even if demand were risk-neutral, long-term contracts would indicate a more efficient outcome.

Therefore, in this context, a long-term market that would supplement the spot market and solve the risk aversion problem should spontaneously arise. The result would be that agent risk management would be entirely determined by market forces.[14]

In a liberalised context, both sides of the market also have to bear part of the risk involved in investment and resource management. A serious problem arises when this is forgotten, as it has by most retailers and, particularly, regulated retailers worldwide. They typically commit to purchasing electricity at the price applicable to their wholesale market operations but have yet to enter into long-term commitments (i.e. longer than a year). This *modus operandi* has primarily been driven by a blind reliance on supply being ensured by "someone".[15]

The need for learning processes in immature electricity markets

After more than two decades of operation, today's electricity markets cannot yet be considered mature. Even in those rare markets where demand is truly exposed to spot prices, long-term contracts (with a duration of more than 1 year) are not common. Most consumers are not sophisticated enough to understand the risks involved in electricity markets and the direct benefits that efficient risk allocation would provide. Demand (not only domestic, but also industrial) usually tends to make decisions using only short to medium-run criteria.

This lack of demand-side response causes the long-term market to malfunction, a problem that has no solution in the short run. The result is both insufficient generation investment and very conservative (and thus inefficient) medium-term resource management, which may well lead to future shortages. Note that the need here is not just for consumers to demand less energy from the market when prices are high, a typical goal of demand-side management programmes, but also, and especially, for them to enter into efficient hedging contracts to express their need for a higher level of generation reliability (i.e. to express their risk aversion).

The most orthodox solution to this problem would be to do nothing.[16] Consumers who fail to conclude contracts would experience high prices and the severe consequences of rotating blackouts. By the following year, some of them would recognise the need to protect themselves against this situation and would enter into contracts. This process would continue until consumers have learnt how to operate efficiently in the long-term market.

---

[14] This does not mean that introducing long-term markets guarantees an efficient outcome. Several experiences (the OMIP in the Iberian market is a clear example) have shown that the regulator's decision to establish (and even provide funding for) a long-term market (power exchange) does not guarantee participation. Efficient long-term markets arise when market actors are willing to participate. Regulators can help by creating a suitable, transparent platform for trade, but if the market structure is unsuitable, artificially implementing a trading floor will not induce participation.

[15] Allowing consumers to change their retailer with no penalisation does not provide the right incentive for them to conclude long-term contracts (see [34]).

[16] Educational programmes on the possible consequences of not concluding such contracts may help reduce the potential impact, however.

This reasoning has been defended by various authors in the literature, who support the contention that no specific security-of-supply regulation is required. The most common example, regarded as paradigmatic of this view, is the supply shock that hit the Nordic electricity market in 2002–2003 [52].

Regulators' risk aversion

Given the international panorama to date, a long learning period, which may include several rationing episodes, would ultimately be regarded as a problem attributable to the market rather than to inefficient consumer behaviour.

Electricity is an essential good that cannot be easily replaced in modern society, and shortages have significant social and political implications. Therefore, politicians, regulators and System Operators are acutely aware of the need for a reliable electricity supply. In most systems, the market rules change dramatically before consumers have time to complete their learning process, as in California and Ontario, for instance. The long-term market will never reach a steady state because it will be completely revamped before that can happen. In fact, underlying this problem is the principle that a wise regulator should not assign responsibilities to any individual who is not prepared to fully assume them. Furthermore, according to a common (although questionable) belief, for the most part, today's demand is not yet prepared to deal efficiently with the problem of long-term generation reliability.

This discussion can, therefore, be summarised as follows: politicians have much greater aversion to risk than almost any power consumer. Regulated rates preclude the need for protection against high prices, and even consumers initially exposed to spot market prices ignore reliability when making their decisions. Consumers appear to believe that the regulator will never allow supply shortfalls or inordinately high prices that would jeopardise their interests.

A consequence of the above is that demand does not respond appropriately on the long-term market. Consumers take no interest in a suitable level of adequacy, primarily because there is no real need to respond, and therefore do not factor this item into the pricing process. This curbs the installation of generation intended to ensure reliability.

12.2.5.2 The Consequences of Introducing Flawed Regulatory Rules that Distort Market Signals

The importance of having an appropriate pricing mechanism available, if the market fails to provide enough supply to meet demand, has long been debated in the regulation literature. Indeed, this is considered one of the cornerstones of the market model.

More often than not, however, regulators continue to distort the short-term marginal signal, trying to limit the revenue that generators can extract from the market. These measures are usually justified by the absence of adequate demand elasticity, and may be intended to achieve a number of objectives:

- One may be the decision of the regulatory authority to establish a regulated value for the value of energy not supplied.
- Another possible objective may be to limit market power since, if the reserve margin happens to be very tight, the generators could eventually bid (and be dispatched and paid) at extremely high prices.
- A further aim may be to artificially lower the inframarginal income of generating units. This approach is currently being used in some Latin American countries, in which, for various reasons, the only generating units that have entered the market in recent years are extremely inefficient, and therefore expensive, fossil fuel plants.

Some experiences in the implementation of this kind of measures are described below.

Price caps and bid caps

Market prices are in many cases explicitly capped. These caps range from quite low to very high values. For example, at the time of this writing, €180/MWh in the Spanish market, $1000/MWh in Alberta, $3000/MWh in ERCOT and $12500/MWh in the NEM.[17] Another mechanism is the "curtailment price" ("precio de falla") adopted in several Latin American power markets. This is a maximum market price defined by the regulator to be paid to generators whose production is committed to cover outages. An exception is made for plants able to substantiate higher production costs, which are allowed pay-as-bid prices.

Bid caps are constraints imposed on the prices submitted by generators. For example, in the PJM, ISO NE and NY ISO systems, there are explicit bid caps in place ($1000/MWh). These constraints are in some other cases implicit in the regulation. As an example, the Spanish law stipulates that generation units are "obliged to make economic bids" [19]. In the Irish market, the Bidding Code of Practice establishes that generators' bids must be based on the "opportunity cost", defined as "the value of the benefit foregone by a generator in employing that cost-item for the purposes of electricity generation, by reference to the most valuable realisable alternative use of that cost-item for purposes other than electricity generation" [1]. In California, the automatic mitigation procedure (AMP), implemented in 2002 to limit the ability of real-time market energy suppliers to exercise market power, basically consists of an automatic comparison with previous bids. If an offer price is too high, the AMP reduces it to a reference price in accordance with the cost of production in that power plant.

[17] The price cap is expressed in Australian dollars. Additionally, on top of this price cap, the Australian market also limits the remuneration a resource can capture on a weekly basis. This way, after exceeding certain thresholds, the maximum price a unit can perceive is further reduced. See http://www.aemc.gov.au/Electricity/Rule-changes/Completed/NEM-Reliability-Settings-VoLL-CPT-and-Future-Reliability-Review.html.

All of the above measures lead to a situation in which, in one way or another, the system's marginal prices are based on generation bids only, precluding the participation of demand in determining these prices.

Out-of-market interventions
Out-of-market interventions are "operating reserve shortage" actions. In some cases, when operating reserves fall below a certain level, SOs adopt measures such as voltage reduction and non-price rationing (rolling blackouts) to reduce demand (see [29]). Measures of this type complicate the price formation process in scarcity conditions.

These rules and their influence on short-term market price formation may affect both the suppliers' medium-term resource management and their long-term investments. Indeed, these regulatory interventions can hamper the recovery of investment costs for already-installed generation units. In the longer term, this may lead to expansion of the generation system in ways far removed from the theoretical ideal described above.

12.2.5.3 The Effect of the Lumpy Investments in Generation

Investments in generation are lumpy, meaning that certain technologies have a minimum feasible size (installed MWs). Moreover, for most generation technologies the production cost per unit of output decreases with the size of the plant, until it stabilises at a value much larger than the minimum feasible size, therefore encouraging investment in larger plants. When the size of the new generation investments is not very small when compared to the total peak demand of the system, then short-term prices may not be capable of sending an optimal signal for investment, since these prices would be strongly affected by the potential investment. If this is the case, the optimal amount of investment is not economically viable.

12.3 Review of Security-of-Supply Mechanisms to Meet Firmness and Adequacy Requirements

12.3.1 Introduction

Left to their own devices, wholesale electricity markets seem to be incapable of ensuring an efficient supply from the firmness and the adequacy standpoints. This section presents a classification and a critical review of approaches that have been used in practice in different power systems, when trying to achieve the desired results in these two dimensions.

The regulatory designs to ensure the security of power supply have evolved much in the last two decades, and, while no universal solution has been

forthcoming, international experience has helped narrow the range of possible measures that should be considered by regulators. Broadly speaking, they have two opposing strategies from which to choose, as described below.

- One possibility is to do nothing, in the belief that the market will provide the efficient medium- to long-term outcome. The regulator's non-intervention would be supported by the expectation that demand will (or will eventually learn to) manage the long-term risk involved in electricity markets (for example, by hedging with contracts to guarantee future needs). This is often known as the "energy-only market" approach.
- The other option is to act on behalf of demand, based on the opposite belief. In this case, the regulator designs a security-of-supply mechanism. This mechanism comprises the definition of a certain security-of-supply-oriented product (referred to hereafter as the "reliability product") aimed at ensuring the security of supply of the system (mainly to avoid scarcities), plus a procedure to determine the providers and the associated remuneration.
  Generally speaking, in this second case, all additional security-of-supply mechanisms require the regulator to first define and then (directly or indirectly) purchase one or more security-of-supply-oriented products from generators (or from demand response resources[18]). By directly purchasing these products on behalf of demand (or compelling demand to acquire them), the regulator seeks to guide the performance of the electricity system (operation, management, planning and/or investment) towards the optimal solution that, for the reasons already discussed, cannot be delivered by the market. This reliability product is generally provided by generators, who receive in exchange the extra income or the hedging instruments they require to both proceed with efficient investments (adequacy) and make resources available when most needed (firmness). The other counterparty is either a demand-side actor, compelled to purchase the product by the regulator, or the regulator itself (i.e. the system, the tariff) acting on behalf of demand. The definition of the reliability product comprises two basic elements, namely:
- The underlying asset: energy, capacity (installed or available), etc.
- The contractual features of the commitment: type of contract (forward or option contract, tradable or otherwise on a secondary market) and its characteristics (guarantees, penalties, force majeure clauses, lag period and contract duration).

[18] One possibility that may be considered is to allow consumers to participate in security-of-supply-oriented mechanisms (which is more straightforward in some than in others) by offering a product that is more or less symmetrical to the product required to generating units. Although this chapter focuses on the generation perspective, there is a growing trend to integrate demand in security-of-supply mechanisms. Indeed, many security-of-supply-related problems can be more appropriately managed with efficient demand-side participation. A conspicuous example of demand-side participation in this type of mechanisms is the Forward Capacity Market of ISONE, the Independent System Operator of New England, USA, see http://www.iso-ne.com/markets/othrmkts_data/fcm/index.html.

Price mechanisms and quantity mechanisms
Security of supply mechanisms can be classified depending on whether the regulator's main objective is to ensure a certain quantity of the "reliability product" or directly to set a price for the product itself.[19]

- Price mechanisms: a monetary sum that is determined by the regulator, and often known as the "capacity payment", over and above the income derived from the energy (spot) market, is devoted to remunerate the totality of the reliability product. In this scheme the reliability product is, in practice, the so-called "firm capacity".
- Quantity mechanisms: the regulator requires the demand to purchase a specific quantity of the reliability product, or buys it itself on behalf of the consumers. This product may adopt a variety of formats, e.g. a long-term forward contract for energy or a short- or long-term capacity credit. Depending on the system, the product may be traded bilaterally, through a centralised or de-centralised auction or by means of additional, organised (short- or long-term) markets.

A multitude of analyses from different points of view can be found in the literature [21, 22, 29, 43, 54].

The classification outlined above, (i.e. do-nothing, price mechanisms and quantity mechanism) is used in subsequent items as a guide for reviewing and categorising the range of approaches to be discussed next.

12.3.2 Do Nothing: The So-called "Energy-Only Markets"

The first alternative is to do nothing, i.e. the regulator makes a long-term commitment to refrain from intervening in securing supply. This method, based on non-interference with the market, has often been termed the "energy-only market" approach [8]. As previously mentioned, the lack of intervention by the regulator would be supported by the expectation that demand will (or will eventually learn to) manage the risk involved in electricity markets (for example, by signing long-term contracts). The regulator's position must remain unchanged even though the outcome may differ from what was initially expected.

As discussed above and in Chap. 2, theoretical microeconomic analysis of power systems shows that, under some ideal conditions, the short-term price resulting from a competitive market provides efficient outcomes in both the short and long run [6, 12, 15, 38, 45, 49]. Inframarginal energy revenues (including the

[19] For the sake of simplicity, the experiences discussed here are classified under one of the two extreme approaches (quantity-based or price-based). Intermediate or hybrid schemes are also possible.

so-called scarcity rents[20]) provide the necessary income for the recovery of both operating and investment costs.

Using this argument (among others), some experts (increasingly fewer in number) suggest that only purely market-based approaches would provide for efficient long-term security of supply performance. Note, however, that what does and does not constitute market intervention is not always clear among regulators and academics. As a result, the meaning of "energy-only market" usually depends on the system, the context or the author's point of view.

Some authors use the term to simply refer to the absence of any kind of "capacity-based mechanism" (such as the widely used capacity markets or capacity payments, discussed below), irrespective of the existence of other possible types of regulatory interventions relating to long-term security of supply. Some examples of these actions are:

- The regulator or System Operator enters into long-term contracts for energy or reserves (not only operating but also "load reserves" for use in scarcity situations, which are described in more detail below).
- The system operator is given full operational control in cases where a period of scarcity is inevitable. In other cases, when operating reserves fall below a certain level, the SO implements actions, such as voltage reductions and non-price rationing of demand (rolling blackouts), to reduce demand administratively without having to set prices to reflect the scarcity of supply [29]. A similar example is the maximum generation service contracted for by the SO in the UK [33]. These types of "out-of-market" measures render pricing complex in conditions of scarcity, and affect the satisfactory and expected recovery of generation investments.
- Under certain circumstances (usually low capacity reserve margins), the regulator is allowed to organise an auction to encourage new investments as a backstop mechanism to ensure security of supply.

Such actions must be included among the mechanisms to ensure long-term security of supply. The use of any one of these mechanisms is a clear indication that the regulator does not fully trust the market to deliver the desired outcome, in and of itself, and all of them clearly distort purely market-based signals. Therefore, to avoid ambiguity, the term "energy-only market" will be used here to refer to the strict "do-nothing" alternative.

Two extreme (and quite paradoxical) situations will help to clarify what an energy-only market is and what it is not: if the market agents spontaneously introduce a capacity market, with no regulatory intervention, the resulting scheme would be an energy-only market. Conversely, if the regulator compels every

[20] This refers to the income received when generation resources do not suffice to supply demand, and therefore, the price (which is this case is set by demand) is higher than any of the generators' variable costs.

demand to acquire firm capacity—or even some prescribed volume of energy—via forward contracting, the result is not an energy-only market approach.

From this perspective, it is very difficult in practice to find a market where the regulator is truly committed to just "waiting and seeing" and has waived the competence to intervene explicitly or implicitly. This stance is particularly difficult when the reserve margin is, or is expected to become, tight or even scarce. Then, the regulator might allow or merely suggest a third party to intervene. Such a third party might be the SO or even an incumbent utility.

While a number of markets, particularly in Europe, lack any explicit security-of-supply mechanism, it may be safely asserted that no system lacks at least an implicit regulatory safeguard to ensure security of supply. In some systems, the incumbent (now in a market-like context but still under partial yet sufficient governmental control) "shares the regulator's concern" about system reliability (France,[21] Italy and Portugal are examples).

"Latent" security-of-supply mechanisms are in place in Europe, with their origin in Directive 2005/89/EC,[22] which states that "The guarantee of a high level of security of electricity supply is a key objective for the successful operation of the internal market" and enables the Member States to impose public service obligations on electricity undertakings, inter alia, in connection with security of supply. It further provides that "Measures which may be used to ensure that appropriate levels of generation reserve capacity are maintained should be market-based and non-discriminatory and could include measures such as contractual guarantees and arrangements, capacity options or capacity obligations. These measures could also be supplemented by other non-discriminatory instruments such as capacity payments".

In some cases, another (not always acknowledged) reason why certain regulators do not implement an explicit security-of-supply mechanism is the existence of horizontal concentration. A concentrated market allows generators to ensure the recovery of a "reasonable" rate of return.

Thus, strictly speaking, whether pure energy-only (competitive) markets exist is open to question. That said, certain countries are clearly more inclined to believe in the market's ability to ensure long-term security of supply. Among the most representative power systems that are classified in the literature under the "energy-only" approach are ERCOT (Texas), NEM (East Australia), Alberta (Ontario), the UK and Nord Pool. However, none of them can be considered to fully rely on the "left-to-its-own-devices" ideal market mechanism approach. Indeed, all have some manner of implicit or explicit security-of-supply mechanism:

---

[21] In the preamble to the March 2010 version, the [36] bill in France included the following justification for the proposal to create a capacity market: "The objective is to ensure that all suppliers accept their industrial and energy responsibilities on behalf of their customers and do not rely on an implicit guarantee of delivery on the part of the incumbent".

[22] Directive 2005/89/EC of the European Parliament and of the Council of 18 January 2006 concerning measures to safeguard security of electricity supply and infrastructure investment. Official Journal of the European Union, 4 February 2006.

- In ERCOT, the emergency programme known as EECP (Emergency Electric Curtailment Plan) allows the System Operator to use reserves and out-of-merit units through "out-of-market" protocols. The objective is to avoid load shedding, which is the fourth and last step of the emergency protocol. The resulting short-term prices during these emergency interventions have been criticised for not reflecting the opportunity cost of providing the service. The System Operator may also conclude reliability must-run contracts with inefficient units for a variety of reasons.
- In Ontario, the Ontario Power Authority can enter into long-term contracts to secure an adequate reserve margin.
- In the UK, under the BETTA (British electricity trading and transmission arrangements), the Transmission Service Operator (TSO) is responsible for the long-term purchase of operating reserves. Operating reserve requirements are known to affect both short-term prices and consequently long-term investment signals. Thus, artificially modifying these requirements above the actual needs to cope exclusively with short-term security issues[23] can alter medium- to long-term market outcomes. This has been the experience in the UK's operating reserve purchasing process. Roques et al. [43], for instance, reported evidence to the effect that under the supplemental standing reserve tender (SSRT) initiated in October 2003 to increase reserve capacity, one of the impacts of this supplementary tender (requiring a much larger quantity than usually deemed necessary to maintain system frequency, with the hidden purpose of resurrecting some mothballed units) was to cause an immediate increase in forward market prices (i.e. in longer term signals). In December 2011, in a radical regulatory U-turn, the Department of Energy and Climate Change decided to adopt an explicit capacity mechanism of the type known as "reliability options" (see later in this section).
- In Nord Pool, the SO takes an active role by resorting to long-term contracting of "strategic reserves", discussed in the item dealing with quantity-based mechanisms.

Moreover, note that a substantial proportion of the equity capital in ERCOT, NEM (Australia), Alberta (Canada) and Nord Pool is publicly owned (on the generation or the retailer side, or both). In NEM in particular, around 63 % of generating capacity is government-owned or controlled [2].

12.3.3 Price Mechanisms

Price mechanisms rely on payments to existing and new generators, in proportion to the amount of the product (usually the firm capacity) they provide. Price mechanisms do not specify or limit the volume of generation capacity that can

[23] Note that a scarcity of generation supply is not a very short-term issue.

receive this payment. Price mechanisms basically reduce to the many varieties of the so-called *capacity payments*.

12.3.3.1 Capacity Payments

Capacity payments are a price-based incentive, in principle geared to achieving both efficient resource management (firmness) and investment (adequacy and sometimes also strategy expansion policy).[24]

However, there are two main problems with this mechanism: the reliability product is not properly defined (e.g. there is no commitment for the generators to do anything and just a weak economic incentive to be available when needed) and there is no guarantee that any desired investment target will be achieved, as by just setting a price the target might be overshot or undershot.

The "reliability product" that is remunerated in price-based mechanisms could be defined as "firm supply". Each generation unit's firm supply should represent its contribution to the system's overall security of supply. In practice, depending on the system, several alternative methodologies may be used to define firm supply. In most cases, it is mainly based on each generating unit's (expected) availability when most needed; however, other parameters are sometimes used in its calculation, such as the unit's variable costs (e.g. the smaller the variable costs, the larger the firm capacity assigned; variable costs are used, for instance, in Guatemala and Brazil).

Firm supply can be estimated ex-ante (using historical data, with or without ex-post corrections) or ex-post (based on actual performance in terms of the unit's contribution to system reliability). Some of the most relevant capacity payment experiences are reviewed in the following paragraphs.

Regulator-determined capacity payments were first introduced in Chile in 1982, together with an obligation to demand to contract firm generation capacity at this regulated price, which can be also seen as an anticipation of the later capacity markets. These payments were credited to each unit based on its firm supply (also known as firm capacity in this particular context). Firm supply was calculated using simulation with probabilistic models.

In the UK market model, from 1990 to 2001 capacity payments were made in disguise to all generating plants available during each half hour. Formally, the electricity price was computed one day ahead for each half hour of the next day. This ex ante price had to include the probability that during any considered half hour the available generation capacity might not be able to meet the entire demand. This component of the electricity price was equal to the loss of load probability (LOLP)[25] for the half hour in question, multiplied by the difference

[24] The theoretical underpinning of capacity payments was explained in Sect. 2.3.3.

[25] The LOLP value represents the probability of rationing. Another relevant measure that can be calculated directly from the LOLP value is the expected number of hours of rationing in a given

between the value of lost load (VOLL, i.e. the regulator's estimate of the cost of energy not supplied) and the plant's bid price (if not dispatched) or the system marginal price (if dispatched). During the time that this mechanism was in use there never was an instance of non-served energy in the UK system and the contribution of this component of the market price to the remuneration of the generators should have been negligible. However, this was not the case, since the method of ex ante computation of LOLP was heavily biased towards providing high values; it could also be easily manipulated by the two large generation companies at the time. In consequence, this component of the energy-only market price was for all practical purposes some sort of capacity payment. Criticism was levelled against this mechanism on several counts. Newbery [35] pointed out that some companies artificially increased the LOLP, and thus capacity payments, by declaring certain units unavailable. Green [24] noted that most of these abnormal payments were, rather, the result of the poor definition of the method used to determine new units' availability factors. Roques et al. [43] conducted a thorough analysis of the major shortcomings of this mechanism. These capacity payments disappeared with the introduction of the decentralised NETA model in 2001 (known as BETTA since 2005).

The capacity payment mechanism in the early design (1995) of the Argentinean system consisted of a regulated price ($/MW) applied to two different values of firm capacity. The first one is the average annual power of each generating unit that is dispatched during non-valley hours by a computer model, assuming an extra-dry hydrological year. The regulated price also applies to the power of each unit that is actually dispatched each day, but only for the amount that exceeds the previous annual average for the unit. It also happens that each generator is not allowed to bid, in the actual day-ahead markets for the following year, above the value of the bid that the generator declared for the annual simulation. Since both remunerated capacity values result from economic dispatches based on declared bids by the generators, the generators are incentivised to bid so that they try to maximise their total energy plus capacity remuneration, instead of bidding on the basis of their variable costs, which would result in the most economic dispatch. This results in loss of income for the generators and an inefficient operation. A subsequent design of the Argentinean scheme has modified the fraction of the remuneration of generators that depended on actual production; this component is based on the capacity of units that are available during the 90 h of highest demand each week [14], which are determined ex ante by the regulator.

The "capacity guarantee mechanism" initially implemented in Spain when the market opened in 1998 provided an extra payment based on the average availability rate for thermal power plants (subject to a minimum yearly production

(Footnote 25 continued)
period of time. Many systems define their reliability standards using this latter measure; for example, US power systems usually establish a maximum cumulative rationing period of 1 day in 10 years.

requirement[26]) and on average historical production for hydro units. This scheme was harshly criticised [40]. On the one hand, it was judged to be oversimplified in terms of firmness, as it lacked effective incentives for generators to be (or penalties for not being) available when needed. And on the other, it proved to be extremely unstable in terms of adequacy: the total volume of payments for this item declined from an initial €7.8/MWh of system demand at market start-up in 1998 to €4.8/MWh in 2006.[27] After two-years of debate [7, 8], the Ministry of Industry introduced a revamped design consisting of two differentiated services [32]: an availability service, aimed at allowing the System Operator to enter into bilateral contracts, with terms of no longer than 1 year, with peaking units (such as, for instance, fuel-fired plants and limited energy hydro plants), and an investment service, for units larger than 50 MW, which receive an annual capacity payment (expressed in euros) per installed megawatt during their first 10 years of operation.[28] This latter investment incentive depends on the value of a ''reserve margin index'' ("*índice de cobertura*" in Spanish, or IC), calculated by the System Operator. When the value of this index is below 1.1, the payment is set at 28000 euros per installed megawatt and year. If the value of the index is above 1.1, this payment is lower.

In Italy, the regulator determines a fixed capacity payment. This mechanism was initially meant to be transitory, but it has been in force for more than 10 years. The payment is designed to provide additional remuneration to all power plants in Italy whose output can be considered manageable (wind or run-of-river generating units are among the technologies excluded). The amount paid depends on availability during "high-critical" and "mid-critical" days, which are designated by the Transmission System Operator. The mechanism has been criticised for not ensuring the recovery of fixed investment costs [11], although, for the time being, it appears to be designed more to avoid mothballing (i.e. to keep existing units in operation) than to foster new investment.

In Ireland, under the SEM (Single Electricity Market) established in 2007, generators receive capacity payments to supplement the revenues from the centralised pool wholesale price. In general terms, annual capacity payments are determined by the regulator on the grounds of three parameters: the system's capacity requirements to comply with security standards, the annual carrying cost

[26] Plants had to produce at least 480 equivalent hours every year to be entitled to receive the capacity payment. The measure was designed to require plants to prove that they were minimally reliable. The rule generated obvious inefficiencies, however, for it meant that high-cost peaking units had to be uneconomically dispatched to receive the payment.

[27] The regulator (i.e. the Government) began to reduce capacity payments as market prices began to rise above expectations because the original purpose of capacity payments, in part, was to remunerate existing generating units for stranded costs.

[28] This 10-year condition is aimed at rewarding CCGTs only, which entered the system after the market opened in 1998. This design is clearly "contaminated" by the windfall profits debate that calls into question the income that mainly nuclear and hydro plants (installed in the former regulated context) receive under the new market scheme.

of the best new entrant (generally, the most efficient peaking unit) and the VOLL. The product of the first two parameters determines the total amount assigned to cover capacity payments, and this amount is distributed among all generating units[29] according to complex criteria aimed at reflecting their contribution to overall security of supply. These payments depend on the declared availability at each hour, and the hours are weighted in accordance with the ex-ante expected LOLP and the ex-post calculated LOLP; therefore this particular version of the capacity payments also rewards firmness, albeit only partly. For further details on the (long) formula used to determine payment distribution, see SEM [46]. The per unit payments are updated on a yearly basis; therefore, this mechanism fails to provide a predictable and stable source of income for generators, and it does not reduce their exposure to risk significantly, since they do not know precisely how the payment will evolve in years to come. Since it has been acknowledged that it would be more appropriate to stabilise the payment for a minimum number of years (at least for new units entering the system), the design, starting in 2013, will reduce the updating frequency to every 3 years, which is not a significant improvement, really, given the times to build and the economic lives of power plants.

Capacity payments have been implemented in other electricity systems as well, primarily in Latin America, in countries whose regulatory schemes were reformed to introduce market-based design. In Colombia they were replaced by the reliability charge mechanism (described below), while they are still in force in others (Peru, Chile and the Dominican Republic). In some cases, these capacity payments coexist with other security-of-supply mechanisms (such as mandatory long-term energy contracting).

12.3.4 Quantity Mechanisms

The security-of-supply measures classified under the heading "quantity mechanisms" differ from the price mechanisms discussed above in that the regulator relies on a market-based instrument to set the price for the reliability product. This approach theoretically solves the main problem posed by price mechanisms: instead of setting a price and then hoping for the right amount of capacity to come into the system, the regulator establishes the desired quantity and allows the market mechanism to set the right price.

This method is based on the belief that only market-based solutions are efficient. Unfortunately, real life usually stands at a substantial distance from the

[29] Although this method has been considered a price mechanism, the regulator really determines the total amount of the capacity payment to all units. Therefore, the price of capacity depends on the separately computed quantity that is entitled to receive the remuneration. There is a (hyperbolic) relationship between quantity and price, since the regulator only fixes the product of both.

ideal (markets are often far from fully competitive, significant entry barriers are in place and regulatory design is flawed more often than not).

Various quantity-based experiences are analysed in the next subsection, beginning with a review of the flaws of the initial mechanisms, the so-called "capacity markets". The present designs, re-worked to solve these problems, are then discussed.

12.3.4.1 Capacity Markets

The term "capacity market" was originally used to describe the pioneering markets introduced by some regulators to trade a particular (reliability) product (roughly, MWs of installed capacity) under certain particular conditions (short-term markets, either bilateral or centralised, and short contract commitments). As for the term "energy-only market", the "capacity market" designation can be also quite misleading, since the underlying characteristics of such a mechanism are not always clear: does it refer merely to a mechanism for establishing the obligation to buy "MWs of installed capacity" or does it also cover short-term markets and commitments?

In the earliest capacity markets, demand was required to contract the capacity (expressed in MW) needed to supply its future consumption. The regulator usually capped the amount of capacity a generating unit was entitled to sell. This cap is to some extent easy to determine in the case of only-capacity-constrained thermal plants. However, the attempt to implement this approach with energy-constrained generating plants poses obvious problems. In thermal only (non fuel-constrained) systems, a given plant's availability can be assumed to be unrelated to the availability of the rest of the plants in the system and also (sufficiently) unrelated to peak demand. This is not at all the case in hydrothermal systems, which are mostly energy constrained. The immediate consequence is that "capacity" is obviously far from having "the ability to produce energy when needed".

The former ICAP in the Eastern USA (PJM, NYISO and ISO-NE)

ICAP markets are the capacity market mechanisms most extensively studied. Even today, they are an inevitable reference, mainly because of the poor results obtained.

These mechanisms consisted of requiring every load serving entity (LSE) to back up its expected peak-load capacity requirements (plus a reserve margin) with "capacity credits". Generating units received credit for their installed capacity (that is what ICAP stands for: Installed CAPacity). Each LSE had to purchase a certain amount of the product (the credits), which was supposed to guarantee that there would be enough installed capacity to meet their expected demand (plus the aforementioned margin) at peak hours.

Capacity is not always available, however, and Independent System Operators (ISOs), beginning with PJM, aware of the firmness problem developed the concept of UCAP (Unforced Capacity). Under the UCAP concept, each ISO was able to

lower the capacity of each unit that was given credit (that is, the quantity of the product that the generator was entitled to sell), depending on its actual historical availability.

In the early designs, UCAP was calculated as an average available capacity over long periods of time (typically a season or even a whole year) regardless of whether unavailability occurred during a period of scarcity. Thus, the incentive to be available when reserves were tight was exactly the same as for any other hour of the year.

The ISOs, in their attempt to encourage generators to make their installed capacity available, sought supplementary rules. As a result, an additional condition was introduced for generators willing to participate in the ICAP mechanism: a must-offer requirement in the day-ahead market [18]. Unfortunately, this did not solve the problem, because of the difficulty inherent in establishing a method that did not entail full reliance on self-reporting by generators (the must-offer requirement was ineffective, as unavailability could be masked by high-priced bids).[30]

Controversial ICAP performance

In PJM[31], the PJM Market Monitor conducted an analysis to assess whether unit fixed costs were covered by the prices received by generators from the PJM markets plus the ICAP payments, and concluded that investment costs were not being recovered.

In addition to the non-recovery of investments, the design of these capacity markets posed another significant problem: extreme price volatility. Capacity market prices tended to fluctuate from very low during long periods when the system had a large reserve margin, and extremely high when insufficient capacity was available [18].

The inelasticity of the demand for capacity was identified as the main reason behind this price volatility, and proposals were soon forthcoming to plot a downward-sloping demand curve that would better represent demand interests. At present, a variable resource requirement demand curve has since been defined (an elastic price-quantity curve) for each of these capacity markets.

Demand inelasticity was not the only or even the predominant reason for these market results, however. Rather, these failures were the inevitable consequence of other design flaws, which are reviewed below.

Why did prices soar from near zero to sky high?

Fixed investment costs for generating units are stranded or sunk costs. Since, according to microeconomic theory, these costs should have no effect on future

[30] Although some sort of monitoring is possible, the only alternative involves conducting random on-site checks on a unit-by-unit basis, the methodology commonly used in Latin American designs, such as in Guatemala.

[31] PJM RPM Filing, Bowring Affidavit, at 15 (2006).

decisions, it is not rational for generators to internalise their investment costs in their bids in (fully competitive) spot markets.

If a capacity auction is called shortly before the delivery date, only units that are up and running can participate. As these units cannot internalise their investment costs in their bids, the generators have to bid the cost associated with providing the reliability product. What, then, is the additional cost of keeping the installed capacity (ICAP) of an existing unit operational? In most cases, it is obviously near zero. This was why prices tended to fall dramatically for a long time.

Conversely, when the capacity reserve margin tightens, the reliability product is scarce, which is reflected in the price. In addition, in such scenarios, the market is more vulnerable to the exercise of market power, especially where the demand curve is inelastic and unresponsive to prices.

These were the conditions prevailing in the early ICAP markets. Consequently, the prices designed to supplement the return on investment in energy fluctuated from near zero to very high values.

Thus, instead of the market providing the stable price signal sought by investors, the end result was that another (even more) volatile short-term market was created. The fear of energy scarcity was replaced by fear of installed capacity scarcity.[32]

In the end, generators received extra income that supplemented their earnings from energy sales, but this remuneration was neither certain nor could it guarantee in advance that fixed investment costs would be recovered.

Was this volatility unavoidable?

Volatility might have been reduced if the lead time, or the time lapsing between auction and delivery, had been lengthened to allow potential new entrants to participate, a solution put forward by [51], in connection with electricity systems in general, and Chandley [18], with regard to the PJM ICAP. Further changes were also recommended, such as providing for longer term contracts to enable investors to obtain financing more readily.

Lack of locational signals for capacity

As network congestion is significant in all these eastern US markets, reflecting the value of energy in terms of where it is consumed is considered a necessary element in market design. In this context, if enough liquidity and market competition can be ensured, these capacity market mechanisms should send signals that reflect the varying price of capacity at different locations.

Subsequent designs in these three systems attempted to correct the most consequential design flaws, as will be seen in later items.

[32] Although the consequences of a scarcity period in this new market also had an undesired economic impact, energy rationing was not one of them, because the reserve margin with respect to the expected peak consumption was defined by the regulator.

France

As mentioned earlier, in March 2010, the bill on the "New Organisation of the Electricity Market" [36] proposed the institution of a capacity market. The plan would require retailers to acquire "guarantee certificates" (*certifications de garantie*), which would be allocated by the SO (together with "elastic demands") among generating units according to their "total technically available capacity". The CRE (*Commission de Régulation de l'Energie*, or Energy Regulatory Commission) would annually calculate the penalties to be paid by market agents in the event of non-compliance, "to provide agents with incentives to invest in new demand response or generation capacity".

This is a preliminary proposal, with the final implementation details still missing; as written, however, the bill apparently includes the main characteristics of the ICAP design discussed above (such as the absence of a lag period and of any specific provision on long-term contracts). The final design will hopefully not contain the same flaws.

Other capacity markets

Other capacity markets are in place in Guatemala and Western Australia.

In the electricity market design in Guatemala, the regulator requires demand agents to hedge their future consumption. These so-called "monomial" contracts consist of two related (but not necessarily correlated) products: their future energy consumption (linked to a profile) plus their "firm offer" needs (capacity credits of a sort).

Reserve capacity procurement was instituted in Western Australia in late 2004. All demand agents must buy capacity credits to cover their share of the system's future capacity requirements. Both the system requirements and the capacity credits assigned to the generating facilities (and also to some demand-side management resources) are determined by the Independent Market Operator (IMO) on a yearly basis. Capacity credits can be traded either bilaterally or in a centralised auction, which is only held (and conducted by the IMO) when bilateral contracts have not fully met the total requirement.

12.3.4.2 Long-Term Auctions for Delayed-Delivery Reliability Products

This approach consists of (often centralised) auctions for longer term contracts, but also features delayed delivery (the aforementioned lag period, measured in years) to give the successful bidders in the auction time to build their plants, after having ensured the reduction on their risk exposure.

Colombia: the reliability charge

The Colombian power system pioneered the wave of change in the regulatory design of security-of-supply mechanisms and served as inspiration, directly or indirectly, for other reworked designs, which are reviewed in later paragraphs.

Highly dependent on hydro-generation, the Colombian electricity system is therefore particularly sensitive to the El Niño-Southern Oscillation phenomenon, a cyclical climate pattern in which, roughly speaking, every 5–8th year is severely dry.

The first scheme adopted in the Colombian market, in use from 1996 to 2006, was a capacity payment that was determined by the regulator. Although no scarcity arose in that period, the substantial narrowing of the reserve margin fed concern about the suitability of the scheme.

Its effectiveness had been called into question almost from the outset. The capacity payments assigned to generation plants in the Colombian electricity market provided neither a suitable incentive for efficient availability management in scarcity periods[33] nor a stable and trustworthy long-term signal for potential investors.[34] A consultation process on the flaws detected in the mechanism was launched in 1999, and several alternatives were proposed as a result. The finally chosen approach consisted of replacing the capacity payment with a quantity mechanism, without the flaws inherent in the PJM arrangement discussed above. The original proposal, put forward in 1999 in response to a request by ACOLGEN (the generators' association), was subsequently described by the consultants who developed it in [51].[35] The two primary features of this proposal were the introduction of the "reliability option" as the new reliability product and its purchase in a centralised long-term auction.

The reliability option[36]

Generally speaking, the so-called "reliability option" is a call option type of contract. An annex has been devoted to this specific method because of a number of reasons. On the one hand, its design is somewhat more complex than most of the other capacity mechanisms and requires a more detailed explanation. On the other hand, presently this is the method that is more seriously considered in different

[33] They had an adverse effect on efficient system planning. Since, the firm capacity of the hydro plants depended critically on the water reservoir level in the "dry season", generators managed their reserves so uneconomically that reservoirs were at their full capacity in that season.

[34] A capacity payment is simply a regulatory commitment. In the Latin American context, in which investors perceive regulatory risk to be high, the regulator realised that providing long-term contracts with the distribution companies as counterparties was a better solution to mitigate such risk aversion.

[35] The inception of this idea took place during internal discussions on a possible approach for dealing with reliability of supply in the Spanish power system in the Spanish Electricity Regulatory Commission between Miguel Ángel Fernández Ordóñez and Ignacio Pérez-Arriaga in early 1998, as well as in a related concept proposed by OFFER in its July-1998 review of the trading arrangements in England & Wales. The idea was later refined in meetings at the Secretary of State for Energy in Argentina, with the participation of Larry Ruff in September 1998. The full concept of reliability options was developed in 1999 during the consultancy job for ACOLGEN of the authors of the paper [34]. Vázquez et al. [50] is an in-depth analysis, by the same authors, of the potential implementation of this approach in the Dutch power sector. A complete description of the method of reliability options is given in Annex A to this chapter.

[36] In the mechanism finally implemented, commissioned to Peter Cramtom and Steven Stoft [20], the reliability product was called the firm energy obligation.

regulatory jurisdictions when trying to design an advanced approach to generation adequacy and firmness. Finally, the complete concept of the reliability options method was originally developed by a team of researchers from the Institute for Research in Technology, at Comillas University, most of which are also co-authors of this book.

The long-term auction

The other main feature introduced in the original proposal of reliability options[37] was the centralisation of reliability option purchases through a long-term auction. The objectives were to intensify competition and benefit from economies of scale (pooling the system's numerous retailers, sometimes small and often regulated, to enable large investors to participate), as well as to enhance transparency (to prevent vertically integrated companies from taking advantage of non-transparent agreements).

The auction design ultimately adopted in the implementation of the Reliability Charge (Cargo por Confiabilidad) in Colombia uses a descending clock format, including (among other details) a downward sloping curve to establish the price dependency of the amount of reliability product purchased. Another relevant characteristic of the process is that the new and existing plants are subject to different rules (e.g. existing plants are price takers in the auction).

Brazil

After a series of rationing episodes in 2001 and 2002, imposed on all types of consumers in a region accounting for 80 % of total national consumption, ensuring long-term security of supply became a genuine priority in Brazil. Several in-depth analyses were conducted, after which experts in the field identified a number of flaws in expansion and contracting procedures. The proposal put forward, partially inspired by the aforementioned solution proposed several years earlier for the Colombian system, resulted in the mechanism currently in place [3].

The main features that distinguish the Brazilian from the Colombian mechanism are summarised below.

- Separate auctions are called for existing units and new entrants. In the former, the lag period and contract duration are significantly shorter (a 1-instead of a 5-year lag, up to 15 years instead of up to 30).
- Two reliability products are defined: a forward financial energy contract for hydro units and an "energy call option" (which, in very general terms, resembles the Colombian reliability option) for steam plants.
- The regulator has a backstop mechanism that allows the government to hold specific energy auctions prompted by energy policy decisions. In 2008, for

[37] This key feature of the method has been kept in the implementations made in Colombia and ISO New England, as well as in the detailed studies for the Dutch power sector in [50] and for the market reform in the UK in 2011.

instance, a special auction was held under this mechanism for 1200 MW of co-generated power produced with sugar cane biomass (see [10]).

ISO New England

In ISO-NE, the forward capacity market (FCM) replaced the earlier ICAP mechanism [21] and ISO New England [27]. This new framework shares the principal features of the Colombian mechanism, barring some of the complexities aimed at coping with different generation technologies and including locational signals, i.e. capacity requirements and clearing prices are calculated separately for different areas.

Since the auction-based mechanism is very similar to the Colombian mechanism, it is not described here. One outcome worthy of note, however, is the remarkable degree to which demand has been integrated as a potential provider of the reliability product, a development observed in designs in place in other systems as well.

PJM's Reliability Pricing Model (RPM) and the new NYISO ICAP

The poor performance of the capacity markets originally implemented in the PJM and NYISO areas led to significant design modifications aimed at correcting most of the shortcomings analysed earlier. Briefly, the new design [41] consists of an auction for a reliability product, which in this case differs from the Colombian mechanism. The product design is a variation on the former UCAP, where the availability is now measured in a much more detailed manner that now does take into account the availability when the production is actually needed. The timing provisions envisage both a longer lag period and longer contract durations.

Others

Other systems where a market based on long-term auctions for delayed delivery reliability products have been implemented include Chile [9, 50] and Panama[38] [17] and [48]).

12.3.4.3 Strategic Reserves as the Reliability Product

This final category of quantity-based mechanisms includes schemes based on purchasing "strategic reserves". Traditionally a particularly controversial type of reliability-based product, these reserves consist of separating off a certain amount of generation capacity that does not participate in the energy market unless the regulator or the SO finds it necessary (according to more or less objective criteria).

A number of papers, such as the remarkable analysis conducted by [22], have argued that if the criteria used to signal the initiation of both the production of

[38] In Panama, distributors must conclude contracts in advance via public auctions for both their expected energy supply and their capacity requirements (peak consumption), taking into account a safety margin determined by the regulator.

reserves and the purchasing process are well designed, this mechanism can yield acceptable[39] results.

In Finland, Norway and Sweden, the SO is in charge of purchasing "load reserves". These reserves (another name for strategic reserves), which are completely separate from the standard primary, secondary and tertiary (balancing) operating reserves used to restore frequency, are earmarked for times when demand comes close to exceeding available production capacity; in other words, they are called upon to supply energy when generation becomes scarce. The main objective of this type of mechanism is usually to prevent the mothballing of old units; in some cases, however, the SO is also responsible for establishing the rules for offering these electricity reserves on the market. Obviously, this may significantly distort price signals (by converting the SO into an irregular market agent).

New Zealand is a hydro-dominated system (this technology accounts for 65 % of the energy produced in an average year). As in most similar systems, the concern has been to ensure enough production resources during dry years.[40] The mechanism designed to overcome energy constraints under such conditions consists of contracting for strategic reserves, which may include either new or old equipment [31].

Contractors are selected through centralised public auctions, and their sole responsibility is to supply energy and capacity during scarcity periods (dry years). The design of the strategic reserve mechanism includes the price at which reserve capacity is to be offered on the wholesale market. This price is set high and, ideally, serves as a threshold for detecting scarcity situations[41]; if the reserve capacity price were too low, it would compete with ordinary generation, which would deter new investment by generators.

Prices above the generation reserve price are expected to be very rare, although not impossible. Spot prices may still peak at very high values if extreme circumstances exhaust the available reserve capacity.

While this trigger price plays the same role as the strike price in the reliability option (as in the New England, Brazilian and Colombian mechanisms), strategic reserves and the reliability option are two very different products. The main difference is that in the former, energy can only be sold if the spot market price reaches (or exceeds) the unit trigger price established by the regulator; in the reliability option, by contrast, production can also be sold at other times. Therefore, strategic reserves represent generation that is kept off the market under normal circumstances.

---

[39] Although, from our point of view, inferior to those the reliability options scheme can provide.

[40] The objective of New Zealand's Electricity Commission is to ensure that supply remains secure even in a one-in-sixty dry-year event, i.e. in a drought of a severity that can be expected to occur once every 60 years.

[41] In the December 2008 update, for instance, the Whirinaki reserve energy trigger price was set at $0.387/kWh ($387/MWh).

12.4 Principles and Criteria for the Design of Security-of-Supply Mechanisms

12.4.1 Introduction

It has been shown in this chapter that the electricity market, without any regulatory intervention, is not able to provide sufficient generation availability when needed, and that the solution to the problem necessarily entails the development of additional signals to ensure firmness and adequacy of supply. Indeed, regulatory mechanisms to secure long-term supply have been designed and implemented in a number of power markets around the world.

Once the regulator has decided to undertake the task of "helping" the market reach what is deemed to be the most efficient outcome, the next key question is how to introduce the necessary adjustments to the market designs in place so as to achieve the long-term objective pursued. This is particularly complex and controversial, because in the end, the responsibility for all medium- to long-term planning may directly or indirectly revert to a central planner. Sight should not be lost of the fact, in this regard, that avoiding the potential inefficiencies of central planning was one of the driving forces behind the wave of liberalisation undertaken a few decades ago.

The items below discuss the steps and elements regulators should consider when designing a mechanism aimed at overcoming inefficiencies in the firmness and adequacy dimensions. The need for such a mechanism is assumed to have been decided already.

12.4.2 Design Elements

The first step in the process of introducing a security of supply mechanism for adequacy and firmness is to verify that the barriers interfering with the proper functioning of the market have been removed as much as possible.

Generally speaking, all mechanisms aimed at solving the security-of-supply problem involve four major decisions to be made by the regulator:

- Identification of the counterparties (buyers and sellers).
- Definition of the reliability product, i.e. the product that the demand side must purchase from the generation side.
- Determination of whether to set the quantity, the price or a quantity-price curve for the purchase of the reliability product.
- Establishment of other details, such as whether contracts are to be bilateral or auction-based, and whether or not the purchasing process is centralised.

12.4.2.1 The Counterparties: Buyers and Sellers

The buyers

The buyers are the part of demand on behalf of which the regulator makes decisions. The regulator must decide whether to act on behalf of all or only some portion of demand.

Care needs to be taken to avoid creating free riding issues and cross subsidies. It may happen that the agents involved in the regulated mechanism are not the only parties benefiting from it.

The sellers

The regulator also has to define who is entitled to act as a seller in the mechanism. In some cases, all types of units are allowed, while in others, only new investments or certain technologies may participate. Depending on the particular case, discriminating among different units may create market segmentation with undesired long-term effects.

12.4.2.2 The Reliability Product

The "reliability product", i.e. the product to be sold by generating units in return for the additional hedge or the source of extra income introduced by the security-of-supply mechanism, must be carefully and precisely defined.

Determining the product to be purchased from generation is of the utmost importance and complexity. In practice, numerous alternatives have been proposed and implemented: fixed or flexible long-term energy contracts (like forward energy contracts or the reliability options), certificates based on installed generation capacity (or reservoir capacity), certificates based on available generation capacity (or available energy), certificates based on the installed capacity of a certain technology, long-term reserve requirements, physical units to be operated by the SO under certain conditions, and financial energy contracts, among others.

Product definition may determine the success or failure of the entire mechanism. Introducing a mechanism that entails buying a reliability product clearly affects generator decisions. Therefore, when defining the reliability product, the regulator must analyse the generators' expected response to determine whether this response leads to an efficient result or otherwise. For instance, if the regulator decides to buy just installed capacity, it will probably obtain the capacity with the lowest investment costs, but perhaps also with low availability. If the regulator decides to pay for the water reservoir level in the "dry season", reservoirs will be filled to their full capacity in that season.[42] When the consequences of the product

[42] A clear example of how these mechanisms can predetermine the design of new investment can be found in Guatemala. In this market, the capacity payment is related to the average production of generating units in the four peak hours of each working day in the dry season (from December to May); therefore, new small hydro plants are designed to have a reservoir for daily regulation whose storage capacity (MWh) is the capacity of the turbine (MW) times four (h).

definition are not thoroughly evaluated beforehand, the results may be highly inefficient.

A certain consensus has been reached around the idea that the reliability product should remunerate the ability to produce energy at "reasonable" prices (whatever "reasonable" might mean, usually well below the so-called non served energy value) in the event of system scarcity. Nor is scarcity a notion that can be readily defined. In this respect, market price seems to be the most appropriate and transparent indicator, but other alternatives, less objective and possibly also less efficient, have been also implemented, such as a low threshold of the value of the operating reserve margin in real time, or the a priori (e.g. 1 year ahead) definition of periods with more or less potential for scarcity, for instance in terms of the anticipated demand level.

Product characteristics: design issues

Some of the design characteristics of the reliability product, which largely determine the results of the mechanism, are listed below.

- Contract duration: large investments usually require long-term contracts to enable investors to obtain project financing under conditions that ensure that the plant will be competitive.
- Lag period (also known as lead time): as previously defined, this is the time lapsing between the date when the commitment is concluded and the date when the product must be delivered. Projects whose construction takes longer than the lag period will not be able to take advantage of the mechanism fully. Extreme situations arise when the lag period allows no new potential investment to participate (as observed in the short-term ICAP markets in place a few years ago in the North Eastern US).
- Penalties: the clauses establishing the penalties to be applied in the event the generator does not fulfil its commitment. An appropriate range of penalties is essential to provide generators with incentives.
- Force majeure clauses: these clauses exempt a party from liability if some unforeseen event beyond the control of that party prevents it from performing its obligations under the contract.

  Some of these unforeseen events may be explicitly mentioned in the contract, although if these clauses are not carefully designed, they may hinder attainment of the reliability objectives. For example, in the former Chilean market scheme (where some sort of capacity market existed), generators that did not have enough energy to supply their contracts had to pay compensation at much higher prices than the average production cost (the cost of the energy not served, as defined by the regulator). Nevertheless, before the 1998–1999 supply crisis, exceptional conditions were also defined in which that compensation need not be paid. A drought that was not in the 40-year statistical record used by the regulator to calculate regulated prices was defined as a "force majeure" condition (in Article "99 bis" of the former regulations). After the 1998–1999 crisis, which led to severe supply disruptions with scheduled load curtailment,

the regulator changed the controversial article, eliminating drought and other circumstances from the "force majeure" conditions [25].

- Guarantees: these are usually required by the regulator as a means of hedging against company credit risks. However, they may sometimes introduce entry barriers.
  Long-term financial contracts for energy usually require physical generation capability (i.e. the physical ability to self-produce the quantity committed) as a guarantee. This physical capability is usually quantified on the basis of the firm capacity (or energy) concept.

12.4.2.3 Price Versus Quantity

The regulator must decide whether a price-based, quantity-based or price-quantity-based curve is to be offered on behalf of demand.

Resorting to a fixed-price mechanism may lead to a security-of-supply level that is either too high or too low. Similarly, using a fixed quantity may result in an overly high price (particularly if the regulator does not expect much competitive pressure or if the possibility of new entry is limited somehow).

Gradual and flexible requirements better reflect the utility that each level of security-of-supply provides the buyer (demand, in this case represented by the regulator). Additionally, they help reduce market power and also provide more information about how imminent a scarcity situation may be.

12.4.2.4 Other Details

The regulator must also determine certain characteristics of the purchasing process. Two relevant decisions in this respect are:

- whether the purchasing process is centralised or otherwise, i.e. whether it is managed by retailers (distributors or load-serving entities when part of the demand is regulated) or the regulator.
- whether the purchasing process is arranged through bilateral agreements or public auctions.

Regulatory intervention

As mentioned earlier, the interaction between demand and generation in a purely market-based context, in which both express their preferences, should lead to an efficient outcome. If a regulatory mechanism is implemented, demand preferences are partially replaced by the aforementioned decisions, assigning the regulator a significant role in the long-term equilibrium.

Regardless of the market-based procedure followed for purchase of the reliability product, the parameters that define it and the offer curve largely determine not only the level of security-of-supply but also the type of generators that will provide it.

A.1 12.5 Annex: The Method of Reliability Options

The reliability option seen as a way to introduce a market-compatible price cap
One possible way of describing the motivation for the reliability options design could start from the idea of implicit insurance. When there is a security of supply problem, and prices for final consumers are high and shortages appear, the situation may become a very difficult political problem for the regulator. Consumers are also hurt but, as we have just seen, they do not react. So it is the regulator who should try to protect consumers (and also himself against the associated political risk), through a change in the market design. He would like to impose a price cap on the market, but this is a problem with the very old and inefficient plants, which may not produce even if needed if the price cap is low, and also with the new entrants, which may be discouraged with the perspective of not being able to see high prices. In financial markets, a buyer who wants to get a price cap on his future purchases can acquire a certain kind of derivatives, known as a *call option*, which gives him the right, but not the obligation, to buy the item at a predetermined price (the strike price) in exchange for a premium fee. This is a way to obtain market-compatible price caps. Accordingly, we propose that the regulator should buy call options from the generators, probably through a centralised auction, and therefore isolate consumers from the high prices.

At the same time, the generator that is selling a call option is giving up receiving the part of the spot price that is above the strike price in exchange for the premium fee or, in other words, he is exchanging some uncertain and very volatile income from the spot market (which only happens in periods of stress for the system) for a certain stable and predictable remuneration. This greatly reduces the risk for generators, which, as has been analysed in this chapter, is one of the major reasons obstructing that the "right" amount of new investment may happen at the "right" time. This is particularly true for peaking units, which are exposed to much larger income volatility.

Thus, two main objectives are achieved with this mechanism: on one hand, consumers do not have to bear the risk of having high energy prices reflected in their bills; on the other hand, efficient economic signals for new investment are being provided.

Two relevant time-related parameters associated to the reliability option
As with any long-term contract, there are two relevant time-related parameters associated to the reliability option contract: the lag period (also known as lead time, or planning period, as defined in the Colombian regulation) and the duration of the commitment. Defining a sufficiently large lag period is essential when the objective is to allow enough time to build the plant, and the latter has also to be long if we want to reduce risk exposure and thus facilitate project financing.

Further sophisticating the product definition: tying a physical obligation to the financial option

It is worth noting that the financial option without any associated physical delivery obligation is a tool that serves exclusively to hedge market agent's risks. Whether the mere efficient allocation of risk, coupled with short-term signals, is sufficient to ensure that the security of supply is solved or not, is a matter that the regulator should evaluate.

If, for some reason, the regulator desires to add further incentives for being available when needed in the short term, one alternative would be to tie a physical delivery obligation to the call option. This means that an option-selling generator that, when the prices are high, fails to provide the committed power has to bear an extra penalty (i.e. on top of the payment because of the financial side of the call option) for each non-delivered megawatt.

The mechanism to trade the reliability option

This reliability product could fit in the context of either a price or a quantity-based mechanism. In the latter alternative, as with any other product, it could be traded either bilaterally or in a (centralised or not) auction mechanism. Although diverse alternative schemes could be devised, the fact is that in practice this product has only been defined in the context of centralised quantity-based auctions. Indeed, when making reference to the "reliability options mechanism" it is usually understood not only the reliability option product, but also these latter features (i.e. the physical obligation and the centralised auction). In the following we will restrict the discussion to this framework, which, in a nutshell, could be described as follows:

- An auction is organised where the auctioneer has to determine, in advance, at least the following parameters:
  - the strike price, s: it should not be too low, since it acts as a price cap for demand and somehow represents the frontier between the "normal" energy prices and the "near-rationing" energy prices,
  - the time horizon: typically 1 year for existing units and longer time horizons for new entrants. The seller can be required to generate the committed capacity at any time during that period,
  - the total amount of power to be bought, Q,
  - and the value of the explicit per unit penalty, pen.
- The generators submit one or several bids to the auction, expressing quantity (the capacity they want to sell) and price (the required premium). Note that a significant additional advantage of the reliability option approach is that it eliminates the need for the regulator to calculate the "firm capacity" of each unit. This is clearly an improvement over alternative methods, especially when there are energy-limited plants involved.

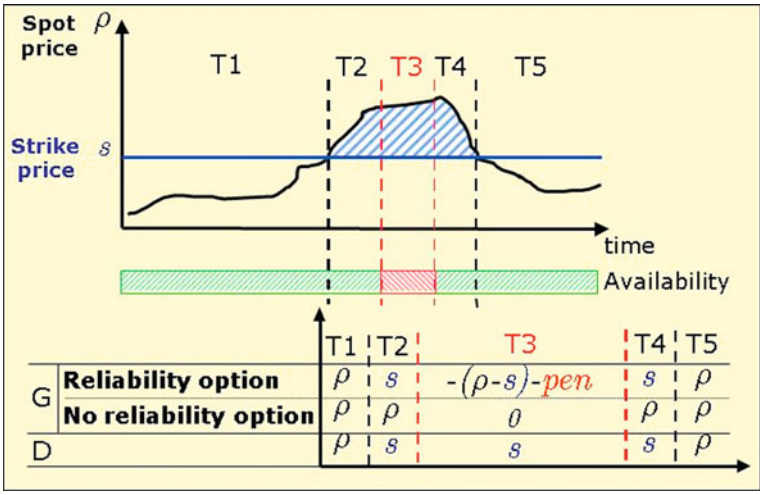


Fig. 12. 1 The reliability product

- The market is cleared as a simple auction and all of the accepted bids receive the premium that was solicited by the marginal bid.
- During the specified time horizon, any time the spot price ρ exceeds the strike price s, the bids that were accepted in the capacity auction will have to refund the regulator—and, indirectly, consumers—for the difference $(\rho - s)$ for each megawatt sold in the capacity market. Henceforth, we will refer to this refund as the "implicit penalty".
- In case the physical obligation is tied to the financial option, then if the spot price is above the strike price and the production g of a certain generator is lower than the committed capacity q, then he would have to pay to the regulator an "explicit penalty", computed as $pen \cdot (q - g)$. An example of this is illustrated in Fig. 12.1 where it is shown how much demand (D) has to pay during each time interval, as well as how much generators (G) who have signed a reliability option should receive.

References

1. All Island Project (AIP) (2007) The bidding code of practice. A response and decision paper. AIP-SEM-07-430, 30 July 2007
2. Australian Energy Regulator (AER) (2007) State of the energy market
3. Barroso LA, Bezerra B, Rosenblatt J, Guimarães A, Pereira MV (2006) Auctions of contracts and energy call options to ensure supply adequacy in the second stage of the Brazilian power sector reform. IEEE PES General Meeting 2006, Montreal, Canada
4. Batlle C, Pérez-Arriaga IJ (2008) Design criteria for implementing a capacity mechanism in deregulated electricity markets. Special issue on Capacity mechanisms in imperfect electricity markets. Util Policy 16(3):184–193. doi:10.1016/j.jup.2007.10.004

5. Batlle C, Rodilla P (2010) A critical assessment of the different approaches aimed to secure electricity generation supply. Energy Policy 38(11):7169–7179, Nov 2010, ISSN 0301-4215. doi:10.1016/j.enpol.2010.07.039
6. Batlle C, Vázquez C, Barquín J (2007a) A critical analysis of the current balancing mechanism in France. IIT working paper
7. Batlle C, Vázquez C, Rivier M, Pérez-Arriaga IJ (2007b) Enhancing power supply adequacy in Spain: migrating from capacity payments to reliability options. Energy Policy 35(9):4545–4554
8. Batlle C, Solé C, Rivier M (2008) A new security of supply mechanism for the Iberian market. Electr J 21(2):63–73. doi:10.1016/j.tej.2008.02.003
9. Batlle C, Rodilla P, Barquín J (2009) Report: policy and regulatory issues concerning security of electricity and gas supply, Florence School of Regulation Training course for senior regulatory staff, Association of Mediterranean Regulators for Electricity and Gas. Available. www.florence-school.eu
10. Batlle C, Barroso LA, Pérez-Arriaga IJ (2010) The changing role of the State in the expansion of electricity supply in Latin America. Energy Policy (2010). doi:10.1016/j.enpol.2010.07.037. Available www.upcomillas.es/batlle
11. Benini M, Cremonesi F, Gallanti M, Gelmini A, Martini R (2006) Capacity payment schemes in the Italian Electricity Market. CIGRE General Session 2006
12. Bohn RE, Caramanis MC, Schweppe FC (1984) Optimal pricing of electrical networks over space and time. Rand J Econ 15(3), Autumn 1984
13. Borenstein S, Bushnell J (2000) Electricity restructuring: deregulation or reregulation. Regulation. Cato Rev Business Gov 23(2):46–52
14. Cammesa (2005) Argentine power sector capacity payments, markets and new generation, Compañía Administradora del Mercado Eléctrico Mayorista, APEX conference Orlando, USA
15. Caramanis MC, Bohn RE, Schweppe FC (1982) Optimal spot pricing: practice and theory. IEEE Trans Power Apparatus Syst PAS-101(9)
16. Caramanis MC (1982) Investment decisions and long-term planning under electricity spot pricing, IEEE Trans Power Apparatus Syst PAS-101(12)
17. Centro Nacional de Despacho (CND) 2008 Reglas comerciales. In: Spanish. Available. www.cnd.com.pa
18. Chandley J (2005) ICAP reform proposals in New England and PJM. LECG, Report to the California ISO 2005
19. Comisión Nacional de Energía (CNE) (2005) Spanish electric power act. Unofficial English translation. Comisión Nacional de la Energía, vol. 7, 3rd edn
20. Cramtom P, Stoft S (2007) Colombia firm energy market. Proceedings of the Hawaii international conference on system sciences. Available. www.cramton.umd.edu
21. Cramton P, Stoft S (2005) A capacity market that makes sense. Electr J 18(7):43–54
22. Finon D, Meunier G, Pignon V (2008) The social efficiency of long-term capacity reserve mechanisms, Utilities policy, vol. 16, Issue 3, Capacity mechanisms in imperfect electricity markets, pp 202–214
23. Finon D, Pignon V (2008) Capacity mechanisms in imperfect electricity markets. Editorial of the special issue on capacity mechanisms in imperfect electricity markets. Utilities Policy 16(3):141–142
24. Green R (2004) Did English generators play Cournot? Capacity withholding in the Electricity Pool. CMI Working Paper 41, March 2004. http://www.econ.cam.ac.uk/electricity
25. Huber ER, Espinoza VR, Palma-Behnke R (2006) Hydrothermal coordination and capacity payment schemes in chile: current discussion and future challenges. Working paper, Mimeo
26. Hogan W (2005) On an 'Energy-Only' Electricity market design for resource adequacy, paper prepared for the California ISO
27. ISO New England (2006) Market rule 1 standard market design, Section III, ISO New England, Inc. FERC Electric Tariff No. 3

28. Joskow PL (2006) Introduction to electricity sector liberalization: lessons learned from cross-country studies. In: Sioshansi F, Pfaffenberger W (eds) Electricity market reform: an international. Perspective 1–32:2006
29. Joskow PL (2007) Competitive electricity markets and investment in new generating capacity. In: Helm D (ed) The new energy paradigm, Oxford University Press, Oxford
30. Ministerio de Economía, Fomento y Reconstrucción (2008) N° 39.048, Diario Oficial de la República de Chile
31. Ministry of Economic Development (MED) (2003) Electricity supply security: questions and answers. 20 May 2003. Available. www.med.govt.nz
32. MITyC Ministry of Industry, Tourism and Trade (2007) Orden ITC/2794/2007 de 27 Septiembre (Order ITC/2794/2007, of Sept. 27). In Spanish
33. National Grid Electricity Transmission (NGET) (2010) The grid code. Issue 4 Revision 2. 22 Mar 2010. Available. www.nationalgrid.com
34. Neuhoff K, De Vries L (2004) Insufficient incentives for investment in electricity generations. Utilities Policy Elsevier 12(4):253–267
35. Newbery DM (1998) Pool Reform and Competition in Electricity chap. 5. In: M. Beesley (ed.) Regulating Utilities: Understanding the Issues, London Institute of Economic Afairs, pp. 117–166
36. Nome (2010) Projet de loi de nouvelle organisation du marché électrique. In: French, New organization of the electricity market bill. Available. www.energie2007.fr/actualites/fiche/2538
37. North American Electric Reliability Council (1997) NERC planning standards
38. Pérez-Arriaga IJ (1994) Principios económicos marginalistas en los sistemas de energía eléctrica (In Spanish). Technical report IIT-93-044
39. Pérez-Arriaga IJ (2001) Long-term reliability of generation in competitive wholesale markets: a critical review of issues and alternative options. IIT working paper IIT-00-098IT
40. Pérez-Arriaga IJ, Batlle C, Rivier M (2006) Diagnosis of the White Paper for the reform of the regulatory scheme of the power generation in Spain. IIT working paper IIT-07-003. Available. www.iit.upcomillas.es/batlle
41. PJM (2008) PJM manual 18: PJM capacity market, PJM forward market operations
42. Rodilla P, Batlle C (2010) Security of electricity supply at the generation level: problem analysis. Working paper IIT-10-027A
43. Roques F, Newbery DM, Nuttall WJ (2005) Investment incentives and electricity market design: the British experience. Rev Netw Econ 4(2)
44. Ruff LE (2003) UnReDeregulating electricity: Hard times for a true believer. Seminar on new directions in regulation. Kennedy School of Government, Harvard University, Cambridge, 1 May 2003
45. Schweppe FC, Caramanis MC, Tabors RD, Bohn RE (1988) Spot pricing of electricity. ISBN 0-89838-260-2, Kluwer Academic Publishers, Boston
46. SEM (2008) Designated SEM Trading and Settlement Code v4.5, Single Electricity Market. Available at http://www.allislandmarket.com/MarketRules
47. Stoft S (2002) Power system economics: designing markets for electricity. IEEE Press & Wiley-Interscience, ISBN 0-471-15040-1, 2002
48. Urrutia VC (2008) Cobertura de la demanda de energía eléctrica en Panamá: Perspectivas y Mecanismos para asegurar el abasto suficiente a los usuarios (In Spanish). XII Reunión de la Asociación Iberoamericana de Entidades Reguladoras de Energía San Luis Potosí, México
49. Vázquez C (2003) Modelos de casación de ofertas en mercados eléctricos (in Spanish). PhD Thesis, Universidad Pontificia Comillas
50. Vázquez C, Batlle C, Rivier M, Pérez-Arriaga I J (2003). Security of supply in the Dutch electricity market: the role of reliability options, IIT working paper IIT-03-084IC, for The Office for Energy Regulation (DTe) of The Netherlands. Presented at the workshop CEPR Competition & Coordination in the Electricity Industry, Toulouse, 2004
51. Vázquez C, Rivier M, Pérez-Arriaga IJ (2002) A market approach to long-term security of supply. IEEE Trans Power Syst 17(2):349–357

52. Von der Fehr NHM, Amundsen ES, Bergman L (2005) The Nordic market: signs of stress? Energy J Special Issue:71–98
53. WCED (UN World Commission on Environment and Development) (1987) Our Common Future: Report of the World Commission on Environment and Development, WCED, Switzerland
54. Wolak F (2004) What's wrong with capacity markets. Stanford University, Mimeo

Chapter 13
Electricity and Gas

Julián Barquín

One of the most interesting regulatory challenges in the energy sector during the next decade will be to cope with the multiple dimensions of the interaction between the markets and the infrastructures of gas and electricity.

This chapter provides a very brief introduction to the natural gas industry and its regulatory structure, in which the focus is on the factors that affect the electricity industry.

While the electricity and natural gas sectors followed distinct parallel courses during most of the twentieth century, they have been gradually converging in the last 25 years. On the one hand, the use of gas as a fuel to produce electricity has risen steeply, albeit from a very low level. On the other hand, the problems arising around electric and gas industry liberalisation are often similar, mostly because both are grid industries.

Natural gas is extracted from the fields where it is deposited, transported to consumer hubs through gas pipelines or as liquefied natural gas (LNG) and finally distributed to end consumers. Taken as a whole, this supply, transmission and distribution system is known as the natural gas chain. Production and transmission to consumer countries and regions, i.e. the upstream end of the chain, is usually distinguished from transmission and distribution within consumer countries and regions, or the downstream end.

Upstream, gas and oil systems are similar. Investment in exploration and production is made primarily in non-OECD countries. In some cases, political risk is highly significant. Gas companies typically enter into agreements with local public companies, which entails taking account of factors such as royalties, local taxes, and the possible existence of State shareholdings. Exploration involves high technical and financial risk and requires a long-term approach. The medium-term horizon is more relevant for development and production and is expertise and capital intense. The upstream portion of the industry often falls outside the competence of national regulators with jurisdiction over end consumers.

J. Barquín (✉)
Institute for Technological Research, Comillas Pontifical University,
Alberto Aguilera 23, 28015 Madrid, Spain
e-mail: julian.barquin@iit.upcomillas.es

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_13,

The downstream system, in turn, is more like the electricity system. Transmission and distribution are natural monopolies. Significant technical constraints (balancing, transmission constraints, quality standards…) must be handled by a System Operator. The business involves a number of very expensive and industry-specific infrastructures, designed to accommodate security of supply issues, among others. Moreover, in both gas and electricity, procurement and supply (i.e., retailing, not a network activity) to end consumers is an activity that can potentially be conducted on a competitive market. Finally, the national regulator for gas is often the same body that supervises the electricity industry.

The first section of this chapter describes the basic technical structure of the gas industry, including the nature of the activity, production, transmission, storage, distribution and consumption, as well as the *modus operandi* in each step along the way. The second addresses industry regulation, focusing on the downstream end of the business, which is the most relevant from the standpoint of electricity industry actors. The third section deals with security of supply, as regards the gas industry per se and its impact on the electricity industry. The last section discusses market power problems stemming from the existence of large companies that conduct business in both industries.

13.1 Technological Aspects of the Natural Gas Sector[1]

Natural gas is a mix of methane and other gaseous hydrocarbons such as ethane, propane and butane. It also contains nitrogen, carbon dioxide and water vapour. Geologically speaking, the origin of gas and oil is similar and they are often found in the same fields. In such cases, the gas is known as wellhead, oil well or associated gas. Non-associated gas is the gas deposited in fields containing gas only.

The nature of a given hydrocarbon depends primarily on the proportion of hydrogen to carbon atoms in its molecule. Hydrocarbons with a high proportion of hydrogen have very low melting and boiling points, lower densities, less combustion energy per unit of volume and more combustion energy per unit of weight, than materials with a lower hydrogen to carbon ratio. Methane (CH_4) and ethane (C_2H_6), for instance, are gases at ambient temperature, while propane (C_3H_8) can be readily liquefied at ambient temperature by raising the pressure. The boiling point for butane (C_4H_{10}) at atmospheric pressure is 1 °C below zero. Since it can be liquefied at ambient temperature by raising the pressure, it can be transported in bottles. All hydrocarbons with five (pentane) or more carbon atoms are always liquid under normal conditions. Pentanes and heavier liquids are known as condensates or natural gasolines, whereas ethane and heavier liquids (including condensates) are natural liquid gases.

[1] For general reference of technical aspects, see [11].

Natural gas is regarded as wet when it contains significant amounts of natural liquid gases (NLGs) and dry otherwise. The NLG content may vary widely from one field to another, from nearly nil to up to 30 %. Hydrogen sulfide, another impurity in natural gas, must be separated because of the corrosion it induces. Gas with high hydrogen sulfide content is called sour, while sweet gas has a low proportion of H_2S.

13.1.1 Reserves and Resources[2]

Prospecting for natural gas is a resource-intensive activity that calls for considerable know-how. Due to technological limitations, however, not all the gas found can be extracted. A distinction is therefore drawn between resources or the total amount of gas in the field and reserves, which is the amount that can be economically extracted.

Reserves depend on both technology and the market price for gas at any given time. Moreover, both resources and reserves are subdivided into additional categories depending on how reliable the estimate is believed to be. Reserves with a probability of recovery of 90 % or higher are proven, when the likelihood of recovery is 50 % or over they are probable, and when the certainty of recovery is 10 % or higher, possible. Nonetheless, these estimates generally entail a certain amount of discretion on the part of the geologist concerned and are often reported by companies or governments with an agenda.

Natural gas reserves are highly uncertain, but much more abundant than oil reserves. The ratio between reserves and output has held at around 60 years over the last 10, with a slightly upward trend, because reserves have risen more rapidly than the amount of gas extracted. The largest reserves are found in Russia (around 24 % of the world-wide total), Iran (16 %) and Qatar (14 %).

The quantities quoted above are for conventional gas. In addition, there are also huge amounts of unconventional gas: shale gas, coal bed methane, tight gas (from low permeability reservoirs) and gas (or methane) hydrates. Total and recoverable volumes are very uncertain, but the IEA estimates that, excluding gas hydrates, they might amount about double of those of conventional gas.[3] Generally speaking, unconventional gas is more evenly distributed than conventional gas.

---

[2] Global information on natural gas reserves, resources, transportation and consumption can be found in the IEA report *World Energy Outlook*. The report is updated every year and can be downloaded from the IEA website www.iea.org.

[3] IEA, World Energy Outlook 2010.

13.1.2 Production[4]

The decision to exploit a field depends on whether the gas is associated or otherwise. If it is not, exploitation depends strictly on profitability considerations. Wellhead gas, however, has to be extracted to bring the oil to the surface. If no gas pipeline is available to transport it, it must be flared or reinjected. The advantage of reinjection is that it enhances oil recovery (by maintaining the field pressure) and allows for future recovery of the gas. In the short term, however, it raises drilling costs.

Non-associated gas may contain over 85 % methane. Under these conditions, it may be injected directly into a gas pipeline. By contrast, where the natural liquid gas content is significant, it must be separated before injecting the natural gas into the pipeline, given its higher economic value and because its possible condensation inside the pipes could hinder transmission.

The location of world-wide production appears, a priori, to be illogical. Output tends to be higher where reserves are lower and extraction most expensive: in Siberia (under particularly severe meteorological conditions), North America (often with high production costs, in light of the small relative size of many fields) and the North Sea (offshore production in an unfavourable climate). One of the main reasons is that gas is difficult to transport to consumer hubs. Nonetheless, the declining cost of shipping liquefied gas on LNG tankers is contributing to the development of fields that have traditionally been only scantly exploited, or not at all, such as in the Near East.

13.1.3 Transmission

Gas transport, unlike oil transport, is complex and costly. For that reason, it is a characteristic of countries sufficiently developed to have invested the capital needed to finance gas transmission and distribution grids. The two main transport media are gas pipelines and LNG tankers.

Pipelines

Pipelines are steel pipes, normally 36–142 cm in diameter, that carry gas at pressures of 80–100 bar. They constitute the primary transport medium and may be hundreds of kilometres long.

The gas moves through the pipe because of the difference in pressure at the two ends. Field pressure itself sometimes suffices to transport it for considerable distances, but normally compressors need to be installed at regular intervals (typically every 100–150 km) to raise the pressure. The energy required is often obtained by

[4] Upstream activities are not the focus of this chapter. However, as a reference of engineering aspects, see [8].

burning some of the gas carried, although electrical compressors are used in some systems. Valves may also be installed to facilitate grid operation.

A gas pipeline may be able to carry on the order of one million or more cubic metres of gas per hour, at normal pressure and temperature. This rate may be raised if the operating pressure is increased, although the trade-off is higher compressor operating costs.

LNG chain

The LNG chain, i.e. the shipment of gas from the field on tankers to markets normally thousands of miles away, comprises the following stages.

- Liquefaction trains are the most technically complex and expensive part of the process. They consist of several cooling cycles to lower the temperature of the gas to –160 °C, thereby reducing its volume 600-fold.
- The LNG is loaded onto LNG tankers, a specific type of vessels normally with a cargo capacity of 140,000 m^3 of LNG, equivalent to approximately 900 GWh and shipped to its destination, where it is unloaded.
- Regasification includes LNG tanker mooring and unloading, as well as the measurement, storage and vaporisation of the natural gas, which may alternatively be loaded onto trucks.

The energy expended in the entire process is on the order of 10 % of the energy of the gas shipped. The chain is flexible, but only to a limited degree. Liquefaction plants are designed to operate with a high and constant load factor and LNG tankers cannot store gas for long periods of time because it evaporates slowly. Consequently, the chain is designed on the assumption of a constant flow of tankers from the liquefaction to the regasification plant. Some flexibility is nonetheless possible (by re-routing a vessel from one regasifier to another, for instance), particularly in the long term, where operating plans can be amended.

As noted, these infrastructures are expensive. The cost of a gas pipeline is approximately proportional to its length. A high pressure line, for instance, may cost on the order of over 1 million dollars per kilometre, although this varies widely depending on the type of terrain involved. LNG transport, in turn, entails both the fixed costs (regardless of the shipping distance) incurred to liquefy and regasify the product and variable costs that rise moderately with volume due to the need for a larger number of tankers. A gas pipeline may be preferable for relatively short distances (1500–3000 km, depending on circumstances), while LNG is more cost-effective for longer range shipping.

As in the case of electricity transportation, the construction of the transmission network may be much postponed because of delays in administrative authorization, either because of environmental concerns or because it involves an international agreement. Construction itself uses to be much speedier (sometimes the pipeline can be built within a year) and it involves high capital intensity and high economies of scale.

13.1.4 Storage

Gas is stored to attain:

- a strategic objective, namely to ensure a reserve from which to draw if imports are interrupted.
- a technical objective, to be able to supply the demand for gas, which is characterised by wide daily and seasonal variations (the demand for hot water, for instance, rises in the morning and at night and heating may be necessary in the wintertime only).

Several types of storage can be identified.

- Linepack is storage in the transmission network. Whilst the volume is small and serves primarily as a daily balancing tool (flexibility), since all users are present, it constitutes a good trading platform.
- LNG is also stored in regasification terminals. The volume involved is larger than in linepack storage. It also constitutes good operational storage, allowing logistic users weekly/monthly flexibility and accommodating changes in demand. It may be used by the System Operator to quickly respond to disruptions in supply.
- Underground storage, which accommodates larger volumes than the other two, serves seasonal purposes, although it takes about 12 h to reverse injection/withdrawal cycles. It comprises strategic reserves and is much more inexpensive than LNG storage. A number of geological or man–made structures can be used for underground storage, including depleted gas fields, aquifers, salt caverns or mines.

13.1.5 Distribution and Consumption

Gas is distributed through pipes that operate at pressures of under 20 bar. Certain large consumers (such as electric power plants) may, however, be connected directly to the transmission grid. Like electricity grids, gas distribution networks are organised hierarchically: the high pressure distribution grid[5] (4–20 bar), which is fed by the transmission grid, feeds the medium pressure network (50 millibar to 4 bar), which in turn feeds the low pressure grids.

The end consumers of natural gas have traditionally been and in most systems continue to be manufacturers and households (particularly for domestic heating). Nonetheless, gas has been increasingly used to produce electricity over the last 20 years, particularly since the advent of combined cycle gas turbine plants.

[5] All the pressures cited are differential, i.e. the difference between the pressure of the natural gas and atmospheric pressure. Residential facilities are typically designed for differential pressures of 15 millibar.

13.1.6 Downstream Gas System Operation

As in the case of electricity, the existence of a meshed network requires the presence of a System Operator to coordinate operations and ensure system security, also known as system integrity in this context. The gas system is simpler in two respects, however.

- As gas transmission is not subject to Kirchoff's second law,[6] gas flows can be directed through specific paths. The path of a certain gas parcel may even be traced from source to destination, a possibility that makes no sense in electricity systems, from the standpoint of their physics. From the perspective of transmission planning, then, one of the most significant sources of network externalities is absent in gas systems.
- Dynamics are much slower in gas than in electricity, because significant amounts of gas can be stored in the network, typically enough to balance the system during an entire day. By contrast, the amount of electrical energy stored in the grid only suffices to "balance" the system for milliseconds, which is why the electricity network must be balanced instantaneously.

Such slower dynamics naturally render system operation easier. The trade-off is that gas System Operators typically have fewer resources from which to draw when operation goes awry.

Operational procedures

Specific operational procedures vary between systems, although they all have certain similarities. The following description is based on the procedures followed in Spain,[7] where system operation is organised further to a "process chain" consisting of several steps.

- The first is programming. All agents using gas system facilities are required to submit a programme to the System Operator and the operators of the facilities they intend to use. They must inform the amount of estimated gas input, output, supply or storage in a given period. Programmes, which are usually merely informative, are drawn up for different time frames: yearly, monthly and weekly. The shorter the time frame (and

[6] Formally, the role played by pressure in gas transmission is similar to the voltage angle function in electric power grids. Valves and bypasses are simple and reliable devices, however, for which there is no inexpensive equivalent in electricity.

[7] The specific procedures can be downloaded from the SO website www.enagas.es (search "Procedures" under "Technical Management of the System").

the closer to injection/withdrawal), the more detailed and realistic is programming.
- Nomination, the second stage, is also required of all agents using gas system facilities but, as opposed to programming, is binding. Agents notify the System Operator and facility operators of the estimated gas input, output, supply or storage during a given day, broken down by gas system injection or withdrawal point. Notifications can call for more capacity than previously contracted by the agent with the facility operator, subject to availability. They may also be rejected, e.g. if specific requirements required for short-term storage are not met.
- Measurement, sharing and balancing are standardised procedures used to measure volumes and qualities, establish each agent's share in the gas transported, regasified, distributed or stored and physically balance the various facilities.

Throughout the chain, agents must maintain a balance; i.e. the amount of gas injected into the system must equal the amount withdrawn plus the inventory difference. Users' inventories must be below their maximum assigned capacity, defined to be the contracted capacity plus any amount allocated by the SO based on a regulated procedure. Otherwise they must buy or sell gas, modify their programming or notification, execute supply interruption clauses, negotiate supply interruptions, use underground storage or modify consumption. Failures to comply are fined or otherwise penalised.

13.2 Structure and Regulation of the Downstream Natural Gas Industry

Upstream structure and regulation in the gas and oil industries are similar and often involve the same actors (companies and regulators). Unlike the oil market, however, the natural gas market is not global, due to its high transport costs.

The three chief natural gas markets are found in North America (essentially US and Canada), Europe and Asia (Japan, South Korea and Taiwan). Certain national systems, such as in Russia, Brazil and China, have dynamics of their own. While mean yearly prices tend to move in the same direction in all these markets, in a shorter time frame trends may vary from one area to the next. The conditioning factors also vary widely: North America produces most of what it consumes, Europe depends heavily on gas pipeline imports from Russia and northern Africa, and the East Asian countries import their gas in liquefied form from Indonesia, Australia or the Persian Gulf. The convergence among these markets due to the falling costs of LNG tanker shipping has led some observers to predict that a single global price will prevail in the long term.

A distinct US phenomenon has been the huge increase in shale gas production that presently amounts to more than 20 % of total US gas production.[8] As a consequence, gas prices are now low in the US (especially when compared with past years' expectations), gas has displaced coal to a very significant degree for the US electricity production, and coal prices have collapsed in the US and elsewhere.

As in electricity, the downstream sector of the natural gas business has traditionally been regarded as a natural monopoly. The reasons are similar in the two sub-industries: economies of scale, capital-intensiveness and the geographic specificity of assets, to name a few. Certain particulars have led to regulations with industry-specific characteristics, however. As in the electricity industry, in some parts of the gas business, which are being liberalised or de-regulated, competition is being furthered, whilst others continue to be regarded as regulated monopolies.

13.2.1 The Traditional Model

The gas industry was traditionally structured around vertically integrated companies that produced gas in their own fields or purchased it on the wholesale market, built, operated and maintained the major infrastructure (regasification terminals, transmission pipelines, storage facilities and even distribution networks), and sold the gas either to local distributors or end consumers. Local distributors were often owned by towns or cities, regions or States.

In these systems, the regulator (usually a ministry, for independent bodies were seldom created for this purpose) was normally involved in long-term central planning, including energy balancing, choice of technologies and determination of the additional capacity needed. Since gas utility capital was often held by the State (region or city), the regulator was also involved in company management. Regulatory authorisation was required to conduct commercial or technical business. One of the regulator's most important tasks was to set the tariff to be paid by end users. Where a market of any description existed in the gas industry, it was restricted to bilateral agreements between producers and buyers. Long or very long (up to 30 years) wholesale supply contracts were the norm, with upstream gas prices being pegged to oil or by-product prices. Despite the use of the past tense here, the regulatory framework described is still in place in many systems.

Some of these features are also characteristic of traditional electricity regulation, which is ultimately the outcome of the fact that distribution grids constitute a natural monopoly. The rationale for others is specific to gas, however. Connecting a gas field to consumers is a capital-intensive endeavour. Therefore, upstream investors require assurance that they will be able to profitably sell the gas for a number of years. Similarly, downstream companies require guarantees that they will be able to sell the gas bought to final consumers. The traditional solution was

[8] Environmental concerns (Europe) and high extraction costs and sophisticated technology (elsewhere) have up to now mostly prevented shale gas production outside the US.

to conclude long-term agreements for a fixed volume (meeting upstream party concerns) and a price pegged to oil (or oil product) prices, which removed any end consumer incentive to switch fuels[9] (meeting downstream party concerns). Such agreements are somewhat misleadingly known as "take or pay" contracts, for what they actually stipulate is a buyer commitment to pay for a given amount of gas.

Such contracts, however, neither guaranteed a profit margin for the supplier (e.g. oil prices could decline) nor protected the buyer from demand swings (power plants might demand less than expected because of unexpectedly high hydro production, for instance). Nonetheless, they normally included price revision clauses to accommodate periodic adjustments in the pricing formula, as well as an arbitration procedure to provide for a solution where no agreement could be reached. That notwithstanding, given that suppliers and buyers were often based in different countries, long and bitter disputes have been known to ensue.

Other clauses of these agreements provided that the buyer would not resell the gas outside its own franchise zone or country, although they were allowed to resell the gas under discriminatory terms, charging residential customers substantially more than fertiliser factories, for instance.

13.2.2 The Deregulated Model

Gas deregulation is a relatively recent development and only feasible where the gas system is mature (i.e., large) enough. First, given the huge volumes involved in gas supply contracts, competition can only be sustained by very large-scale systems. Network investment by comparison is typically much smaller, and much more linear (looping[10] and other incremental upgrades may be preferred to huge investments reflecting economies of scale), reducing both the need for long-term commitments[11] and the likelihood of hold-ups.[12]

As in electricity markets, liberalisation is advisable only where competitive pressure is sufficiently strong. Under such circumstances, competition should yield

[9] Gas prices have been typically always slightly lower than oil prices (taking into account switching costs, technical efficiencies, and so on).

[10] Looping consists of building a bypass along a given section of pipeline (e.g., the first 20 km of a 100 km pipe). Because the volume for gas transportation is greater in the "looped" section, a smaller pressure differential is required to move a given quantity of gas in that length of pipeline. The resulting greater pressure differential for the rest of the line raises transmission capacity.

[11] When initially developing the system, investments are made to accommodate both the existing demand and future demand growth due to the huge economies of scale involved, but this further exacerbates the problems discussed in the preceding item.

[12] These consist of the exercise of market power created when a specific facility is needed for system operation. For instance, if the gas transmission grid is not densely inter-connected, an LNG importer may be forced to use a specific regasification facility. Although third-party access (TPA) provisions are usually in place to address these concerns, facilities with reduced or no TPA obligations may also exist. Such measures at least lower the incentive to hoard capacity (see below).

more efficient prices, higher quality and innovative products. Concerns arise, however, around the possible decline in reliability and in bargaining power with supply side oligopolies in heavily import-dependent systems.

Effective competition calls for unbundling of the businesses involved. Distribution is a natural monopoly that must be conducted by regulated companies. By contrast, wholesale gas procurement and retail gas supply are potentially competitive activities. As in electricity systems, transmission, which lies in between procurement and distribution, must be regulated and is the natural platform for wholesale gas trading.

Transmission

Wholesale market agents buy gas from producers, sell it to consumers or distributors and hire the services required to ship it from the entry to the exit points. Both the access to transport facilities (mainly pipelines and regasification plants) and the tariffs to be paid must therefore be regulated.

Access rights or transmission capacity hired by agents may be defined in three ways, broadly speaking.

- In point-to-point access, both entry and exit points are specified in advance. The right consists, for instance, of transporting 10 GWh of gas from Entry harbour to Metropolis.
- In entry/exit, entry and exit rights are granted separately, i.e. neither the origin nor the destination of the gas need to be specified. A right may be acquired to inject 10 GWh at Entryharbour, for instance, regardless of whether the gas is to be shipped to Metropolis or Gotham.
- Zonal access entails purchasing the right to inject or withdraw gas at any node inside a zone.

Ratemaking or tariff setting can be similarly classified.

- Point-to-point charges are based on the established entry and exit points and typically computed with a distance-related formula (such as a distance matrix).
- In entry/exit arrangements tariffs are computed independently for each point pursuant to a pre-established methodology.
- The zonal charge is a flat rate levied on transactions anywhere in the zone.

Different methodologies may be used for defining access rights and tariffs. In electricity transmission, connection access rights are typically defined on an entry/exit basis (the generator can deliver any amount of power to the grid, up to its rated capacity). The use of system tariff for electricity transmission is typically a flat charge (postage stamp),[13] although some systems apply charges with locational components, as explained in Chap. 6 of this book.

[13] Entry/exit tariffs are therefore analogous to electricity locational pricing (a different locational component for the electricity price in each bus). However, in gas there is nothing analogous to spot pricing for electrical energy, as gas tariffs with locational components are computed from long-term transmission infrastructure costs.

Entry/exit tariffs
Entry/exit tariffs may be computed in one of two ways [2].

- Long-range marginal cost arrangements are associated primarily with network expansion. The tariff can be computed from a transmission model that computes optimal expansion during peak hours. This is a sensible approach if expansion costs are linear or quasi-linear, which is more likely to be the case in mature systems. The resulting marginal costs can be adjusted to maintain an equal split of revenue between entry and exit or to attain a revenue target. As tariffs depend on network expansion, this system is appropriate wherever significant growth is expected, i.e. in congested systems.
- In average accounting cost schemes, the goal is to allocate the fixed costs of prior investments to system users and is more appropriate if no significant further expansion is anticipated.

In both cases, the same formal methods can be used to obtain consistent entry and exit tariffs, given measures of usage (e.g., average participations) and elements (pipelines, compressors, etc.) and costs (long-range marginal or accounting ones).

A primary market for capacity arises around the access rights or contracts sold by pipeline and regasification facility owners at regulated prices. In liberalised systems these contracts can be re-sold on secondary markets.[14] The characteristics of these secondary markets depend on the nature of the capacity rights. Entry/exit capacity booking may be regarded to favour competition, since it enables new entrants to book capacity without specifying the contractual path followed by the gas. Incumbents may have an advantage in point-to-point systems because, thanks to their large capacity portfolios, they can optimise their gas flows, therefore lowering their average transport costs. Entry/exit systems may also favour market development, since financial players should prefer anonymous trading. Be it said that despite the foregoing, the most highly developed gas market (in the US) is organised around point-to-point transport contracts. The existence, on the one hand, of regulated tariffs that set price caps only and on the other of significant pipe-to-pipe competition may constitute the critical features of that market.

[14] Although the primary rights holder may still be liable for notification and other obligations vis-à-vis the System Operator.

The US transmission market[15]

Since the entry into effect of FERC Order 636, issued in 1991, the US pipeline companies are no longer allowed to deal in the gas commodity itself. Rather, they are required to offer unbundled transmission services to other gas owners. Firm transmission capacity must be offered at a price capped by a regulated formula. Interruptible services may be also offered at a capped price.

Transmission services are purchased from pipeline companies during the so-called "bid-week", usually the third week of each month. Shippers notify the gas volumes they plan to transport in the following month, specifying the injection and withdrawal points and the volume, which is limited to the amount of their firm transmission rights. Unused firm transmission capacity reverts to the pipeline, which sells it as interruptible transmission ("use it or lose it" clause).

The US market is characterised by competition among pipeline companies that offer alternative routes between two markets. At the same time, other companies offer storage services enabling actors to compete for different time slots. As a result, negotiated tariffs are often lower than the regulatory ceilings.

Hubs

Hubs are platforms for wholesale gas trading. They may be divided into physical hubs, typically placed where several pipelines meet and are directly connected to storage facilities, and notional hubs, also known as virtual trading points. Examples of the former are Henry Hub in the US and Zeebrugge in Belgium, and of the latter the National Balancing Point (NBP) in the UK and the Title Transfer Facility (TTF) in The Netherlands.

Transition to a deregulated system

Most physical assets in downstream gas systems are regulated facilities. The role played by generation plants in electricity systems is played by long-term procurement contracts in gas markets. Like the former, long-term contracts are huge long-term investments whose recovery, planned under a regulatory regime, undergoes dramatic change when markets are liberalised. Unsurprisingly, these contracts have generated a good deal of controversy.

Discriminatory clauses are much more difficult, not to say impossible, to enforce in a liberalised market: trading activities tend to equalise prices. Existing contracts therefore come under stress and may be re-negotiated. At the same time, for competition to be effective, a suitable number of agents must supply gas to the system. Since the volumes provided in long-term contracts are likely to account for

[15] See, for instance, [22].

a large share of future needs, gas release programmes may be engineered to oblige incumbent companies to sell part of their contracts to new entrants.

These developments are likely to lead to incumbent company downsizing, although other responses include expansion into foreign markets, the electricity industry or upstream activities.

Henry Hub and the National Balancing Point

US physical hubs were greatly expanded in the wake of FERC Order 636 of July 1991 and the subsequent unbundling of the gas transmission system. As pipelines no longer offered some of the services required by transmission system users, such as storage or balancing, new companies arose to meet these needs. Parallel administrative services and trading platforms became available. About 25 hubs are presently in operation in the US.

Foremost among these junctions is Henry Hub, located on the Louisiana coast that interconnects 14 pipeline systems. That, in addition to its proximity to a large salt-dome storage cavern facility means that huge volumes of gas are physically exchanged. It also owes its notoriety to being the pricing point for natural gas futures contracts on the New York Mercantile Exchange (NYMEX). These derivative contracts enable parties to hedge against price changes at Henry Hub and, given the high price correlation among all the US hubs thanks to a competitive transportation market, in the US as a whole.

The National Balancing Point was created in 1996 as a virtual hub operated by National Grid, the System Operator. Under British regulation, all gas injected into or withdrawn from the transmission system is assumed to pass through it. Trades are not required to be balanced. If a shipper is unbalanced at the end of the day; however, it is required to buy or sell the required amount to balance its position. National Grid is also responsible for keeping the system as a whole balanced by trading on the NBP. Trades are anonymously placed on an electronic platform operated by APX-ENDEX. The NBP price on the International Petroleum Exchange of London is the underlying value for futures and other derivatives.

13.2.3 Interactions with the Electricity System

In a number of systems, recent gas demand growth has been mainly due to increasing penetration of gas-fired power plants. As a consequence, gas and electricity systems have become interlocked and subject to new stresses because of the unusual requirements that each one imposes on the other one.

From the point of view of the gas-fired power plants, the constraints imposed both by gas supply contracts (e.g. “take or pay” clauses) and gas network access (e.g., nomination requirements) are unknown for more traditional thermal generators.

Efforts have been done in order to model both kinds of effects.[16] Gas network constraints can be the source of additional externalities to the operation of gas-fired power plants.[17]

From the point of view of the gas system, demand for electricity production is both volatile and difficult to forecast when compared to the more predictable traditional residential and industrial demands. Actually, special operation requirements are sometimes imposed on gas-fired power plants because of this reason.[18] In any case, electricity generation tends to require more flexibility of the gas system than average. However, flexibility is costly, because it requires additional transportation capacity to provide the needed operational margin. On the other hand, gas system design regulations are traditionally focused in a situation in which the infrastructure is used almost ever close to its maximum capacity.[19] The appearance of new large shale gas fields in locations with low gas demand—presently in the US and perhaps in other parts of the world—will require a tight coordination of electricity and gas transmission network planning.

13.3 Security of Supply

13.3.1 Natural Gas Security of Supply

While gas security has been the object of growing concern, often associated with geopolitical issues, an analysis of actual supply disruptions leads to a rather less troubling view. Further to Stern [16] and [17] incidents can be classified into source, transit and facility events, depending on where the cause lies. Source and transit incidents tend to draw more public attention. Examples are the cut-off of Algerian gas to Italy after the explosion of a device on the Trans-Mediterranean

[16] Example [3], where an electric utility company that owns some gas-fired power plants decides its optimal supply portfolio of different natural gas products considering its risk preferences; or the extension of the previous work to the decisions related with the gas nomination made in [23]. The drawback of "take or pay" clauses that may origin an excess of natural gas in a centralised hydrothermal dispatch has been discussed in [18], in which natural gas flexible contracts for industrial natural gas consumers are introduced. In the short term, [7] propose an optimisation model in order to solve jointly the unit commitment of thermal power plants and the flows in the gas network.

[17] Example if the output of a gas-fired power plant is limited, because the gas network has no capacity to deliver all the required gas due to functioning of other gas-fired power plant feeding from the same gas system.

[18] Example in several national regulations gas-fired power plants are required to submit nominations for each hour instead of for each day as customary for other consumptions.

[19] An in-depth analysis of this question is made in [6]. An open question is how to provide to the different agents with the incentives that lead to a socially optimal gas and electricity expansion. In particular, gas agents should provide the optimal flexibility and electricity generation agents should pay for the full cost that impose on the gas system.

Pipeline (a "terrorist" incident) and the recurring crises in connection with the transport of Russian gas across Ukraine and Belarus.

Facility incidents have also been known to occur, however. The liquid contamination at the UK Interconnector pipeline in 2002, the fire at the Algerian Skidka liquefaction plant in 2004 and the fire at the Rough storage facility in the UK in 2006 are a few examples.[20] "Engineering" risks may be contended to be especially high in stressed and ageing gas systems.

Reliability analysis is more developed in electricity than gas systems, from both the academic and the regulatory standpoints. Nonetheless, similar simulation techniques can be applied to both, see e.g., [14]. In addition to the specific technical characteristics of the models used, the assumptions made and results required must be carefully defined. The specific questions that should be addressed are listed below.

- How is reliability to be valued? Do indexes such as the loss of load probability or energy not supplied suffice, should priority be given to economic indicators such as the expected loss of social welfare, or should both be taken into consideration? Where a highly developed market is in place, the loss of welfare attached to gas supply interruptions might be estimated from market data,[21] although this is not usually the case and specific methodologies must be deployed.
- What events should be considered? No model can possibly cover all the scenarios leading to security of supply incidents. Rather, a list of incidents must be drawn up a priori. Not only engineering-related events (such as pipeline failure), but also geopolitical incidents (such as disruption of supply due to transit disputes) can be modelled, often in a similar fashion. The difficulty lies in determining the respective probabilities, although the existing operation research techniques can be used for the systematic analysis of subjective probabilities.
- What measures can be implemented? Models are generally used to compare the merits of different strategies, making this a critical issue, as discussed below.

Improving security of supply

Security of supply can be improved in a number of ways, including the non-exhaustive list of measures given below by way of illustration.

- Construction of additional infrastructure is one such measure. Redundancy in gas systems is typically lower than in electricity systems, particularly as regards transmission. Certain types of infrastructure, such as storage and regasification facilities, may impact reliability heavily, however. The availability of sufficient storage capacity is critical to deal with disruptions in gas supply. Concerns about

[20] The incident would have had more dire consequences than the price spikes observed if it had occurred earlier in the winter.

[21] From the prices and volumes specified in supply contracts with interruptible clauses or the risk premiums attached to forward contracts, for instance.

over-dependence on a single source of supply can be eased considerably by the installation of regasification facilities.

- Another measure is enhancement of demand-side response. Even moderate response can lead to significant decreases in energy not supplied. This objective may be attained by requiring or incentivising dual-fuel capabilities in gas-fired electricity plants and other large industrial facilities, instituting especially tailored tariff systems, or applying real pricing and other "smart gas" applications.
- In security-driven system operation, storage facilities can be operated very conservatively, keeping as much gas as possible available for possible contingencies. As in water management in electricity systems, however, a trade-off exists between security and economic efficiency.
- Capacity mechanisms consist of imposing or incentivising contracts for greater amounts of deliverable gas than is expected to be needed. As when building additional generation capacity in electricity systems, this approach raises a "missing money" issue that must be dealt with. Capacity mechanisms may be implemented in several ways, through shippers or the System Operator, for instance, for domestic only or total demand.
- Balancing requirements may need to be fine-tuned if shippers contract less gas than required to optimally meet demand, because they perceive that the cost of not being balanced is less than the marginal outage cost. This may occur if the cost of imbalance is inappropriately set by the authorities or if shippers perceive that in the event of a contingency they will not bear the full cost because of politically motivated action taken by the authorities.
- Lastly, inter-system connection may be enhanced. Unlike electricity system interconnectivity, interconnection between gas systems is seldom if ever based on reliability considerations. A larger interconnected system is intrinsically safer than its separate parts, however. Measures such as providing for bidirectional gas flow in pipelines or un-impeded access to storage capacity in neighbouring systems can improve security.

Market regulation and security of supply

A number of legal and regulatory considerations must be addressed to attain an adequate level of security of supply. One possibility is to allow markets to decide on the appropriate level with minimum regulatory intervention. This is the system in place in the US.

The US market is characterised by intense spot trading at many hubs, a robust forward market that has dispensed with most traditional long-term contracts and a competitive gas transport market. In that country, most gas is pipeline (as opposed to LNG) gas, see [9]. Pipeline construction has grown to meet needs since the advent of deregulation in 1985.

Adequate pricing facilitates security of supply. The graph below shows gas prices at Henry Hub since 2004. Note the peaks in the second half of 2005, associated with the disruptions caused by hurricanes Katrina and Rita. Parallel movements can also be observed in the forward price curves and in the

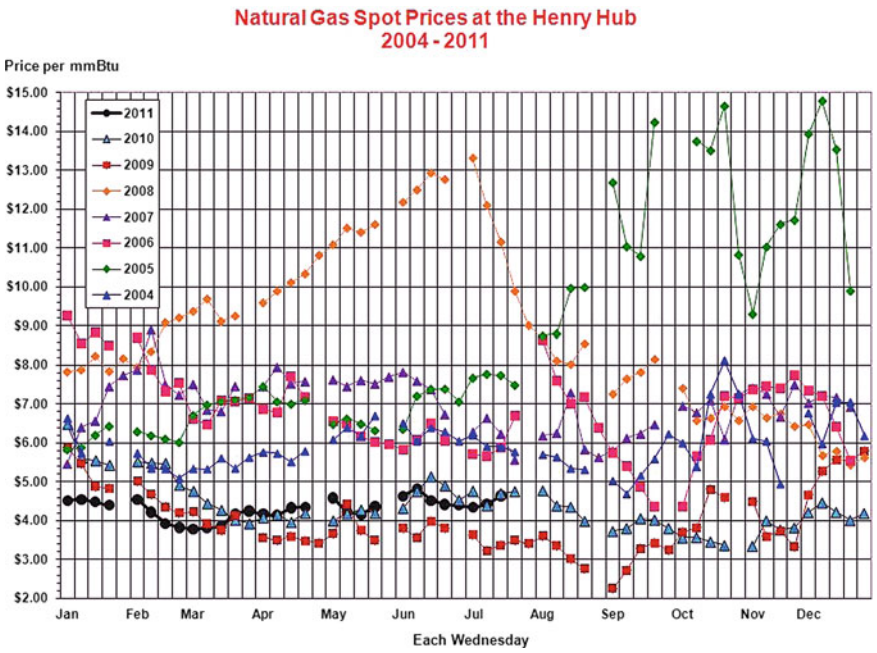


Fig. 13.1 Natural gas spot prices at Henry Hub (www.neo.ne.gov)

differentials between hubs. These price signals incentivise changes in consumption patterns, e.g. gas use by gas-fired electricity plants (Fig. 13.1).

The first factor to be borne in mind with regard to the construction of new transmission capacity is that pipeline companies must be fully unbundled from the shipping and distribution businesses.[22] Long-distance (inter-state) transmission is regulated by a single authority (the FERC). Primary capacity is remunerated under cost-of-service arrangements, while secondary capacity trades are liberalised. New capacity projects must show that they are able to support their own regulated costs (by submitting a portfolio of letters of intent from committed shippers, for instance, that therefore acquire long-term capacity rights).

The reliance in the US on light-handed regulated markets is unique. Physically, Europe's gas system is almost as large and complex as the US's. European gas networks are mainly regulated by each National Regulatory Agency (NRA); however, unbundling is much less thorough than in the US[23] and spot and forward trading is considerably less intense.

[22] This makes it difficult for any one shipper to monopolise a given transmission route because a well informed market unveils such attempts and the pipeline company is both entitled and has an incentive to sell unused capacity in secondary markets.

[23] Even if the Third Energy Package provisions are fully enforced.

Decision making on transmission facility construction is incumbent upon national or even sub-national Transmission System Operators (TSOs) and approved by the NRAs. The resulting cost is added to the regulated assets base and passed on to users as an access charge. European regulations require TPA provisions. One concern expressed around this sort of regulation is that it may potentially result in certain users subsidising network expansion needs created by others (e.g. a new regasification facility), particularly in the absence of an effective zonal price system.[24]

Facilities straddling several jurisdictions require an agreement among the TSOs and NRAs involved, a development that has been historically slow in materialising. Most of the pipelines used for third country (mainly Russia and Algeria) provision of gas in Europe were built by vertically integrated utilities and are subject to long-term contracts. The European Commission and many NRAs have consistently called for revision of these contracts ("gas release programmes") on the grounds of concerns about market foreclosure. The Commission has also encouraged more comprehensive unbundling of pipeline networks. Incumbent companies have systematically contended that such a measure would weaken their bargaining power *vis à vis* large foreign producers and ultimately compromise security of supply [19].

LNG facilities, whether built by TSOs or commercial companies, are initially subject to TPA provisions. The frequent exemption from TPA obligations granted by the commission, however, narrows the difference in status between these facilities and their US counterparts. The commission has considered the impact on market competitiveness and security of supply when granting such authorisation. Large numbers of LNG facilities have been built in Southern Europe (Spain in particular), often driven by electricity companies.

Japan depends wholly on regasification facilities for its supply. The Japanese Government (through its MITI and Japan's Export–Import Bank) has orchestrated the financing of gas trains see [1, 21]. The infrastructure consists of a cluster of LNG terminals that supply a number of relatively isolated markets. Therefore, each local monopolist is protected from competition and can invest under cost-of-service arrangements. The unavailability of nuclear electricity is seen as a relevant concern and national gas infrastructure expansion including storage and pipeline capacity as possible countermeasures [12].

Government involvement is even greater in the rest of the world,[25] as most gas systems outside the US and the EU are heavily regulated.

Import security

Traditional wisdom regards domestic supply as "secure" and imports as "unsecure". Nonetheless, international gas trade has been growing continuously despite certain incidents. In Europe, gas trade has survived both the collapse of the Soviet

[24] Such a system should, moreover, be compatible with the electricity pricing scheme to prevent investment decisions from being distorted.

[25] For a review, see [21].

Union and Islamic unrest in Algeria unscathed. More dramatic political upheavals in the near future cannot be readily envisaged.

Governments should nonetheless pursue a policy of import diversification. Discussion has been particularly intense in the European context that is described in the following.[26]

In the EU, the overall import mix for gas and oil differ very little, although the former is somewhat less diversified.[27] Unlike the global oil market, however, gas markets are still highly segmented along national borders. Eastern European Members States, for instance, are almost wholly dependent on Russian gas, while the Iberian countries depend heavily on Algeria for their supply. Gas from Russia has become a very divisive issue in the EU, particularly because the largest consumers are Germany and Italy, countries with a much wider diversity of supply than the smaller but highly dependent economies in Eastern Europe [13].

Diversifying the source of gas, whether by building new pipelines to tap resources in the Near East (the Nabucco project) or new LNG facilities, will increase security of supply in Europe.[28] But internal action, which may be more cost-effective, should not be overlooked. New intra-European transmission capacity (via the intensification of bidirectional gas flows such as in the Eastern EU or allowing gas transit from Iberian LNG terminals to Central Europe) should ease security concerns.[29] Strategic gas storage guarantees between neighbouring Member States might be another effective measure in this regard,[30] along with fairly simple action, such as coordinating EU government and regulator emergency plans.

To the extent that the costs of these and other measures are to be shared by the parties concerned in inconspicuous ways,[31] implementation is very challenging and arguably requires the supervision of a pan-European agency.

---

[26] Although diversification is sought by most governments in import dependent countries. For the Japanese case, see [10].

[27] Eurostat, *Panorama of Energy,* 2009.

[28] The ability to change supplier is an additional advantage in LNG facilities. LNG shipping is not nearly as flexible as oil shipping, however. Harmonisation of technical standards among European regasification facilities might be an effective strategy for building the EU's internal energy market and enhancing security of supply (by enabling tankers to berth in as many terminals as possible).

[29] It might also contribute to greater gas market competition, for European consumers would have access to a larger number of shippers and importers.

[30] In other words, the host Member State should allow this gas to be shipped to the State storing it, irrespective of any security concern on the part of the host Government.

[31] Pipeline cost allocation proportional to the length of the pipeline in each Member State is unlikely to reflect the benefits and incentives deriving from such a facility for the users in each State.

13.3.2 Gas and Electricity Security of Supply

Gas now constitutes a significant and even a predominant part of the generation mix in a number of electricity systems. The need for reliable gas plants may be reinforced in systems where intermittent energy penetration is high, for in such systems gas-fired steam plants, which constitute a more reliable technology, can stabilise most of the fluctuation. In light of this, electric system reliability depends not only on electric system components, but also on the reliability of the gas system.

Natural gas plants do not store large amounts of fuel, but are fed by high pressure gas pipelines. Consequently, incidents in these gas pipelines or in general in the gas grid or system severe enough to obstruct the supply determine unit shut-down. Such incidents may affect more than one generating set or unit. Unavailability in a regasification plant, for instance, affects all the gas plants fed by it. Such failures may naturally impact the system very significantly.

Where the unavailability is purely technical, its duration depends on the respective repair time. In some cases, the situation may last for several weeks. Interruptions in supply may be the outcome of other issues, however, such as problems in transit countries or simply a colder than normal winter, leading to higher than expected residential consumption and consequently less gas available for generating electricity.

From the regulatory standpoint, these should be among the issues dealt with in the mechanism in force to guarantee security of supply in both the electricity system (to correctly assess the contribution made by gas and other technologies) and the gas system (to correctly assess gas plant security of supply). No procedures to confront this problem are presently in place. Nonetheless, electricity system operators in various areas of the world are beginning to factor gas system reliability into their analyses [5]. The mechanism to handle system-wide incidents may be particularly difficult to design, however, for while the event is even less likely than individual plant failure, its impact is much greater.

In the long term, electricity and gas grid design should be coordinated, in part for purely economic reasons. Determining whether gas or electricity transmission is more suitable is seldom a clear-cut issue. This, incidentally, means that the respective transmission tariffs should be developed in a coordinated manner to provide consistent incentives to generation investors. The other reasons for coordinating electricity and gas networks are more closely associated with security of supply, which is normally one of the main reasons for expanding the system. Computer models have been developed to plan such joint expansion [20], although they are not yet being routinely used by regulators, operators or transporters.

13.4 Multi-Commodities Utilities and Market Power

In the wake of liberalisation, a sizeable number of energy companies decided to expand into industries traditionally unrelated to their line of business. The strategy consisted of using their privileged relationship with their electricity or gas customers to offer them gas, electricity, water, or telecommunications, television and internet services. The contention was that the economies of scale in customer management that these multi-utility companies could reach would afford a significant competitive advantage [4].

Such predictions have not always held, however.[32] Part of the reason may lie in the need to master highly specialised technical businesses. One significant exception has been observed, however: so-called gas-electricity convergence. No small number of former electricity companies have successfully entered the gas market, and vice versa. The explanation may lie in a number of circumstances.

- The type of electricity generation plant favoured most by investors in recent years is based on natural gas combustion. Natural gas combined cycle plants have been particularly popular, along with open cycle and co-generation or CHP facilities. Their profitability depends on the conditions of gas supply.
- Insofar as a substantial part of electricity generation is powered by natural gas, opportunities for arbitration between the two markets may arise, although they call for an in-depth knowledge of both.
- Industrial consumers may be offered more comprehensive service, particularly for flexible heat units or co-generation units.
- The same authority generally regulates the two markets and usually broaches issues in comparable ways. Moreover, the regulation of gas and electricity grid expansion should be coordinated. All these factors imply regulatory synergies for companies engaging in both businesses.

With the inter-relations resulting from the convergence between the electricity and gas industries, market power problems in one may spill over into the other. Consequently, market power analyses must address not only horizontal concentration in each or vertical concentration between wholesale and retail markets, but also the "diagonal" relationships between gas and electricity. Problems only appear, however, if the dual company holds a predominant position in one of the two industries. Some of the situations that may arise are listed below (see [15]).

- Input foreclosure arises when the dual company uses its predominant position in the gas market to hamper its electricity industry competitors' access to gas. The basic idea is that by raising the gas price it might induce an electricity price

[32] Significantly, a considerable number of European electricity companies (VIAG, VEBA, RWE, Scottish Electricity, United Electricity, Endesa, ENEL) entered the telecommunications market, while disappointing results have since determined their exit in most cases.

increase and possibly a deterioration of the competitive position of its electricity competitors from which it profits as an electricity company.

- Customer foreclosure is the use by the dual company of its predominant position on the electric power market to buy gas only from its gas division, limiting competition on that market.
- Conglomerate effects on the retail market are the result of electricity or gas distribution companies' position of privilege with respect to their customers, which may often exclude other retailers. Their most powerful competitor, and consequently their greatest incentive to keep prices competitive, is the rival gas or electricity distributor. Therefore, it is arguable that companies should be prevented from simultaneously being gas and electricity distributors in any given franchise.

References

1. Abdulkarim R (2009) Natural gas security of supply in Japan. will Qater's LNG be the solution? Shingetsu Electron J Japanese-Islamic Relat, vol 5, March 2009
2. Alonso A, Olmos L and Serrano M (2010) Application of an entry-exit tariff model to the gas transport system in Spain. Energy Policy 38(9) 5133–5140
3. Chen H and Baldick R (2007) Optimizing short-term natural gas supply portfolio for electric utility companies. IEEE Trans Power Syst, 22(1) 232–239
4. Chevalier JM (2004) Les Grandes Batailles de l'Energie. Gallimard, Folio Actuel
5. Fedora PA (2004) Reliability review of North American gas/electric system interdependency, Proceedings of the 37th Hawaii international conference on system sciences
6. Hallack MCM (2011) Economic regulation of offer and demand of flexibility in gas networks. Ph D Thesis. Université Paris-Sud 11
7. Liu C, Shahidehpour M, Fu Y and Li Z (2009) Security-constrained unit commitment with natural gas transmission network constraints, IEEE Trans Power Syst, 24(3) 1523–1536
8. Lyons WC and Plisga GJ (2005) Standard Handbook of Petroleum and Gas Engineering, Elsevier, USA
9. Makholm J (2010) Seeking competition and supply security in natural gas: the US experience and European challenge. In: Lévêque F et al (eds) Security of energy supply in Europe, Edward Elgard, UK
10. Mehden F von der and Lewis SW (2004) Liquefied natural gas from Indonesia: the Arun project, Program on energy and sustainable development, working paper no. 25, Stanford University
11. Mokhatab S, Poe WA, Speight J (2006) Handbook of natural gas transmission and processing. Elsevier, The Netherlands
12. Morikawa T (2006) Changes at LNG chain and challenges for Japan. Institute of energy economics. Available at http://eneken.ieej.or.jp
13. Noël, P. (2008) Beyond dependence: how to deal with Russian gas, Eur Counc Foreign Relat
14. Oxera (2007) An assessment of the potential measures to improve gas security of supply. Report prepared for the department of trade and industry
15. Rey P, and Tirole J (2007) A primer on foreclosure. In: Armstrong M and Porter R (eds) Handbook of industrial organization, Vol III, North-Holland, Amsterdam

16. Stern J (2002) Security of European natural gas supplies. Royal institute of international affairs. Available at www.chathamhouse.org
17. Stern J (2006) The new security environment for European gas: worsening geopolitics and increasing global competition for LNG. Available at www.oxfordenergy.org
18. Street A, Barroso LA, Chabar R, Mendes ATS, Pereira MV (2008) Pricing flexible natural gas supply contracts under uncertainty in hydrothermal markets. IEEE Trans Power Syst 23(3):1009–1017
19. Tönjes C and de Jong J (2007) Perspectives on security of supply in European natural gas markets. Clingendael international energy programme, Working paper, August 2007
20. Unsihuay-Vila C, Marangon-Lima JW, Zambroni de Souza AC, Perez-Arriaga IJ, Balestrassi, PP (2010). A Model to Long-Term, Multiarea, Multistage, and Integrated Expansion Planning of Electricity and Natural Gas Systems, IEEE Trans PWRS 25(2) 1154–1168
21. Victor DG, Jaffe AM and Hayes MH (2008) Natural gas and geopolitics, Cambridge University Press,Cambridge
22. Walls WD (2009) Natural gas and electricity markets. In: Joanne Evans and Hunt LC (eds) International handbook on the economics of energy, Edward Elgar, UK
23. Zhuang DX, Jiang JN, and Gan D (2011) Optimal short-term natural gas nomination decision in generation portfolio management, Eur Trans Electr Power 21(1) 1–10

Chapter 14
Challenges in Power Sector Regulation

Ignacio J. Pérez-Arriaga

> *Continued deregulation is the proper way to go, to the extent feasible… The central institutional issue of public utility regulation remains finding the best possible mix of inevitably imperfect regulation and inevitably imperfect competition.*
>
> *All competition is imperfect; the preferred remedy is to try to diminish the imperfection. Even when highly imperfect, it can often be a valuable supplement to regulation.*
>
> *But to the extent that it is intolerably imperfect, the only acceptable alternative is regulation. And for the inescapable imperfections of regulation, the only available remedy is to try to make it work better.*
>
> Kahn [15], "The economics of regulation", MIT Press, 1988

Regulation has to respond to economic trends, unremitting environmental concerns, changing political priorities and long-term objectives rising from climate change or energy security issues, but also to unexpected short-term events like the Fukushima accident, or technology breakthroughs such as shale gas extraction becoming competitive or the dramatic reduction in solar photovoltaic electricity generation costs.

Regulation is always walking the thin line between markets and governments. As the Alfred Kahn's sentence that opens this chapter says, no activity within the power sector can be entirely left to competitive forces without any supervision or regulatory support, nor is it advisable to give up the use of market forces and incentives completely, even when facing activities that are hardly amenable to competition. But regulation—when it has to be used—has to be sound, and market forces—when an activity is trusted to them—should not be tampered with unnecessarily.

The author thanks Carlos Batlle, Pedro Linares and Tommy Leung for their numerous and useful comments in the development of this chapter.

I. J. Pérez-Arriaga (✉)
Instituto de Investigación Tecnológica, Universidad Pontificia Comillas,
Alberto Aguilera 25, 28015 Madrid, Spain
e-mail: ipa@MIT.EDU

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3_14,

This is not an easy task. After the revolution of restructuring and liberalisation of the 1990s and early 2000,[1] power sector professionals have realised that designing and implementing successful electricity markets is much harder than initially thought. We are still experimenting with numerous topics to make electricity markets work properly, as profusely described in this book, from the design of wholesale market prices to capacity mechanisms, ancillary services and so on. Much progress has been made and much is still pending with incentive-based regulation to promote efficient distribution network investment. Disparate solutions have been adopted worldwide with respect to retail unbundling. No consensus has been reached yet on most regulatory topics in electricity transmission. And then, in the midst of this revision of the previous regulatory reform, a new revolution erupted, from several fronts simultaneously.

This time it is not only a question of markets versus governments. It is also not only the still open issues in the usual topics, which have been already examined in the previous chapters of this book. Now it is the new dimension to regulation—both in the content and role of the regulatory institutions—that must accompany the revolutionary challenge expected of power systems around the world in the next few decades to contribute to the transition to a sustainable energy model. Power systems will have to be decarbonised, neighbouring electricity markets will merge or coordinate to cover larger geographical areas, large amounts of intermittent renewable generation will be deployed everywhere, information and communication technologies will be integrated at all levels in power systems facilitating demand response and advanced control schemes, new business models will appear and universal access to electricity supply will be (hopefully) finally achieved.

14.1 Introduction

There is no doubt that the power sector is facing what is perhaps the major challenge in its history of less than 150 years. According to institutions such as the International Energy Agency [9], the International Panel on Climate Change [12] or the European Commission in its Energy Roadmap 2050 [2], the electricity industry will have to move from a generation mix that is mostly based on fossil fuels to a virtually decarbonised power sector by 2050, while supporting the electrification of transportation and heating. And this will have to take place in the midst of the still on-going process of a two-decade long regulatory reform, meant to introduce more competition and consumer choice and less governmental interference in this industrial sector.

[1] With some earlier exceptions, as the Chilean restructuring process in 1981 and the independent power production initiated in the US in 1978 with the PURPA Law.

Drastic regulatory and structural changes are not new in the power industry, as it has been described in Chap. 3 of this book. The early power sector developments at the end of the nineteenth century and at the beginning of the twentieth century were mostly driven by private initiative and competition. This was soon replaced by strong governmental intervention in the form of public ownership or treatment of the electricity companies as regulated monopolies, either publicly or privately owned. During most of the past century and until the 1990s, the electricity industry regulation worldwide was based on a quasi-standard regulatory approach involving heavy State planning and intervention, with the State being the sole regulator. A new regulatory paradigm, announced by the pioneer reform in Chile in 1981, was adopted by several countries in the early 1990s and quickly swept the world, although many power systems remain with the traditional regulatory framework. Many electric utilities were restructured; the potentially competitive activities of generation and retailing were unbundled from the network activities of transmission and distribution, which remained under regulatory control while offering open access. Independent system operators were given the responsibility of running the power system securely and guaranteeing the provision of ancillary services. Competitive wholesale and retail markets were created to improve efficiency and responsiveness to consumer preferences. Consumers had the freedom to choose their supplier. Incentive regulation was introduced in the network activities to improve their efficiency. Independent regulatory agencies were created to monitor market behaviour and to implement the regulation of the diverse activities. And, in theory, although it has never been truly achieved in most countries, the role and the direct political influence of governments was reduced. The jury is still out regarding how beneficial this reform has been for consumers and electricity providers.

Before the liberalised model has had time to consolidate (for instance, the European Union has set the target of 2014 for the completion of its Internal Electricity Market for wholesale transactions), we are facing again a new paradigm change that will lead to a future power system that is very different from the present one. The serious and justified global concern about climate change is affecting energy policy and power sector investments profoundly all over the world. Intense political oversight and intervention are already taking place in the form of all sorts of new energy policies, and much more is anticipated. In the future, the criteria of security and sustainability will have at least the same priority as efficiency in the regulatory design.

Without trying to prejudge the future, it seems clear that some features will characterise the power sector during the next decades.[2] First, we can anticipate a strong presence of renewable generation—with high variability and uncertainty,

[2] The MIT Professor Fred Schweppe, in his 1978 article "Power Systems 2000: Hierarchical Control Strategies", IEEE Spectrum, July 1978, pp. 42–47, was able to imagine the future more than 30 years ahead of his time and provided a futuristic description of a power system which very accurately describes the current trends, some of them already a reality. This chapter only attempts a modest extrapolation of emerging features of some actual power systems.

and distributed to a large extent—in many power systems. Second, the availability of communication and control technologies, combined with current trends in regulation and consumer behaviour, suggest a vigorous future for active demand participation, tightly related to the expected massive efforts in energy conservation and efficiency. In the absence of a technically and economically viable and widespread storage option, and with massive penetration of intermittent renewable generation, the future role of demand response cannot be over-emphasised. Third, the combination of the first and second features will stimulate the creation of new services and business models, possibly leading to a reformulation of the role of the traditional electric utility. Fourth, political developments, economic rationality and network reinforcements inexorably lead to an integration of existing power systems and markets into larger entities. Finally, in developing countries it is expected that during this period universal access will be finally achieved, and electricity consumption will grow to reach minimum standards of quality of life.

The existing experience with electricity regulation makes it easy to anticipate that a key *enabling factor* for a successful transition to efficient and reliable power systems that will contribute to the sustainability of the future energy model will be *across-the-board regulation* that:

- will foster research, development and deployment of low-carbon technologies for electricity generation, despite the fact that they might not be presently competitive in the narrow economic terms of existing electricity markets,
- will make possible, in particular, that adequate flexible resources (suitable generation, demand response, storage devices and interconnection facilities) are deployed and contribute to the reliable and efficient operation of the system in the presence of large intermittent renewable penetration, and
- will incentivise innovation efforts and investment in deploying advanced metering, monitoring and control systems in distribution and transmission networks.

Regarding the future evolution of regulation and regulatory institutions, the still unanswered questions are: Are the current electricity markets, the structure of the power sectors and the regulatory frameworks ready to meet the challenges for an efficient, secure and clean supply? In the context that can be anticipated of strong sustainability—and security-oriented policy measures, how can we improve or redesign power system regulation to facilitate these policies and reach their objectives efficiently? How can we make these policy measures compatible with the functioning of electricity markets? What is the role of regulation and regulatory institutions during this process?

In this chapter we shall focus first, in Sect. 14.2, on understanding the reasons that demand a revision of the current regulation, the unavoidable trade-offs that make designing a satisfactory regulatory framework difficult, the new context that this regulation will have to be applied in and the objectives that this regulation will have to achieve. Then, in Sect. 14.3, we shall venture into the mostly unchartered territory of providing some elements that could be useful in the search for answers to the questions posed in the preceding paragraph. Some miscellaneous comments

will be presented in Sect. 14.4. This will conclude the book, leaving the interested reader with the task of continuing the fascinating job of developing the regulation that will guide the future of power systems.

14.2 The Regulatory Challenges

In a nutshell, we can summarise the envisioned regulatory challenges in two blocks. First, the need, in a short time span, for an urgent transition to a sustainable energy model where the power sector must play a key role. Second, to make this transition in the current widespread context of liberalisation of the wholesale and retail electricity markets.

14.2.1 Transition to a Sustainable Energy Model

"If we do not change direction soon, we will end where we are heading…" Chinese proverb cited by the International Energy Agency [6].

The context: Sustainability concerns of the present energy model

The most often quoted definition of sustainable development was provided in the report "Our common future", also known as the Brundtland Report, from the UN World Commission on Environment and Development, in 1987. It is defined as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs."

Energy is key for sustainable development. Society relies on increasing supplies of energy to meet its need for goods and services.[3] Within the broader context of sustainable development, we can define a *sustainable energy model* as one that: (a) meets the minimum requirements for tolerable environmental impact and basic universal access *(worldwide)* to modern forms of energy supply; (b) facilitates a lasting, affordable and dependable energy supply that makes it possible to maintain or increase the total aggregated capital (economical, physical—either natural or

[3] The world primary energy demand has exceeded 500 EJ (exajoules) and will continue to increase, due to the growth in world population and the improvement of living standards. According to the middle projection in [11], global per capita gross domestic product (GDP) will increase by 2 % a year on average through 2050, mostly driven by growth in developing countries. Global population size is projected to plateau at about 9 billion people by 2050. Energy systems must be able to deliver the required energy services to support these economic and demographic developments. A conservative projection of the World Energy Outlook of the International Energy Agency [9] estimated a 60 % growth in primary energy demand between 2010 and 2030.

built—, social structures, knowledge) associated with human prosperity; (c) allows and promotes an equitable sharing of any surplus energy supply above the required minimum.

As acknowledged by any reputable energy organisation, the current path of world energy production and consumption, *even with presently expected policy measures*, is not sustainable since: (a) the environmental impact is intolerable[4]; (b) there are major concerns about access to lasting, dependable and affordable energy sources; and (c) there is an unacceptable worldwide disparity of levels of energy access and consumption.

Global warming is the most serious human-caused current change. The Intergovernmental Panel on Climate Change [12], in its fourth assessment of global warming, released February 2, 2007, used its strongest language yet in drawing a link between human activity and recent warming. "Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice, and rising global mean sea level." "There is new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities." "Most of the observed increase in globally averaged temperatures since the mid-twentieth century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations." It is a well-established scientific fact that humans are dangerously changing the climate of the planet.[5]

Climate change can affect us in multiple ways: heat waves, loss of water inflows to rivers, more frequent and more intense abnormal atmospheric phenomena, fires, melting of permafrost, changes in seasons and in animal habitats, illness vectors propagation to previously safe territories, rise of the sea level, large human

[4] In preparation for the Rio 2012 World Summit, the United Nations Development Programme (UNDP), with the collaboration of more than 600 scientists, published the report GEO5, Global Environment Outlook: Environment for the future we want, which starts with this statement: "The currently observed changes to the Earth System are unprecedented in human history. Efforts to slow the rate or extent of change—including enhanced resource efficiency and mitigation measures—have resulted in moderate successes but have not succeeded in reversing adverse environmental changes. Neither the scope of these nor their speed has abated in the past five years. As human pressures on the Earth System accelerate, several critical global, regional and local thresholds are close or have been exceeded. Once these have been passed, abrupt and possibly irreversible changes to the life-support functions of the planet are likely to occur, with significant adverse implications for human well-being."

[5] "When some conclusions have been thoroughly and deeply tested, questioned, and examined, they gain the status of "well-established theories" and are often spoken of as "facts." […] Climate change now falls into this category: There is compelling, comprehensive, and consistent objective evidence that humans are changing the climate in ways that threaten our societies and the ecosystems on which we depend." Quote from a letter sent by more than 200 members of the US Academy of Sciences and published in Science, vol. 328 7, May 2010 (see footnote 6).

migrations, etc., plus other effects of a nonlinear nature that are very difficult to predict, such as alterations in the ocean currents.

The currently accepted international goal is to contain the global mean temperature increase to less than 2 °C above the preindustrial level.[6] According to the best available science, this goal requires halving the world's GHG emissions by 2050—a reduction of something between 80–95 % in GHG emissions in developed countries—if we want to be within the 50 % probability band of not exceeding the threshold of 2 °C by the end of the century [12].

Although reaching a comprehensive international agreement to curb the GHG emissions at a safe level does not seem possible in the short term, at the time of writing this book, many countries around the world are adopting or considering stringent measures in this direction. For instance, the European Union has prepared a Climate Change Roadmap 2050, followed by an Energy Roadmap 2050 [2]. Figure 14.1 shows the European Union planned GHG emissions reductions until 2050 in different sectors. The power sector is presently the largest emitter of GHGs in the EU, and, according to the mentioned EU roadmaps, it should become fully decarbonised by 2050 in practical terms. Moreover, by 2050 heating and transportation must be heavily electrified, therefore posing a considerable additional burden on the electricity sector. The EU Energy Roadmap has examined several plausible scenarios until 2050 and has concluded that in all of them renewable energy sources move to centre stage, providing at least 55 % of gross final energy consumption in 2050 and at least 60 % of electricity production.

These figures are in agreement with those by the International Energy Agency [9]. The IEA estimates that for the 2 °C scenario, the global CO_2 emissions from the power sector in 2050 will have to be cut by almost 80 % from today's level of 12 GtCO2. In this scenario, according to the IEA, "more than 90 % of the global electricity demand in 2050 will be supplied by low-carbon technologies: renewable technologies will reach a share of 57 % in the world's electricity mix, nuclear power will provide around 20 %, and power plants equipped with carbon capture and storage (CCS) will contribute 14 %".

In June 2012, the US National Renewable Energy Laboratory (NREL) published a report [17] showing how renewable energy resources, accessed with commercially available renewable generation technologies, could supply 80 % of total U.S. electricity generation in 2050 reliably—of which 50 % consists of intermittent generation (wind and solar PV)—while balancing supply and demand at the hourly level. China has implemented a massive programme of energy efficiency that includes the manufacturing and deployment of solar and wind

[6] The IEA in the World Energy Outlook 2011 warns that "We cannot afford to delay further action to tackle climate change if the long-term target of limiting the global average temperature increase to 2 °C, as analysed in the 450 Scenario, is to be achieved at reasonable cost. In the New Policies Scenario—i.e. the policies that is presently foreseen that will be implemented—, the world is on a trajectory that results in a level of emissions consistent with a long-term average temperature increase of more than 3.5 °C. Without these new policies, we are on an even more dangerous track, for a temperature increase of 6 °C or more."

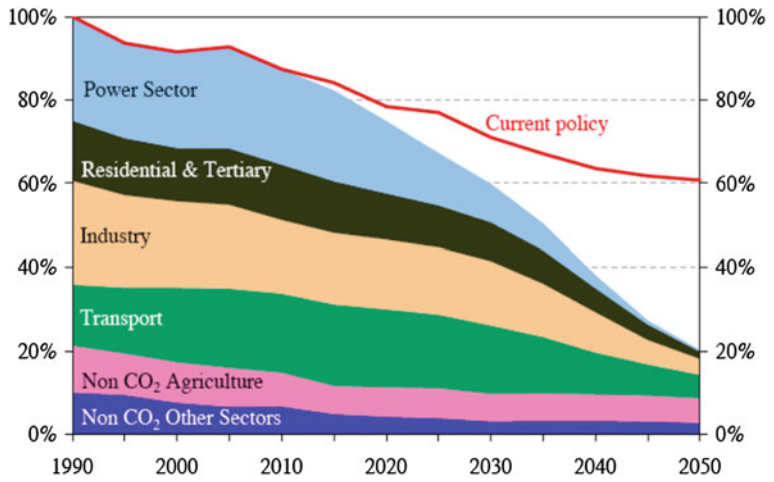


Fig. 14.1 EU GHG emissions towards an 80 % domestic reduction (100 % = 1990) [2]

generators. Many other countries are following suit, each with its own approach. All this will lead to a drastic departure from the current generation technology mix in a few decades and also to a profound revision of how power systems will be operated and planned around the world.[7]

The challenges: The need to reformulate energy regulation

As explained in the introduction to this chapter, these serious (and justified) global sustainability concerns will affect energy policy and the power sector, in particular, profoundly. As a consequence, justified intense political oversight and intervention are anticipated. Sustainability will have at least the same priority as efficiency and security of supply in the regulatory design. New and emerging clean technologies will be crucial in attaining a sustainable power system model, but their development and commercial deployment will typically need regulatory support. Power sector regulation looked much easier in the early 1990s, when the liberalisation movement started, since we were ignorant of many of the implementation difficulties and there were more degrees of freedom (less security and environmental constraints).

The current regulatory paradigm has to be reconsidered in this new context, where public energy policy will play a major role. This adds a new perspective to the present deliberation on the energy sector regulatory model and presents a dilemma for regulators: when facing the challenge of decarbonisation of the power sector, what is the role that regulation and regulatory authorities should adopt?

Ofgem, the UK Office of the Gas and Electricity Markets, has been quick in responding to this question. On its website, one can read: "Protecting consumers is our first priority. We do this by promoting competition, wherever appropriate, and regulating the monopoly companies which run the gas and electricity

[7] See [7] for a futuristic but credible long-term vision.

networks. *The interests of gas and electricity consumers are their interests taken as a whole,*[8] including their interests in the reduction of greenhouse gases and in the security of the supply of gas and electricity to them." Ofgem's duty to contribute to the achievement of sustainable development was introduced in 2004. In 2008, the Energy Act promoted this duty, placing it on an equal footing with the duties to meet reasonable demand and financing authorised activities. The Act also highlighted that Ofgem's principal objective, to protect the interests of consumers, refers to both future and existing consumers. These changes underline Ofgem's important and developing role in shaping the future of gas and electricity industries in a sustainable manner.[9]

14.2.2 The Classical Regulatory Dilemma: Markets Versus Governments

"There is a place for the market, but the market has to be kept in its place", Arthur Okun, Equality and efficiency, the big trade-off, Brookings Institution Press, 1975.

The context: Restructuring and competition in the power sector. A work in progress

More than 20 years of liberalisation and restructuring have taught us that creating well functioning, competitive wholesale and retail markets for electricity is very challenging, both technically and politically, and cannot be applied everywhere [22]. Where properly implemented, wholesale markets have led to improved performance and have attracted significant investments. This has required a firm political commitment to the reform. Despite significant failures and implementation difficulties, most liberalised power sectors are continuing with the process of reforms.

The logical sequence of events is to first create the legislative and regulatory framework and institutions, and then to restructure and/or privatise the power sector. Joskow [14] identifies some key components of the "textbook" architecture for restructuring, regulatory reform and the development of competitive electricity markets:

- Privatisation, to enhance performance and reduce the interference of the government.
- Vertical separation of competitive and regulated monopoly activities.

[8] Emphasis has been added.

[9] Ofgem's sustainable development work focuses on five themes: Managing the transition to a low carbon economy, eradicating fuel poverty and protecting vulnerable customers, promoting energy saving, ensuring a secure and reliable gas and electricity supply and supporting improvement in all aspects of the environment. See http://www.ofgem.gov.uk/Sustainability/Pages/Sustain.aspx.

- Horizontal restructuring to create a level playing field for competition.
- Independence of the System Operator with respect to the competitive activities.
- Energy and ancillary services markets and trading arrangements.
- Open access to the transmission network, plus adequate locational signals.
- Free choice of supplier with an adequate design of retail tariffs.
- Creation of independent regulatory agencies.
- Provision of transition mechanisms between the former traditional regulation and the competitive one.
- … and nothing more!

The numerous failures have had multiple causes, sometimes because of fatal departures from orthodox regulatory principles (frequently triggered by misguided opportunistic governmental interventions seeking to protect specific industrial sectors or domestic fuels, to create or to defend national champions against foreign takeovers, or to avoid consumer discontent at time of elections) and other times because life is much more complex than what theories can anticipate. Most of the times failures happened because of the inadequate structure of the power sector to accommodate competition at wholesale or retail levels, because of excessive horizontal concentration in the competitive activities, insufficient unbundling of competitive and regulated activities or lack of volume to hold competition, or lack of a suitable institutional framework. In other cases, the problem was an incorrect allocation of risk in the regulatory design that exposed some parties to unacceptable risk levels, as happened in California. Poor design of default tariffs (sometimes purposely kept low because of governmental intervention) may kill the retail market. And flawed market pricing rules, or the absence of adequate compensatory mechanisms, may lead to insufficient remuneration and lack of generation investment. Frequently, the absence of clear business models and rules for cost allocation, as well as the absence of expeditious siting procedures, have hampered transmission investment and prevented it from keeping up with demand growth and generation expansion. The success of regulatory reforms requires the adoption of an orthodox regulatory approach. However, the reforms will not succeed whenever the underlying structural and institutional conditions are not adequate or if there is no firm political commitment to the reform. Regulatory models are not easily transferred to countries facing different sets of conditions. Finally, nationalistic energy policies may ruin the best efforts in achieving effective supra-national electricity markets and even also competitive domestic markets.

Although the major regulatory themes of restructuring and liberalisation of the electricity sector have been properly identified, we are still quite far from having a consensus on how to approach some of the most important ones. Significant open issues remain in the design of power markets, in achieving a satisfactory level of investment and performance in the regulated network activities and in the interaction between markets and networks. They have been identified and abundantly discussed in this book and include:

- In wholesale market design, the organisation of trade, determination of prices, incorporation of network effects, provision of ancillary services, mechanisms to promote generation adequacy and firmness and demand participation.
- In the retail market, the impacts of default tariff design, interferences because of improper unbundling, sharing customer information and adequate switching processes.
- In the regulation of transmission networks, the planning criteria and responsibilities, available business models, allocation of costs or siting.
- In distribution networks, the remuneration in the face of anticipated or present challenges, network charges for embedded generation, quality of service, losses and innovation incentives.
- And in system operation, how to handle intermittency with a variety of existing and new resources, the potential need for new ancillary services and the integration into wider regional markets.

As we shall see below, many of the existing loose ends and regulatory imperfections in the listed topics will be amplified by emerging new factors of change that demand a revision of the present regulatory compact. For instance: intermittent renewable penetration will test market design and the rules of price formation; the extended geographical scope of regional markets significantly increases the complexity of transmission network capacity expansion planning; the need for flexibility will require levels of activation of demand response that have not been explored yet; the pressure on energy efficiency and conservation will reveal the conflicts of interest of distributors and retailers and will require ad hoc incentive schemes and the participation of new dedicated market agents, as well as specific attention to the interests and priorities of the end consumers [23]; and the standard approaches to remuneration of distribution networks will have to be modified with the strong presence of distributed generation.

The Challenges

A. *The limitations of energy markets*

The advantages of markets—when the appropriate conditions hold—for the efficient allocation of resources are well known and do not need to be proved. However, while the regulatory trend during the last two decades has been towards a stronger presence of energy markets, the capability of markets to face some of the major future regulatory challenges is now subject to close scrutiny. The debate about the limitations of energy markets has multiple aspects to be considered:

- First, we need to distinguish between generation and retail activities—well suited to a competitive setting—and those related to the transmission and distribution of electricity in networks, which have natural monopoly features and where competition may only be introduced marginally (through incentives to improve the performance of firms or the quality of service) or exceptionally as in the case of merchant transmission lines.

- Within the energy sector—electricity generation in particular—a debate that has resulted in much controversy and many publications exists about whether or not an energy-only market model can provide an adequate level of investment for a reliable supply. In the end, the ultimate responsibility is of the regulator, who has established the regulatory framework. The question, therefore, is about what is a sound market design: Should generation investment be left entirely to market forces, or should some mechanism be established to promote or to enforce some level of security of supply that the regulator considers to be satisfactory? Note that, in addition to adequacy and firmness of generation investment—already discussed in Chap. 12—the regulator may also wish to introduce some mechanism to promote additional flexibility in the system (both from the demand and the supply side, as well as from storage) to cope with an increasing presence of intermittent generation.
- Markets have difficulties taking into consideration uncertain future events such as the long-term availability of energy resources—their eventual depletion, affordability and reliability of supply—as well as the implications of the diversification of energy sources for the energy dependency level of any considered power system. The issue of concern regarding security of supply here is not dependency by itself, but the physical and economic vulnerabilities that might result from that dependency. This issue is directly related to the "strategic expansion policy" dimension of the security of supply in power systems, as explained in Sect. 12.1.1.
- Because of this shortsightedness, it is also difficult for energy markets to promote the development of those technologies that are most suitable under a long-term strategic perspective. Due to several reasons (which mostly have to do with political acceptability, but also with imperfect information), current energy prices fall short of internalising the total actual incurred costs of the different technologies. The agents of an energy market will not undertake costly investments in new technologies—typically with useful economic lives of at least 30 or 40 years—in a highly uncertain regulatory, technical and economic context. Investors, quite understandably, are not convinced that governments will internalise externalities to their full extent and that they will provide a sufficient regulatory long-term commitment. Investments in renewables, clean coal or nuclear are particularly salient in this respect, but all generation technologies end up being influenced by this uncertainty.

It is becoming increasingly clear that the current market paradigm, relying only on the correction of market failures through internalisation, is not providing an adequate response to the pressing environmental, economic and security of supply concerns. Because of political unacceptability issues, most governments are not willing to commit to sufficiently strong and long-term internalisation in present market prices of externalities with mostly long-term impacts, such as climate change [19]. Therefore, it seems necessary to bestow the market with some kind of long-term vision, so that, while minimizing the interference with the efficiency of the allocation mechanisms of markets, market agents receive additional signals to

steer them in the right direction. Energy policies supported by indicative planning could be of help in this regard.

B. *The need for a long-term vision*

The issue here is how to strike the right balance between competitive markets and regulatory intervention, private initiative and indicative or even mandatory planning. It is the role of an energy policy to establish the basic criteria to be met by a future sustainable energy model, as well as the specification of the major targets to be achieved, such as CO_2 emission levels, energy efficiency, penetration of renewables or minimum requirements of diversification in primary energy sources, and a basic nuclear energy strategy. It is also necessary that this energy policy includes a definition of the regulatory instruments that will make all this possible, while minimising interference with the functioning of energy markets.

Indicative planning is a term that has been used to define this ensemble of measures, although it is a misnomer, since some of them have a mandatory character. Indicative planning is more than just prospective analysis (find what could happen) and rather has a normative character (identify what has to be done to ensure that the future occurs with some desirable features). Indicative planning explicitly considers future energy alternatives and sets objectives for what has to be regulated. For example, while many aspects of electricity markets could be regulated, perhaps they should not all be regulated. The list of potential regulations include, for instance, targets for penetration of renewable energies or other clean technologies, objectives for energy efficiency and savings, support schemes to improve the security of supply, goals for sectorial carbon emissions, or priorities and resources for research and development.

National laws typically provide States with indicative planning as an instrument to set up their energy strategies.[10] The UK government white paper "Our energy future. Creating a low carbon economy" [5] begins by stating in the introduction "This white paper is a milestone in energy policy. It is based on the four pillars of the environment, energy reliability, affordable energy for the poorest, and competitive markets for our businesses, industries and households. This white paper sets out a strategy for the long term, to give industry the confidence to invest to help us deliver our goals—a truly sustainable energy policy". The French government commissioned a similar document, [1] which among its conclusions includes: "It is necessary to undertake without delay actions allowing us to be in 2020 on a virtuous trajectory to face the different long-term scenarios (2050 and further) … It is also required in parallel to become prepared for the long-term challenges, by devising from this very moment structural policies which will only bear fruit in the long term … This being a long-term perspective, it is evidently good policy to envisage very ambitious, even extremely ambitious objectives". Many other countries have followed suit. It can be argued that it is easy to set targets to be met 40 years from now, when nobody presently responsible to make

[10] This is the case of the current Spanish Electricity Sector and Hydrocarbon Acts, for instance.

this commitment will be in office. But this is only partly true. We are now one investment cycle away from 2050; therefore our present investment decisions are already partially locking in the behaviour of the power system in the 2050s. If these long-term plans are the outcome of a broad political consensus—across boundaries of the significant political parties—they will provide some of the security that investors in the power sector so highly value.

Energy policies should provide the integral vision for all stakeholders, plus the constraints to be respected, the targets to be met and the incentives to achieve both. The implementation of these policies should not curtail the freedom of installation for electricity and gas utilities, which would continue operating in a free market environment. But some sort of indicative planning should provide the framing conditions that should be known by all agents that may be affected, and should set up goals and well-determined resources for everything that is regulated.

Motivated by the pressing needs of finding solutions to security of supply risks and climate change, governments are trying to implement approaches along these lines, which mix centralised global objectives with policies and measures at national level. For instance, in the EU long-term targets—such as CO_2 emission reductions or penetration levels of renewables—are set at EU level, while leaving to subsidiarity of the Member States and to ad hoc global markets associated with each target—like the EU emission trading scheme or the possibility of exchanging renewable certificates among Member States—the actual task of deciding how to meet these commitments.

The main issue here is to assess, for each type of infrastructure or activity, the right amount and kind of regulatory intervention, so that the investment that takes place is compatible with the long-term sustainability requirements that have been previously identified. What is the borderline that the regulator should not cross? Perhaps too naively, at the beginning of the liberalisation process it was believed that all kinds of investment decisions should be "left to the market". Now, the more recent realisation of the serious shortcomings of our energy model demands a shift in paradigm. An energy policy will provide the identification of the objectives. Then, orthodox principles of microeconomics and regulation should be employed to determine the nature and intensity of the regulatory measures (quotas, incentives, or cross-cutting policies), if any, to be applied in each case: renewable energy penetration and the corresponding support schemes; energy efficiency and savings targets and how to achieve them; any support schemes required to improve the security of electricity and gas supply; the development of gas and electricity network infrastructures; priorities and resources for R&D in energy; carbon allowance allocations; and the practical implementation of any guidelines resulting from the public opinion on the future of nuclear energy.

Markets should be used as much as possible, with the prices of energy, emissions and green or white certificates sending the correct economic signals for investment in adequate technologies or consumption. However, while the long-term and sustainability implications of the energy model are not duly internalised in these prices (for reasons already explained), market instruments will need to be supplemented by other measures, such as R&D support, and also, especially in

those sectors such as energy efficiency and savings where externalities are more difficult to internalise fully and behavioural issues are more prevalent, by more traditional "command and control" measures such as standards.

The leadership to determine the national energy policy should come from governments, while complying with any higher level energy strategy (at a global or regional level), and avoiding any direct interference with decisions at the company level. But the role of governments should be limited to providing any necessary regulatory measures to energy markets that will make it possible for those markets to achieve the objectives of the agreed-upon long-term energy policies. Tensions and ambiguities will always exist regarding the fuzzy borderline between markets and governments. But in the energy sector, they must be seen not as opposite but complementary forces.

14.3 Elements for a Regulatory Response to the Challenges

How can power systems become fully decarbonised by 2050? How could they also support the electrification of transport and heating? How could they successfully integrate large amounts of intermittent and/or distributed generation? How could they make full use of Information and Communication Technologies? Should they incorporate consumer response and choice fully? How could they integrate neighbouring markets so that efficient commercial trade encompasses increasingly larger regions? Should the huge volume of new required infrastructure of generation and networks be planned? How to achieve universal access to electricity supply? How to encourage innovation and new business models? How to make markets and governments compatible? And, how to accomplish all this, reliably, efficiently and with acceptable environmental impact?

Here, we shall focus our analysis on a reduced number of very relevant topics that can help in providing a partial vision of what can be expected of power systems engineering, economics and regulation during the next decades: rethinking electricity market design, enhancing network regulation, providing universal access to electricity and investing in clean technologies.

14.3.1 Rethinking Electricity Markets Design

The design of future markets will have to be done within the limited space of freedom left by energy policy requirements, as explained in the past Sect. 14.2.2. Hopefully these policies will be preferably implemented by means of market-based instruments or, at least, by regulatory mechanisms that interfere with the market as little as possible. Wholesale electricity markets will have to coexist with environmental charges and constraints, instruments that correct market failures—such as capacity or perhaps flexibility mechanisms—, incentives to different technologies and levies or straight prohibitions imposed to others.

We cannot predict how the electricity markets will be a few decades from now, but it can be anticipated that some major issues will have to be addressed sooner or later, in one way or another: market performance and prices, the need for flexibility, attracting adequate investments, changes in system operation and geographical expansion. Some useful references on power systems prospective analysis are [2, 3, 7, 9–11, 17].

Market performance and prices

The examination of markets with strong penetration of wind and solar generation, and a parallel regulatory analysis, reveal some interesting features, not all of them obvious, as explained in Sect. 11.4.4: for example, more frequent cycling and efficiency losses of conventional mid-merit and even based-load plants, as well as the appearance of zero or negative market prices and curtailment of the output of generation with zero variable costs, such as wind and solar. An increment of wind or solar penetration has several simultaneous effects on market prices via reduction in demand to be met by conventional plants, increased operation costs because of cycling, and induced changes in the generation mix. The final impact depends much on the specific rules used to compute market prices in the considered power system. A future strong presence of intermittent generation would reduce the energy-providing role of the conventional fleet and increase its reserve-providing role.

Several conclusions may be extracted from this analysis. First, regarding the need for flexibility: electric sector modelling shows that a more flexible system is needed to accommodate increasing levels of renewable generation. Power system flexibility "expresses the extent to which a power system can modify electricity production or consumption in response to variability, expected or otherwise" [10]. In other words, a power system's flexibility expresses the capability of that power system to maintain a reliable supply in the face of rapid and/or large imbalances, whatever the cause. Flexibility must be present for different aspects of system operation and different time frames.

Electricity systems need flexibility and employ a range of resources to meet it within their technical, regulatory and market frameworks. The need for flexibility, resulting from variable renewables, demand and contingencies, can be met by four flexible resources: generation, demand response, storage and interconnections. Moderate volumes of intermittent generation can be handled by the inherent flexibility that already exists in most power systems [8]. Detailed simulation models show that rudimentary rules that specify the amount of "back-up capacity" that must be associated to wind or solar investments or the "capacity credit" of intermittent renewable generation, are inadequate to represent the diversity and the complex behaviour of power systems with large ratios of penetration of wind or solar generation.

A second conclusion is that the technical and economic conditions for conventional generators, in particular those that will be subject to more frequent cycling, will be tougher in the presence of large intermittent generation penetration: more expensive operation, very volatile prices and more uncertainty

regarding income, technical performance and regulation. The expected implications of this fact will be discussed later.

Storage and active demand response will be natural companions of the anticipated very strong penetrations of wind and solar PV generation. The deployment of *storage* depends on the progress of storage technologies to reach competitive costs, the expected spread of energy prices, the possibility to participate in the provision of ancillary services and perhaps of some regulatory support that remunerates the firmness of the storage response when system security is at stake. *Demand-side* options must play a crucial role in a high renewable electricity future. To make demand response options a reality, power systems will need to deploy advanced metering, real-time tariffs and behind-the-meter load management. The presence of competitive and focused energy service companies will also be necessary. Electrification of mobility will increase the demand for electricity and could be a significant contributor to demand response, in particular if "vehicle-to-grid" technologies are massively deployed. Electrification of heating will be another potential source of demand response.

Attracting adequate investment

Strong intermittency penetration is already pushing the present market rules to their limits in some power systems, and it is starting to reveal existing flaws in market designs and market failures that had remained unnoticed so far.

One of these failures will probably be the inability of current market pricing rules (with possible shortcomings in their design) to attract investors to install the kind of generation capacity that the future generation mix will need. The situation might be similar to what motivated the introduction of capacity mechanisms of some kind in most power systems around the world: to remedy a perceived market failure. Here, the problem could be that both the income and the operation costs of the units that will be subject to heavy cycling will be very volatile, uncertain and very much dependent on regulatory decisions (such as the volume of mandatory targets of penetration of renewable energy). Moreover, the "flexibility product" that is loosely perceived as necessary has not been properly characterised yet, and an ad hoc market or remuneration scheme does not exist for it. Under these conditions the prospective investors may not come in the desired volumes and the power system will end up with a shortage of flexible generation.[11] It is immediate to extend this discussion so that it also includes the other three sources of flexibility: demand response, storage and interconnections. Unless these sources of flexibility are properly characterised and, if this is the case, properly remunerated there will be a shortage of flexibility in power systems in the future. Clear, stable and long-term regulatory policies are needed to encourage investment in the appropriate technologies. "Clear statements of investment objectives and priorities

[11] The IEA report "Energy Technology perspectives, 2012" maintains that the "Provision of flexibility from dispatchable generation technologies is sometimes hindered by the absence of any market or regulatory structure to compensate such services."

at the political level can provide a framework for collaboration among diverse market players, while also helping to allocate roles and responsibilities".[12]

Therefore, some important regulatory questions remain open: How should a well-adapted generation mix, with a strong presence of intermittent generation, look like? Does this expected optimal mix,—which must include flexible but efficient generation, with much cycling and low capacity factor—need any regulatory support (e.g. some ad hoc ancillary services or capacity instruments) under market conditions? How do power systems plan for the "worst case scenario" consisting of several consecutive days with very low wind and solar output?

Revision of bulk power system operation

A strong presence of intermittent generation *(and the associated need for storage and active demand response)* will require a new paradigm in short-term system operation, from several points of view: operating reserves, monitoring and control, security analysis and stability. A "business as usual" system operation is not adequate to deal with large volumes of intermittent generation, integration of demand response and seamless coordinated congestion network management in large interconnected power systems.

These are the challenges. The standard operation procedures might need to be redesigned in the presence of:

- Significantly increased margin of error in the estimation of the equilibrium point between supply and demand, due to the large uncertainty in the forecast of intermittent and/or distributed generation, particularly one day ahead of real time, or longer.
- A richer variety of wholesale agents, such as aggregators of micro generation, of electric vehicles—either able of injecting power to the grid or not—, of storage and of active demand; increased trade with neighbouring systems also with large amounts of intermittent generation capacity.
- Participation of multiple non-conventional agents—such as aggregations of loads or of distributed generators—in the provision of ancillary services. Need for new types of ancillary services, for instance some sort of "flexibility" product.
- Formerly demand nodes in the transmission network that may sometimes become generation nodes because of the embedded generation within the corresponding distribution network.
- New stability concerns, due to the potential strong presence of generators with low or no mechanical inertia, or without voltage control capability, or with the potential risk of massive aggregated loss of production in a short amount of time—as when wind generators in a region must shut down because of excess wind speed, or by a voltage drop if they are not properly equipped with voltage-ride-through capability—.

[12] Ibídem.

And these are some tentative responses to the system operation challenges. They depend on the physical characteristics of the system, but also on coordinated efforts between the regulator and the system operator. For instance:

- How much penetration of intermittent generation is possible? In a hydro-dominated system with ample multiannual reservoir storage capacity, like Brazil, intermittency is not material[13]; moreover, the short installation time of wind generation (about 18 months), as compared to at least 5 years for hydro plants, is a guarantee against load growth uncertainty, since the threat of any forecasted medium-term energy shortage can be eliminated in a matter of months.
- Mostly thermal systems, on the other hand, need to adopt a suite of measures. Spain has organised a hierarchical system of monitoring and controlling wind generation, whereby each generator at each wind farm is supervised from a satellite control centre that reports to the national control centre; in case of need, each wind generator can be individually dispatched or curtailed.
- Technical standards in design and operation can be established for wind and solar PV generators to provide voltage control, inertial response or voltage-ride-through capabilities.
- Wind and solar output forecasts can be improved. Market rules can be changed to reduce the scheduling intervals in markets and to get the market clearing closer to real time, when deviations in wind and solar generation with respect to forecast values are much smaller.
- Conventional thermal technologies can be adapted to provide more flexibility. New kinds of ancillary services or other innovative products can be defined, for instance one that would specifically ask for flexibility. Here, flexibility comprises not only quick response, but also efficient operation, since frequently cycled plants must operate for a significant number of hours in excess of expectations for a typical peaking plant.
- Additionally, power systems can rely on the dispatchability of certain renewable technologies (e.g., biopower, geothermal, CSP with storage, hydropower and, within certain limits, also wind generation, for instance providing balancing support).
- Promote demand response, as described elsewhere in this chapter; and establish a sound regulation for the storage activity, so that a solid business model—perhaps subject to security-minded operation rules—can be defined once the technology has surpassed the competitive threshold level.
- The System Operator must resort to all agents that can provide useful services, like the aggregation of consumers, electric vehicles and distributed generators, including the surprisingly large number of back-up generators that stand idle in factories, office buildings, hospitals and many other installations, which could benefit from providing security support services without neglecting their primary objective.

[13] Only until the extreme case when intermittent production and run-of-the-river hydro may exceed demand, in which case wind or solar must be curtailed or water spilled.

- Future power systems must also feature new approaches for interaction between transmission and distribution control centres with enhanced capabilities, transmit greater amounts of power over longer distances to smooth net electricity demand profiles and meet load with remote generation, and leverage the geospatial diversity of the variable resources to smooth output ramping.

All of these needs will likely require technology advances, new operating procedures, evolved business models and new market rules.

Geographical extension and regional markets

There is a universal trend towards the integration of power systems in increasingly larger areas or regions, encompassing vast territories and multiple utilities, which previously had operated in virtual isolation. This has been made possible by the existence of more and more meshed interconnections among power systems with more transfer capacity.

This integration has multiple consequences, most of them very positive, as has been amply described in Chap. 10 of this book.[14] The integration of neighbouring markets into a regional one requires the coordination or harmonisation of several functions that now must be addressed at the regional level, even if some local individual aspects are respected. These extended regional functions include transmission network planning, a common regional trading platform at the wholesale (and also perhaps retail) level and coordinated transmission congestion management. Harmonisation is needed regarding the design of incentives to renewable generation or the adoption of capacity mechanisms.

It is interesting to compare at a high level the approaches that have been chosen by the European Union and the USA regarding regional market integration. As described in Chap. 10, the EU is in the process of adoption and implementation of a very simple EU-wide trading platform where network constraints are only represented in a very crude form; this allows the fast implementation of a regional power exchange, but does not use the transmission network efficiently and may create operation problems for the system operators.[15] On the other hand, the ISOs in the US use nodal prices at the local level, but this complexity (and the lack of a strong coordinating authority) have so far prevented a fast integration of the local markets into a regional one. A bi-annual EU-wide transmission planning exists, although it is not fully mandatory; meanwhile, only bilateral transmission planning agreements have been required by the federal regulator FERC in the US. There is still much to do until large regional market models can be said to work satisfactorily and region-wide economic signals exist. The design of the much smaller Central American Electricity Market (MER) is good, although it has run into other

[14] In particular, and more directly related to the topic of this chapter, the smoothing effect that integration over wider areas has on the volatility of the total output of intermittent generation in the entire region is important.

[15] Details about the current Framework Guidelines and Network Codes can be found at http://www.acer.europa.eu/Electricity/FG_and_network_codes/Pages/default.aspx.

problems, because of lack of willingness of the countries of the region to let this regional market function without interferences.

14.3.2 Enhanced Networks Regulation

Electricity networks present a different set of questions, since they are typically regulated as monopolies, i.e. some mix of cost-of-service and performance-based criteria. In the case of networks we want to know if the present regulation of transmission and distribution is adequate to support the anticipated changes that have been described in terms of distributed generation and storage, intermittency, electrification of mobility and heating, active demand response and storage, geographical expansion of markets and the creation of new services and business models. Reference [16] is a comprehensive study of these issues.

The challenge is to find the contribution that an enhanced electric grid could make to meeting the future energy needs in an efficient, reliable and environmentally responsible manner. How can regulation help to attain this goal? Does the regulation itself pose any obstacles? How "smart" should regulation be?

The transition from the present electricity grids to transmission and distribution networks with enhanced capabilities requires very significant volumes of investment in new facilities as well as in innovation efforts. Most of them are mainly related to the implementation of much more complex and sophisticated information, communication, and control systems. In addition, investment in grid infrastructure will be also needed to replace old assets, to increase network redundancy and to connect new generation sites and demand users. Finally, operational and maintenance costs should be re-evaluated taking into account the new structure and functionalities provided by these enhanced grids.

Existing electricity grids are already smart. But they need to become much smarter to cope with the new realities of a much more complex, decentralised and interactive power sector, in its way to facilitate an efficient, reliable and carbon-free electric supply. It will be a long, evolutionary process that will use and expand existing network capabilities and add new ones. The design and implementation of adequate regulation at both distribution and transmission levels will be essential in guiding the financial resources and technical capabilities of private firms towards this common objective.

Successful development of the multitude of network enhancements using state-of-the-art technologies that are included under the broad term of "smart grid" requires the application of sound economic and regulation principles: (a) recognise the specific physical and economic characteristics of networks (mostly natural monopolies) and the potentially competitive associated activities (retailing, metering, energy services, distributed generation); (b) find adequate remuneration schemes while always maintaining economic incentives for well justified investments; (c) find instruments to incorporate the deployment of effective innovative technologies in the remuneration schemes; (d) shortcuts and ad hoc rules that do not respect sound economic and regulatory principles will not do the job.

As mentioned previously, regulation and regulators cannot remain neutral when facing the abysmal lack of sustainability of the present energy model and the enormity of the required transformation of the power sector in the next decades, starting immediately. It is necessary that regulators accept their part of responsibility in this collective task. Once again, OFGEM's statement regarding their objective in regulation of electricity and gas networks is inspirational: "The overriding objective of a future regulatory framework for energy network companies is to encourage them to play a full role in the delivery of a sustainable energy sector and deliver long-term value for money network services for existing and future consumers. RIIO[16] is designed to promote smarter gas and electricity networks for a low carbon future." "It is in the interests of consumers that a company that delivers these outcomes is rewarded. Delivery will require significant investment and we will ensure that network companies that deliver efficiently are able to raise the required finance at a reasonable cost to existing and future consumers."

Transmission

In transmission networks, it seems that the challenges posed by the sheer size of the interconnected power systems and the anticipated large presence of intermittent generation will require careful consideration of, and possible substantial updates to, the current transmission planning criteria, definition of the responsible institutions for interconnection-wide planning, cost allocation methods, business models for transmission developers and siting procedures. As was described in the previous section, transmission system operation also must undergo a major renovation, but here we shall only focus on the network itself.

Specific questions are: How can we determine the nature of the transformation that will be needed in the transmission networks to accommodate the anticipated strong penetration of wind and solar generation? A transmission overlay or gradual reinforcements? What are the criteria to make a decision? Is it possible to make such a crucial decision now, given the level of uncertainty? Who should be in charge of finding an answer? What major barriers exist to find an answer and to implement it? How to address the problem of the two-way interdependence between generation and transmission investments[17]? How should we ensure that whatever is planned is finally built? How do we determine who pays? Transmission investment and planning for large interconnected systems (such as the EU or the two US interconnections) is still an open issue, since we do not know how to plan transmission expansion at region-wide level. Also, most regulatory authorities still resist

[16] RIIO stands for Revenue set to deliver strong Incentives, Innovation and Outputs. See the OFGEM report "RIIO A new way to regulate electricity networks", Final report of the RPI-X@20 project, 2010.

[17] This is an example of the well-known poultry dilemma. For instance, investments in wind or solar resources far away from load centres require transmission connections, which will never be built unless the wind and solar farms exist. In the US, Texas and California have enacted innovative regulations that solve this "chicken-and-egg" circular problem.

accepting the simple paradigm that "System Operators plan, regulators authorise, beneficiaries pay", and they fail to define adequate business models for investors.

Sound criteria for transmission expansion should be the basis of any transmission planning process and the subsequent allocation of transmission costs. In theory these criteria should address a variety of goals, such as reliability, cost reduction, market building, market power mitigation, and implementation of energy policies. But in practice just reliability and (not always) cost reduction are employed. The challenge is to go beyond "minimise cost subject to reliability constraints", starting by including under "cost" the estimated loss of utility of any non-served energy to consumers. Several interesting advances are taking place, both in the EU (with transmission planning by European Network of Transmission Systems Operators for Electricity (ENTSO-E)) and US (with US Federal Energy Regulatory Commission (FERC) Order 1000), but we are still far from being able to incorporate these ideas into an implementable transmission planning process.

We do not know how to allocate network costs of large investments that impact an entire regional network fairly and efficiently. Most regulators have not yet adopted the basic transmission pricing principles of having beneficiaries pay, establishing charges that are independent on commercial transactions, determining charges ex ante and, in a regional context, organizing cost allocation hierarchically (see Chap. 6). Moreover, these basic principles are becoming questioned under massive applications for grid connection from wind and solar farms.

A hierarchical scheme of cost allocation (first to inter system operators, then intra system operators) makes sense for lines that cross multiple regions or for any transmission facilities that are used in regional trade. Note that it is not strictly necessary to harmonise the internal cost allocation procedures of the different system operators, but it is necessary to agree on an inter-system operators cost allocation procedure.

Distribution

Many new technologies such as distributed generation, electric vehicles, and advanced metering infrastructure will connect, or are already connecting, to the distribution grid. Integrating these technologies will necessitate changes to the way the distribution grid is operated and maintained. Large capital investments in both distribution grids and the technologies themselves will be needed to ensure that the technologies are installed and operated effectively to realise the economic and environmental benefits they promise. Distribution regulation schemes will need to evolve appropriately to support the required investments. New, more sophisticated incentives may be needed to remove disincentives to modernisation and environmental sustainability that exist in current regulatory systems, and to ensure that new technologies are used optimally. These new technologies and the regulatory issues they raise constitute the challenge to be met in the next decades.[18]

[18] A detailed discussion of the future of electricity distribution regulation and regulatory proposals to address the anticipated challenges can be found in the doctoral thesis of Rafael Cossent, "Economic regulation of distribution system operators and its adaptation to the

Contrary to transmission, distribution networks originally were not designed to accommodate generation. However, their design, operation, control and regulation will have to be adapted to allow potential massive deployment of distributed generation (DG). Significant DG penetration generally results in additional costs of network investment and losses, an effect that increases with penetration levels. Global remuneration of distribution companies is frequently based on the volume of electricity distributed and distribution charges are generally collected through volumetric (€/kWh) charges. Since DG reduces the amount of electricity distributed, distribution utilities in general will be biased against DG and may create barriers to its deployment unless these regulatory shortcomings are properly fixed. Also note that most of the current regulatory mechanisms are focused on cost reduction and lack "natural" incentives for innovation.

Programmes, technologies and economic signals meant to conserve energy and improve efficiency from the consumers' side (the ECE measures discussed in Chap. 9) typically lead to reductions in the regulated revenue of distributors. As with DG penetration, these measures may not be welcomed by distributors, unless the efficiency gains are shared with them and the remuneration of the distribution activity is more precisely computed and made less dependent on the simplistic "distributed energy" metric.[19]

The conventional objectives of the distribution grid regulatory system have been to ensure that the distribution grid company efficiently makes the necessary distribution grid investments to provide a reasonable level of quality of service, and it is remunerated adequately and its costs are fairly allocated among and recovered from network users. Major regulatory changes in the early 1990s led to the development of incentive regulation that targeted improvements in distribution companies' cost efficiency. While challenged by classic problems such as information asymmetry between the regulator and regulated companies and the difficulty of regulating losses and quality of service, these conventional regulatory schemes have generally worked well.

But recent years have seen a greater public, political and regulatory awareness of electricity distribution networks' social, environmental and economic impacts, as well as the need to modernise the networks. This awareness has led to a push for specific social, environmental and economic objectives to be met by distribution networks. In some cases, these objectives are different or wholly new compared to the objectives of conventional regulation. The use of appropriate regulatory tools must help meet such goals.

One objective is environmental sustainability. Distribution networks should be capable of accommodating renewable generation from less polluting sources such as wind and solar generators. Distribution networks should also be mindful of

(Footnote 18 continued)
penetration of distributed energy resources and smart grid technologies", Comillas University, Engineering School, 2013. Adequate design of network charges is also of essence to avoid significant economic distortions, as explained in Sect. 8.5.6.

[19] As explained in Sect. 9.5, the retailing activity is also impacted by ECE measures.

reducing line and transformer losses, so as to reduce the production of pollutants from electricity generation. And enabling the large-scale connection of electric vehicles, distribution grids can indirectly support the reduction of gasoline consumption and air pollution. Many distribution networks might need to be upgraded to be able to accommodate these new challenges.

Another objective is modernisation for greater economic and operational efficiency. Improved power flow management systems can defer necessary investments in network capacity. Improved network control systems can minimise power outages and losses, improving reliability. Networks that can incorporate demand response programmes allow customers to modify their power demand on the fly, and can also defer necessary capacity investments while saving customers money.

14.3.3 Provision of Universal Electricity Access

Solving the lack of access or insufficient access to electricity of a significant fraction of the world population is a key component of the future sustainable power system model. To this purpose, rural electrification has to be explicitly considered a key element of the energy policy in developing countries, with specific support instruments and financing and business models that have to be able to attract large volumes of private investment, since this is a formidable task in terms of volume and organisation. This issue is included in this selection because it must represent a major effort during the next few decades and will make a very significant contribution to the sustainability of the current energy model.

The technical, economic and regulatory issues related to the rural electrification activity were presented in Sect. 5.8. It is a well-known problem, since almost every country has had to create and subsidise some kind of rural electrification programme to reach segments of the population that live in remote areas, are widely scattered in a territory, have low potential electricity demand needs and also low economic purchasing power.

The challenge for the next decades is to end this situation forever and to reach universal access to electricity. The International Energy Agency has estimated the costs of achieving universal electricity access by 2030 [6] and the strategies to finance them [9]. These costs are very small when compared to the total costs of electricity supply worldwide—less than 3 % of the global energy investment projected to 2030 (over $26 trillion in 2010–2030)—but staggering in absolute terms: $700 billion in the period 2010–2030, or $33 billion per year.[20] Therefore

[20] Adding 0.003 $/kWh, some 1.8 %, to current electricity tariffs in OECD countries could fully fund the additional investment. The total incremental electricity output of achieving universal electricity service by 2030 is around 950 TWh. This additional electricity generation represents some 2.9 % of the nearly 33,000 TWh generated worldwide in 2030 in the New Policies Scenario. To generate this additional electricity output would require generating capacity of 250 GW [6]

the key regulatory issue to be addressed is the viability of the economic business model or models—where the guarantee of regulatory stability is a major component—so that the required massive private investment could flow to these projects.

The most suitable models will probably depend much on the specific circumstances. When delivered through an established grid, the cost per MWh is cheaper than of mini-grids or off-grid solutions, but the cost of extending the grid to sparsely populated, remote or mountainous areas can be very high and long distance networks can have high technical losses. This results in grid extension being the most suitable option for all urban zones and for around 30 % of the rural areas, but it is not cost-effective in more remote rural areas. IEA estimates that 70 % of rural areas should be connected either with mini-grids (65 % of that 70 %) or with small, stand-alone off-grid solutions (the remaining 35 %). These stand-alone systems have no transmission and distribution costs, but higher costs per MWh.

Whatever regulatory models are adopted, they must be able to attract big private capital (on top of public governmental funds or international aid that is needed to cover the difference between the actual costs and the charges to final consumers) to guarantee a sustainable business model, so that an electric supply that meets some prescribed quality of service standards is permanently maintained. This may require relying strongly on the local social infrastructure to minimise the logistics, in particular in isolated rural communities. The design of the complete approach is a very considerable regulatory challenge.

14.3.4 Investments in Clean Energy Technologies

Answering to the formidable challenge of climate change calls for a quick transition to a future economy with a drastic reduction in GHG emissions. And this in turn requires the development and massive deployment of new low-carbon energy technologies as soon as possible. Although presently a number of promising technologies have been identified, the critical issue is whether or not to support them, and how to materialise this support—if this is the case—at a local and global level, possibly by integrating this effort into a global climate regime. This is the same classic discussion of markets vs. governments but disguised differently. Of course, establishing targets for penetration of technologies or groups of technologies amounts to picking winners, in one way or another, with well-known associated risks. But leaving it entirely to the market probably means a total absence of activity in this area, at a time where the precautionary principle strongly indicates that some action should be taken. This is a similar discussion to that in Chap. 11 on the justification to the support to renewable generation technologies and the most suitable methods to implement this support.

It can be safely stated that an adequate portfolio of existing and new clean energy technologies will not develop spontaneously under today's conditions. Current measures, such as the EU emission-trading scheme, do not provide the sufficiently high, credible and predictable future carbon price trajectory. Moreover,

there are important additional market failures that undermine the private incentive to invest in clean energy innovation [18], notably the spillover effect of R&D: since innovation has a large element of pure public good, it is unlikely and may be undesirable that innovators capture all the learning benefits. All this contributes to creating high uncertainty about future market revenues from the exploitation of new clean technologies, resulting in the failure to deliver an adequate and timely level of private investment for R&D, development and commercial deployment.

Thus, there is a need for regulatory support. However, as the sentence that starts this chapter indicates, any regulatory fundamentalisms must be avoided. Sound energy policies are typically compromises between competition and regulation.

Clean technologies can be demand oriented—such as improvements in energy efficiency or conservation, or behavioural changes by consumers—or supply-oriented—such as low-carbon generation technologies or the use of fossil fuels with carbon capture and storage. These technologies may not be competitive and they can be at different levels of technological and commercial maturity. The available regulatory instruments can be classified as "market pull"—i.e. creating market conditions where these still non-competitive clean technologies may have a chance with some help, such as a carbon tax, a feeding tariff or a mandated quota—, or "technology push"—i.e. a direct support to innovation in the form of loan guarantees or contribution with public equity capital.

Since for the most part the deployment of clean electricity is policy driven so far, in this policy-driven market the regulation itself is a major risk factor. To unlock finance for clean technologies, investment-grade regulation is necessary. This means a compelling vision, supported by a precise, clear and stable policy [4].[21]

The topic of investment in clean energy technologies exceeds the scope of this book on electricity regulation, although the multiple linkages are obvious. More specific issues pertaining to investment in adequate electric technologies have been already discussed in Sects. 14.3.1 and 14.3.2. A broader discussion on the preferred approaches to foster low-carbon energy technologies from a regulatory point of view can be found in [20] and [21]. An excellent technical and economic assessment can be found in [10] and [11].

14.4 Miscellaneous Comments

Electricity regulation in theory and practice

This book has presented the theory of electricity regulation, interspersed with many experiences of its implementation in diverse international settings. The book, still, leans more on theory than in practice. Theory is good to set a reference,

[21] "A compelling vision, backed up by precise, simple, clear policy, needs to be implemented if larger institutions and investors are to create the argument internally that a greater proportion of the balance sheet needs to be available for sustainable energy" [4].

a standard against which one has to compare any actual regulatory instrument to be applied. But most of the time real life departs from the predictions of regulation theory, possibly because regulation theory has overlooked some essential facts of real life. These are some of the reasons:

- The actual behaviour of the agents in the energy sector is usually too complex to be captured by theoretical analysis. Human imagination is boundless, and the responses of the agents to ambiguous, incomplete or flawed regulatory instruments are really unbelievable. The agents would stretch the interpretation of rules to the limit and exploit the tiniest inconsistencies to serve their own interests. On the other hand, most often the dominant players in a market will not exercise all the market power they have, in order to maintain a favourable status quo as long as possible.
- There are numerous ill-known, hidden barriers that prevent the complete and successful implementation of sound and well-intentioned regulation. Most times this happens because of the absence of an institution in charge of eliminating these barriers, or because nobody thought of them and they are perfectly legal.
- Governments interfere with regulation—typically because of spurious short-term electoral reasons—and take over responsibilities that should be left to independent regulatory commissions. This problem is minimised when the responsibilities of the government and the regulatory authority are clearly specified and the latter is truly independent.
- Complexity. The devil is in the details. What in theory seems to be equal (e.g. the outcome of competitive markets and centralised generation planning, or quantity-based mechanisms versus price-based mechanisms in the promotion of renewables) is actually very different in practice, typically because the theory has overlooked something. Even experienced regulators are frequently puzzled by the rich diversity of new situations that they have to address and that do not fit within any existing jurisprudence.
- A sound and successful regulatory instrument may become obsolete and even detrimental if it is not updated for a long time. Regulation needs to be periodically checked for obsolescence, since the power sector is constantly evolving. A good example is RPI-X regulation: while some countries are still trying to adopt and master it, other pioneering regulators are trying to introduce new features to cope with its lack of focus on technology innovation, among other things.
- Attempts to achieve a perfect solution for any specific regulatory topic in one single step are frequently fruitless. There are so many hands intervening in the design, approval and implementation of a regulatory measure that the outcome is never a well-rounded piece of legislation. Most times it suffices with making some progress in the right direction and waiting for the next opportunity to advance a bit more.
- Regulation theory may lack a comprehensive vision of reality, and it might be excessively focussed on the performance of a specific sector. This may be one the lessons and the leitmotif of this chapter: the need to realise of the lack of

sustainability of the present energy model and the multiple regulatory challenges associated to this long-term threat. Regulation has to be reinvented to explicitly incorporate this fact into its DNA.

- Finally, as it has been universally verified on multiple occasions, effective regulation has to be "loud, long and legal" [4]. *Loud*, i.e. strong enough, with the capacity of making a difference, so that investments in clean energy become commercially attractive. *Long*, i.e. policy instruments that are sustained for a period of time that is consistent with the financial characteristics of the project; stability—not rigidity—is an essential feature of good regulation. And *Legal*, so that policy instruments are based on an orthodox, clear and well-established regulatory framework, as a guarantee of stability. The policy objectives must be unambiguous, with clear enforcement provisions; the regulation must be streamlined across all relevant factors within the boundary of the matter at hand; and complexity and variables that might add risk must be reduced to a minimum.

Effective action in spite of uncertainty

"Energy is central to addressing major challenges of the twenty first century, challenges like climate change, economic and social development, human well-being, sustainable development and global security. … Without question a radical transformation of the present energy system will be required over the coming decades. … A new found appreciation by policymakers of the multiple benefits of sustainable options and their appropriate valuation will be critical for the transformation to occur. … Policy, regulations and stable investment regimes will be essential. A portfolio of policies to enable rapid transformation of energy systems must provide the effective incentive structures and strong signals for the deployment at scale of energy-efficient technologies and energy supply options that contribute to the overall sustainable development."[22]

Electric power systems are the most critical component of the present energy model and its importance will only grow in the next decades as it will have to become almost completely decarbonised while having to provide clean electricity to much of the transportation and heating sectors. This must have profound implications for electricity regulation. However, future electricity regulation is shrouded in uncertainty. It is not clear whether the most convenient approach for the task of reviewing the regulatory framework of existing wholesale markets, when confronted with the challenges that have been exposed, should be a mix of mitigation measures or a complete market overhaul and redesign. It probably will be an evolutionary process, with a long-term adaptive vision becoming clearer and gradual steps that allow the progress happen towards the moving target.

The most relevant regulatory issues to be considered in the review of the present market design have been highlighted in this chapter. A similar situation

[22] *Source* "Global energy assessment. Toward a sustainable future". 2012. International Energy Agency [9–11].

exists for electricity networks, which also need a new regulatory framework that shifts the current emphasis from cost reduction towards a longer-term role as the indispensable enabler of the major changes to come. As described before, energy experts presently debate whether it is better to set mid-term penetration targets for clean technologies, just support R&D for technologies that are far from maturity, or rely on carbon prices and internalisation of any other environmental externalities.[23] Still, despite of this uncertainty, regulators must adhere to the mantra of "loud, long and legal".

Power system modelling and simulation tools

The proper analysis and implementation of the diverse suite of measures listed in this chapter requires enhanced *modelling and simulation tools* of power systems that will handle transmission network planning with the required level of dimensionality and uncertainty (mostly regarding the technology, location and volume of future generation investments). It also requires the ability to quantitatively analyse electricity markets under new circumstances. Analysis topics include optimal generation mixes, market pricing rules and reliability assessment, the integration of demand response, electric vehicles, distributed generation storage and advanced automation schemes in distribution networks, and security analysis, which comprises monitoring and controlling the bulk power system while taking into account the interactions and coordination among system operators. This is an ensemble of tasks of vast proportions that will require a close collaboration between universities, public research centres and the power industry.

New future electric utility formats

All of these future power sector developments will likely take place in a different utility context. The traditional utility will have to evolve and transform into something different, while new services and business models will emerge and take its place (at least partially). Imagine, for instance, a future scenario where decentralised residential solar PV becomes truly competitive with centralised power, without any explicit or implicit subsidies.[24] No doubt a new brand of utilities would emerge that: (a) would install solar panels and also perhaps some batteries plus a low-temperature solar thermal system (all owned by the firm); and (b) would charge a monthly fee covering all the electricity and heating services, including any charges for back-up electricity or network utilisation. The conventional vertically integrated utility (or its centralised generation and monopolistic network businesses, if they are unbundled) would be increasingly reduced to the provision of back-up power and centralised network maintenance. The transmission and

[23] This author strongly supports that carbon prices, because of the inability of decision makers to reach global agreements and make citizens understand why they are necessary, will take a long time to reach a minimum level of effectiveness in promoting clean technologies. Therefore, other measures have to be adopted in parallel.

[24] See Chap. 8 on tariffs, where it is explained how volumetric network charges plus net metering amount to an implicit subsidy to distributed generation.

distribution system operators would have to manage all the new situations described above, as indicated.

Prosperity without growth

And finally, a few words about the dilemma of prosperity without material growth. The perspective provided by the observation of the energy system shows that, in the present circumstances, the objective of a developed society should be to grow better and not to grow more (in terms of physical output and corresponding material inputs). We have to evolve from the industrial paradigm of continued statistical growth to a new paradigm of deliberate restraint and moderation. This message has to guide the careers of young energy professionals in our societies, who are the ones who will have to design and apply this new paradigm.

Quoting [13], "Prosperity is not obviously synonymous with income or wealth. There is a growing recognition that, beyond a certain point at least, continued pursuit of economic growth does not appear to advance and may even impede human happiness. … Any credible vision of prosperity has to address the question of limits. … The modern economy is structurally reliant on economic growth for its stability. The idea of a non-growing economy may be an anathema to an economist. But the idea of a continually growing economy is an anathema to an environmentalist. No subsystem of a finite system can grow indefinitely, in physical terms. In short, we have no alternative but to question growth. … Prosperity consists in our ability to flourish as human beings within the ecological limits of a finite planet. The challenge for our society is to create the conditions under which this is possible. It is the most urgent task of our times."

References

1. CAS (2008) Perspectives énergétiques de la France à l'horizon 2020–2050. Centre d'Analyse Stratégique. Rapport de la Comission d' Énergie presidée par Jean Syrota
2. European Commission (EC) (2011) A roadmap for moving to a competitive low carbon economy in 2050. EU Commission (DG Climate), COM(2011) 112 final, March 8 2011
3. Eurelectric (2009) Power choices. Pathways to carbon-neutral electricity in Europe by 2050. http://www.eurelectric.org/media/45274/power__choices_finalcorrection_page70_feb2011-2010-402-0001-01-e.pdf
4. Hamilton K (2009) Unlocking finance for clean energy: the need for 'investment grade' policy. Chatham House, London
5. HM Government (2003) Our energy future. Creating a low carbon economy: white paper Cm5761, The Stationery Office, London
6. IEA International Energy Agency (2011a) World energy outlook 2011
7. IEA International Energy Agency (2011b) Solar energy perspectives
8. IEA International Energy Agency (2011c) Harnessing variable renewables
9. IEA International Energy Agency (2012a) World energy outlook 2012
10. IEA International Energy Agency (2012b) Technology perspectives
11. IEA International Energy Agency (2012c) Global energy assessment. Toward a sustainable future

12. IPCC (2007) The intergovernmental panel on climate change. Fourth Assessment Report. http://www.ipcc.ch/
13. Jackson T (2009) Prosperity without growth. Economics for a finite planet. Earthscan, Routledge
14. Joskow PL (2006) Introduction to electricity sector liberalisation: lessons learned from cross-country studies. In: Sioshansi FP, Pfaffenberger W (eds) Electricity market reform: an international perspective, p 392
15. Kahn AE (1988) The economics of regulation: principles and institutions. MIT Press, Cambridge
16. MIT (2011) The future of the electric grid. Massachusetts Institute of Technology. http://mitei.mit.edu/publications/reports-studies/future-electric-grid
17. Newbery D, Olmos L, Rüster S, Liong SJ, Glachant JM (2011) Public support for the financing of R&D activities in new clean energy technologies. Final report. THINK project. Florence School of Regulation. http://www.eui.eu/Projects/THINK/Home.aspx
18. NREL National Renewable Energy Laboratory (2012) Renewable electricity futures study. June 2012
19. Pérez-Arriaga JI, Linares P (2008) Markets versus regulation: a role for indicative energy planning. Energy J (Special Issue):149–163
20. Pérez-Arriaga JI (2009a) Regulatory instruments for the deployment of clean energy technologies. Proceedings of the annual meeting of the national academies of engineering, Calgary, Canada, July 2009, and Working Paper 09-009, Centre for Environmental and Energy Policy Research, MIT, July 2009
21. Pérez-Arriaga JI, Linares P (2009b) Promoting investment in low-carbon energy technologies. Special issue on Incentives for a low-carbon energy future of the European Review of Energy Markets, Sept 2009
22. Sioshansi FP (ed) (2008) Competitive electricity markets: design, implementation, performance. Elsevier, Amsterdam
23. Sioshansi FP (ed) (2011) Energy, sustainability and the environment. Technology, incentives, behaviour. Elsevier, Amsterdam

Annex A
Case Example of Traditional Regulation of the Electric Power Sector

The traditional system for regulating the electricity industry is the outcome of many years of analysis, effort and regulatory experience from which much can still be learnt.

The electricity industry is not only enormously complex, both economically and technically speaking, but is a key factor in the economic and social development of any society. From its inception in the early twentieth century and throughout the history of the industry to its present maturity, enormous effort went into improving and optimising decision making and operating processes to develop what is known today as traditional regulation. In order to introduce the traditional regulatory framework avoiding generalisations, this annex has taken the electricity industry in the USA in the early 1980s as a specific example.[1]

The information contained in this annex is particularly relevant for the following reasons: First, the electricity industry in the USA, especially in the years mentioned, was an archetype of traditional regulation. Indeed, for many years prior, it had (both federal and state-wide) regulatory bodies that had conducted in-depth analyses to optimise electricity industry organisation and develop its regulatory components. Moreover, the size of the industry in the USA and the large number of companies and regulators involved made it possible to devote enormous resources to such optimisation and continuous regulatory upgrading. Traditional regulation, with a diversity of formats, currently is the regulatory framework of choice in many countries of the world, including about one-third of the states in the USA.

The time frame chosen for analysing the industry is particularly relevant, since in the early 1980s not only had traditional regulation reached its zenith in the USA, but at the same time was the subject of the initiatives and studies that would later pave the way for the present liberalisation of part of the electricity business.

[1] This text has been adapted and translated with support from the Florence School of Regulation from the report "The electricity tariff in the USA", written in 1984 by Ignacio Pérez-Arriaga, Carmen Illán and Andrés Ramos of the Institute for Research in Technology (Instituto de Investigación Tecnológica, IIT), Comillas University, for the Spanish utility Unión Fenosa. Juan Rivier, also from IIT, and Ignacio Pérez-Arriaga have prepared this revised version.

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3, © Springer-Verlag London 2013

Indeed, the PURPA[2] act, approved only a few years before, was the first step towards what would develop into a general questioning of traditional procedures.

Expansion and operating planning in traditional regulatory environments, not dealt with in this annex, were briefly explained in Chap. 1 of this book, which introduces the concept of decision-making hierarchy based on long-term, medium-term, short-term and real-time horizons. That brief discussion provides an overview of the general framework in which such decisions are implemented.

This annex addresses the analysis of the methods usually employed in North American electricity companies to establish electricity rates. Be it said from the outset that the industry in the USA was characterised in those years by enormous variety, geographically and historically speaking, and that the present discussion focuses on the most representative forms of the traditional method, without taking more than a cursory look at the novelties that had then begun to shape the industry as it is known today. Consequently, this annex is structured as follows.

The Sect. A.1 is an introduction to the electric power generation industry in the USA in the early 1980s: structure, organisation, regulatory bodies and basic ratemaking issues. The four parts that follow are devoted to specific issues.

The Sect. A.2 describes the revenue requirements method, which was universally used in the 1980s (and still now in numerous states) in the USA to determine cost of service, although with variations in its specific application.

The Sect. A.3 deals with ratemaking for the sale of electric energy to a utility's customers (end consumer sales), a process that was subject to the legislation in effect in the respective state.

The Sect. A.4 discusses the establishment of rates for wholesale energy sales (energy allocations and purchases in inter-utility energy exchanges and sales to distributors) among electric utilities, regulated by the federal (i.e., central) government.

The Sect. A.5 describes the organisation of power pools, more or less formal groupings of the US electric utilities with varying levels of integration, and intra-pool energy exchanges. These pools, the predecessors of today's regional transmission organisations in the US, which were governed at the time under a traditional regulatory scheme, now operate as competitive markets run by independent system operators.

The Sect. A.6 and last section offers a critical evaluation of the traditional method of electricity regulation.

[2] *Public Utility Regulatory Policies Act:* federal law enacted in 1978 in the USA that initiated the transformation of the electricity industry, indicating the need for more efficient rate design, authorising qualifying facilities (distributed generators using renewable sources or CHP systems remunerated on the basis of avoided cost), and so on.

A.1 The Electric Power Supply Industry in the USA

A.1.1 Introduction

This Sect. A.1 of the annex contains a brief description of the most prominent characteristics of the complex American electric power supply system as it stood in the 1980s. The physical features of the overall US system of generation, transmission and distribution and the number, ownership and organisational structure of the different utilities operating at the time are described in Sect. A.1.2. Electricity industry regulatory institutions and their powers, particularly as far as ratemaking is concerned, are discussed in Sect. A.1.3. Section A.1.4 summarises the basic characteristics of electric power rates in the US and introduces the subjects addressed in the following four sections: cost of service, wholesale energy sales, rates by each type of consumer and power pools.

The two figures show the typical configuration of vertically integrated utilities, sometimes with embedded distribution companies connected to them, the spontaneous relationships between them (Fig. A.1) and the structured and coordinated transactions that take place when companies choose to function within a "power pool" (Fig. A.2).

A.1.2 Basic Characteristics

The most characteristic features of the American electricity industry in the 1980s can be summarised as follows[3]:

- There were nearly 3,500 utilities, whose ownership was private, public or co-operative. The numbers of each and their relative weight in the system as a whole are shown in Table A.1.
- These utilities' generation plants and transmission grids were grouped into three large non-connected geographic systems: Eastern Interconnection (76 % of the total), Western Interconnection (17 %) and the state of Texas (7 %).
- The American electricity system, which comprised the above three physically independent but synchronically operated systems, had an organisational structure that can be broken down as follows:

[3] The most authoritative sources for this type of information are the statistical surveys published from time to time on the US electricity industry, such as [1–5].

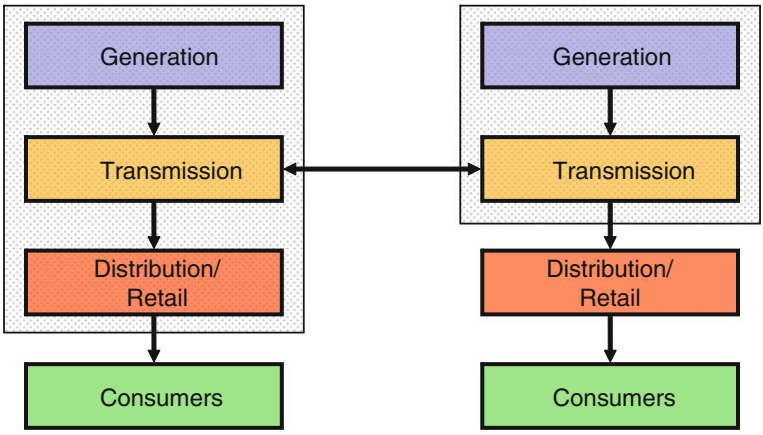


Fig. A.1 Typical configuration of vertically integrated utilities

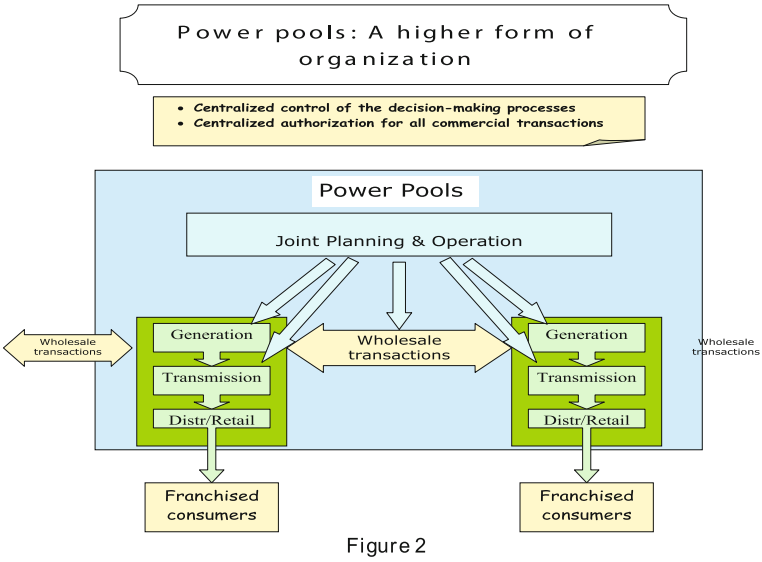


Fig. A.2 Power pools: a higher form of organisation

– The nine regions pertaining to the (NERC), an organisation created in 1968 to enhance the reliability and security of electric power supply in the USA. The electric utilities in each of these regions co-ordinated the planning and operation of their respective systems to attain a suitable level of reliability.

Table A.1 Ownership structure of the US electricity industry in 1980, taken from [8]

Ownership	Number of companies	% share Generating capacity	% share Total generation
Private utilities	237	78.0	78.0
Co-operatives	960	2.5	2.8
Federal systems	6	9.6	10.3
Municipal utilities	2,248	5.6	3.8
State projects	–	4.5	5.2
Total	3,451	100.0	100.0

- Natural groupings (not necessarily power pools) of electric utilities closely inter-related through energy exchanges, strong interconnections, geographic proximity, co-ordinated operation and so on. The overall system was divided into 26 such regions, further to the NERC classification [4].
- Power pools, i.e. formal organisations established by two or more electric utilities to improve their economic performance and short-, medium- or long-term security and/or reliability. The degree of inter-utility integration ranged from simple rather unspecific agreements on energy transactions for reasons of economy to detailed arrangements for co-ordinated operation and planning among pool members. At the time there were around 30 power pools in the USA.
- Individual electric power generation, transmission and/or distribution companies (see Table A.1). Some private and most of the municipal and co-operative utilities were mere distributors of electricity, with no generation or transmission facilities of their own. There were six federal agencies (the Tennessee Valley Authority, the largest electric utility in the USA, among them) that generated electric energy in federally owned facilities, all of which was then sold wholesale to other electric companies.

- The combined generating capacity of these three systems was on the order of 590,000 MW. The 10 largest utilities accounted for 25 % of this capacity, the 30 largest for 50 % and the 100 largest for a little over 80 % [4].
- The 100 largest utilities had capacities ranging from 1,500 to 30,000 MW. The generating capacity of the 100 smallest that generated power was under 1 MW in all cases. Of all the other utilities that generated electricity, around 100 had capacities ranging from 250 to 1,500 MW and around 700 fell in the 1–250 MW category [4].
- The peak demand in 1979 was 400,000 MW, and peak consumption 2.4×10^{12} kWh; 47 % of this energy was produced with coal, 14 % with nuclear energy, 16 % with fuel-oil, 12 % with gas and 10 % with hydraulic power [4].
- A substantial volume of electricity was sold in inter-utility wholesale transactions. In 1979, such sales in the private sector came to 18.5 % of the

Table A.2 Financial information on 31 December 1980 and 31 December 1979 (billion dollars) [6]

	1981	1980	% increment
Operating account			
Electricity industry revenues	118.1	100.8	17.3
Electricity industry expenses	101.2	88.4	17.0
Electricity industry profit	17.0	14.4	18.5
Other profits and deductions	4.7	3.9	20.3
Net profit	12.7	10.7	19.2
Source of funds			
Net profits	12.7	10.7	19.2
Non-cash credits	8.3	7.4	11.5
Total working capital	21.0	18.1	16.1
External funds	23.3	22.0	5.8
Other	3.5	3.2	11.6
Total funds obtained	47.6	47.3	10.5
Application of funds			
Construction and investment	28.1	26.4	6.8
Dividends	9.8	8.8	14.4
Debt repayment	5.9	4.8	23.9
Other	3.9	3.5	11.8
Total funds applied	47.2	47.3	10.5

net power generated by that sector, and in the system as a whole they accounted for 30 % of total net generation.

- Capital expenditure in 1979 amounted to $\$34 \times 10^9$: 73 % in generation, 10 % in transmission and 13 % in distribution. In 1969 these percentages were 50, 19 and 28 %, respectively [4].
- The following tables shown here provide a quantitative overview of the most prominent features of the American electricity industry in the early 1980s. Most of these tables were taken from [6] (See Tables A.2–A.5).
- In 1981 the average rate of return on American private electric utilities' rate base came to 9.1 %.

A.1.3 Regulation

In the market economy environment prevailing in the United States in the 1960s and earlier, the electric power supply industry was one of the most heavily regulated in the various areas of its business. This trend intensified in the 1970s due to growing public interest in issues such as the environment, security, quality of service, scarcity of fossil fuels and cost of electricity.

Table A.3 Tangible fixed assets on 31 December 1981 (billion dollars) [6]

Steam plants	5.41
Nuclear plants	2.78
Hydroelectric plants	0.08
Other fixed assets, generation	0.09
Total fixed assets, generation	8.36
Transmission	2.16
Distribution	4.06
General	0.64
Experimental plants	0.03
Total fixed assets for production	15.25
Nuclear fuel	0.65
Fixed assets in progress	9.00
Other	0.89
Total tangible fixed assets	25.79

The following areas of the electricity industry were regulated by local, state and federal institutions in the USA:

- Rates for the wholesale and retail sale of electric power.
- Land use, in particular with respect to facility siting.
- Environment.
- Financing.
- Fuel.
- Electric utility organisational structure.
- Electricity network characteristics and location.
- Power Pools.

The specific regulation of these areas was incumbent upon the respective local, state or federal level.

Federal

The federal government is constitutionally empowered to regulate interstate trade and control federal land. Federal regulation of the electricity industry was governed by a series of laws enacted by Congress, namely: the Securities Act (1933), Securities Exchange Act (1934), Public Utility Holding Company Act (1935), Federal Power Act (1935) and Public Utility Regulatory Policies Act (1978, i.e. at the dawn of the period considered here). These laws were applied and enforced by independent agencies, whose powers and functions at the time are described below:

- Federal Energy Regulatory Commission (FERC) is an independent agency under the aegis of the Department of Energy with broad powers to regulate interstate energy transmission and wholesale electric energy transactions. The FERC had authority over the rates charged in interstate electric power sales and the rates for energy generated in federal projects were also subject to FERC

Table A.4a Ownership structure (private utilities) on 31 December 1981 (billion dollars) [6]

	Billion dollars	%
Long-term debt	115.42	50.3
Preferred shares	26.57	11.6
Share capital	61.19	26.7
Subsidiary reserves	21.27	0.6
Reserves	24.99	10.9
Total	229.47	100.0

Table A.4b Installed generating capacity in the USA, 1972–1981 (MW)

Year	Electricity industry	Class A and B private companies
1972	399,606	312,067
1973	439,675	345,469
1974	475,888	375,348
1975	508,252	394,850
1976	530,999	414,265
1977	557,012	434,467
1978	579,157	447,447
1979	598,298	456,619
1980	613,546	478,268
1981	634,808	480,325

approval. When deemed to be in the public interest and at the request of an electric utility or state agency, the FERC could order electricity system interconnection and energy sales. Power pool agreements, intra-pool energy exchange prices and specifications, as well as the rules on internal pool organisation had to be approved by the FERC. The commission was the key federal regulatory body for electricity ratemaking (Table A.5).

- Securities and Exchange Commission (SEC) has jurisdiction over holdings (a holding owns several electric utilities, which it usually operates jointly) and regulates their organisation. This commission is also empowered to establish financial regulations for private electric utilities in areas such as ownership structure, bond issues, mergers and asset control.
- Economic Regulatory Administration (ERA) is a Department of Energy agency that regulates electric energy imports and exports, the procedures to be followed in emergency situations, voluntary co-ordination among electric utilities and long-term electricity industry planning.

Table A.5 Private electricity company revenue structure, 1981 [6]

Sale ratios for domestic users	
Percentage of customers	88.2
Percentage of revenues	31.8
Percentage of sales in kWh	26.4
Average yearly sales in kWh per customer	8,277
Average yearly turnover per customer ($)	513.11
Average revenues per kWh sold ($/kWh)	0.0629
Sales ratios for commercial users	
Percentage of customers	10.9
Percentage of revenues	25.9
Percentage of sales in kWh	21.3
Average yearly sales in kWh per customer	53,987
Average yearly turnover per customer ($)	3,393.40
Average revenues per kWh sold ($/kWh)	0.0629
Sales ratios for industrial users	
Percentage of customers	0.5
Percentage of revenues	26.5
Percentage of sales in kWh	32.1
Average yearly sales in kWh per customer	1,674,392
Average yearly turnover per customer ($)	71,390,56
Average revenues per kWh sold ($/kWh)	0.0426

- Environmental Protection Agency (EPA) establishes rules on the environmental impact of electric power generation and transmission.
- Nuclear Regulatory Commission (NRC) regulates the construction and operation of nuclear plants through its licensing process.

State

The powers of state public utilities commissions (PUCs[4]) with jurisdiction over the electricity industry, including regulatory development and enforcement, depended largely on the laws in force in each state. As a general rule, their chief functions included regulating: retail electric power rates, control over bond issues and environmental protection. Of the 47 state commissions with authority to regulate private electric utilities' retail power sales, 16 were also authorised to regulate public and municipal companies and 25 to regulate co-operatives [4]. In such cases, the state commission was authorised to require approval prior to the entry into effect of new rates, suspend rate modifications and initiate audits of rates in place.

[4] *Public Utilities Commission*: state public service regulators in the USA. They are competent in the area of end consumer rate design within their respective states and are largely independent of federal regulators.

Generally speaking, state commissions only authorised rate rises when a utility's net operating investment (rate base) increased or when operating costs took an upward turn. Rights issues and long-term debt generally had to be approved by the respective state commission, which had to be persuaded that the investment in new facilities was justified and the company's subsequent financial situation acceptable. The decision-making procedure included public hearings.

Local

Local level regulation of the electricity industry was uneven, ranging from nil activity to strict control of aspects such as the environmental impact of electric energy transmission and generation or the rates charged for electric power by municipally owned utilities, local public companies and co-operatives.

Reference [2] contains a detailed discussion of the specifics of electricity industry regulation in the USA in those years. A critical analysis can be found in [7, 8].

A.1.4 Electric Power Rates: Key Questions

In the USA, as in many other countries with a privately owned electricity industry, a series of general criteria was followed to determine "fair" electricity rates. Further to these criteria, some of which may clash in practice, electric power rates were supposed to [9]:

- Compensate electric power suppliers for the service provided and expenses incurred.
- Equitably distribute costs among all users, as far as practical given the limitations of metering and similar facilities.
- Provide a reasonable return on capital and attract new resources to finance any new facilities needed to cope with the demand growth.
- Reward service quality and system operating efficiency.
- Promote revenue stability over time to facilitate planning for user and utility future.
- Be sufficiently simple for utilities to readily apply them and consumers to readily understand them.

Ratemaking can usually be broken down into two stages: obtaining the average rate and formulating the rate structure for each type of user. US regulatory commissions traditionally focused on determining the total revenues a utility was to receive for energy sales, along with its cost of service, to attempt to ensure that it obtained a reasonable—neither excessive nor insufficient—return on its capital. This process typically comprised the following steps: (1) establishment of the rate of return on capital and the rate base; (2) determination of the costs incurred: operation and maintenance, depreciation, return on capital and taxes; (3) computation of the average rate level that would ensure the recovery of these costs. This is known as the revenue requirements procedure. Regulatory commissions also usually took an

interest in the question of the allocation of total costs to different types of users: domestic, industrial, commercial and so on; nonetheless, as a general rule they did not actively exercise their regulatory powers in this connection.

The revenue requirements method for determining cost of service was applied to obtain both retail and wholesale rates. Nonetheless, the rates for a number of wholesale energy transactions were set by other *ad hoc* procedures tailored to each type of transaction (see Sect. A.4) in a manner such that the above costs were not all fully reflected in the selling price, or at least not with their full impact. This was the case for transactions that involved no present or future obligations for the utility comparable to the obligations inherent in its own load, in particular with regard to maintaining suitable reserves to guarantee its short-and long-term commitments. Emergency situations were a typical example of this type of operations, and so-called economic transactions another. The latter consisted in the exchange of unconditionally interruptible power between two companies whose marginal costs of generating electricity differed. The duration of such exchanges was usually 1 hour, during which time the purchasing utility had to maintain sufficient reserves of its own, given the interruptible nature of the transaction. Conversely, the wholesale rates for the sale of power to distributors with no generating capacity of their own or for firm power sales were derived directly from the cost of service.

The chief issues to be addressed in obtaining rate levels and structures for direct to user sales are discussed in the following two sections.

A.1.4.1 Rate Level

The rate level determined the extent to which revenues covered operating expenses, provided for a return on invested capital and were able to attract new funding. In practice, determining the cost of service was an extremely complex exercise, since a given level of revenues could be calculated in any number of ways. The most controversial aspects of this process were as follows (see [7] for instance):

- Use of actual current or estimated future values (fuel or capital costs for instance) to prevent rates from lagging behind real costs (regulatory lag), a problem that was intensified by the duration of the regulatory process.
- Establishment of a fair return on shareholder equity.
- Valuation of assets on the basis of their historic cost (usual practice in the USA), replacement cost or at a “fair value”.
- Inclusion or otherwise of the fixed assets in progress in the rate base.
- Accounting treatment for the deferred taxes generated when accelerated depreciation methods were applied.

A.1.4.2 Rate Structure

The rate structure determined the way that the cost of the overall service was distributed among different users. There were countless numbers of possible rate structures. The criteria traditionally followed when establishing rate structures are set out below:

- Simplicity, ready comprehension, public acceptability and practicality.
- Non-controversial interpretation.
- Provision of appropriate revenues.
- Revenue and rate stability over time.
- Equitable distribution of costs among users.
- Effective adjustment of the rate structure to the actual cost of providing each user with service, so rates guide consumers towards the economically optimum use of the existing resources, both as regards how much total energy is consumed and when it is consumed.

The traditional procedure for determining retail rates for electric power in the USA comprised the following steps (see [3, 9–11]:

1. Type classification of users with similar supply requirements and similar contribution to total cost. Only a short number of types were defined, typically including domestic, commercial and industrial users. The latter were divided, in turn, into small- (typically commercial) and large- (typically industrial) scale users, street lighting and government, railways and wholesale sales.
   These categories could be further subdivided by supply voltage or other considerations, such as facility reliability (see Sect. A.3).
2. Breakdown of total service cost, which included both yearly system operating expenses and rate base costs, into four items: generation, transmission, distribution and tangible fixed assets, in accordance with very precise criteria [3, 10, 11]. Only the costs related to these four items were broken down.
3. Breakdown of each of the above items into three categories: demand component (costs associated with installed capacity), energy component (costs associated with the amount of kWh produced) and customer component (costs associated with the number and type of users, irrespective of their consumption figures). Very precise criteria were also used for this breakdown.
4. Division of the sum found for each of the items described in the preceding paragraph among the different types of users. This could be readily calculated in the case of the customer component, in light of the definition of the term; the energy component was distributed among the various types of consumer in proportion to their respective consumption in kWh, although certain refinements, such as line loss for each type, were often accommodated as well. Several methods were proposed to distribute the demand component

[3, 9–11], but none were regarded to be wholly satisfactory. In its Cost Allocation Manual, the (NARUC) specified no single method to the exclusion of the others, although it did suggest that the coincident peak load method (see Chap. 3) should be applied whenever possible. Reference [12] recommends a specific distribution method for each item.

5. Establishment of the rate structure for each type of user from the information obtained in the preceding steps. A compromise was sought between the objective of adapting rates to the actual cost structure and certain practical considerations such as the intrinsic limitations in metering and billing systems. Reference [12] gives a historic description of the different rate structures used in the USA. During the years studied here, the one most widely used for domestic consumers was the decreasing block structure, in which the price of each extra kWh decreased by steps as total consumption increased; this rate usually included a fixed monthly charge unrelated to consumption; the variable charge essentially covered the capacity and energy components. The rate used for commercial and industrial customers had separate energy (usually in the form of decreasing blocks) and demand (depending on the maximum demand used, as well as the load factor in some cases) components, in addition to the fixed user charge.

The traditional rate structure described above was widely criticised for not suitably reflecting the variations in the real price per kWh in different seasons of the year and/or times of day [10, 11, 13, 14]. Some American electric utilities [10, 11, 15] had already modified their rate structures to take these variations into account, adding a further level of complexity to the above procedure.

A.2 Electric Power Rates: Determining Cost of Service

A.2.1 Introduction

This section describes the procedure used by the US electric utilities in the early 1980s to determine the average cost of supplying electric energy, known as "cost of service". The average cost of service was obtained by dividing the electric utility's total costs by the amount of energy supplied. Total costs essentially included operating and maintenance costs, return on capital, depreciation and taxes. As indicated in Sect. A.1, the cost-of-service concept was only applied directly to energy sales that could be regarded to constitute the selling company's own load, i.e. its direct sales to end users (or retail sales) and distributors with no or only token generating facilities of their own. Power supplied in other wholesale

energy transactions and the related revenues were subtracted from the respective values of energy and total costs before computing the average cost of service. Cost of service constituted the grounds for determining the rate structures for both retail sales of electricity and sales to non-generating distributors.

American companies used the revenue requirements method for determining cost of service. This procedure is described in numerous papers, such as [16–19]. Countless variations arose, however, in the practical implementation of the method by electric utilities and regulatory commissions (see [7, 20]).

A.2.2 The Revenue Requirements Method

The core concept in the revenue requirements method is that cost of service must be such that it enables the company to recover all its costs as well as a fair return on its capital. The question of what exactly was understood to be a fair return was obviously intensely debated, and will be turned to later; see [16, 17] for an in-depth discussion of the issue.

In mathematical terms, revenue requirements (or total cost of service) can be expressed as follows:

$$\begin{aligned}\text{Total Cost of Service} = {} & \text{Operation and Maintenance Cost} + \text{Depreciation} \\ & + \text{Tax} + \text{Rate of Return} * \text{Rate Base} \\ & - \text{Additional Revenue}\end{aligned}$$

and the average cost of service is the quotient found by dividing the total cost of service by the power sales to which cost of service is applicable (as described above).

A succinct definition of the terms of the above equation follows. See the American standards on systems of accounts for electric utilities [21] for further details.

- Operating and maintenance costs: the cost of fuel, material and replacement parts, energy purchases, supervision, personnel and overhead.
- Depreciation: the linear method was generally used.[5] Fixed assets in progress were not depreciated
- Tax: all the taxes for which the utility was liable, i.e. on profit, revenue, property, social security and construction (except as owing to fixed assets in progress, since such tax was added into the value of the asset).
- Rate base: net fixed assets (plants, transmission and distribution facilities, other tangible and intangible fixed assets and nuclear fuel, less the cumulative depreciation for all of them), plus current assets (fuel and other material and

[5] See the comment on standardised accounting below.

replacement part inventory, advance payments and deferred revenue, research and development expenses and current asset requirements).
- Rate of return: average weighted interest rate on the company's long-term financial resources (bonds, debentures, shares and preferred shares).
- Additional revenue: expenses/revenue deriving from the sale of the company's property, revenue from wholesale energy sales and other revenue not directly related to the production of electric power.

A.2.3 Discussion

The regulatory commission could question any of the sums presented by utilities when determining cost of service. If the commission decided that a given expense was excessive or unsubstantiated the respective amount was eliminated from the above equation.

The rate base was defined to encompass current assets and net fixed assets, with the latter accounting for the major share of the base. Net fixed assets were evaluated to be the sum of the non-depreciated property used for company operations. The treatment for net fixed assets varied widely from one regulatory commission to the next. The essential questions were: what should be included, when it should be included and at what value. In assessing fixed assets it was common practice to use the historic or original value, which was usually much lower than the current replacement value; certain regulatory commissions allowed utilities to use the replacement cost (computed from extrapolations based on the costs actually incurred [22]), or a "fair" value, which each commission established at its discretion. One particularly controversial area was the inclusion or otherwise of fixed assets in progress in the rate base. When inclusion was not permitted, no provision was made for remuneration for the fixed assets in progress, and the utility was authorised to include the respective interest costs of the capital invested in the works in progress item. If all or part of the fixed assets in progress was included in the rate base, the interest on the part in question was not charged. The trend in the years discussed here was for a growing number of regulatory commissions to allow the inclusion of the fixed assets in progress, to avoid sudden hikes in the rate base, undue interest fund growth during construction and excessive haste in commissioning facilities [7]. A number of different procedures were in use to appropriately account for variations in the rate base throughout the year.

A peculiar American accounting practice known as standardised accounting lay at the root of another variation in the way the rate base was computed. This procedure consisted basically in keeping fictitious parallel accounts used exclusively to determine the company's tax liability each year: instead of the usual linear depreciation used in the real accounts, these parallel accounts provided for accelerated depreciation to reduce taxes in the earlier years and increase the liability in subsequent years. This deferred tax accumulated in a fund.

Most American electric utilities used standardised accounting. A thorough discussion of the effects of this practice on the determination of cost of service and the tax treatment can be found in [23].

One consideration of enormous practical importance for utilities in the reference period was the time that frequently lapsed between when a rate entered into force and when it was actually applied. This is what was known as "regulatory lag" in the US terminology, and arose as the combined effect of two factors: the practice of calculating rates on the basis of the cost and energy sales figures prevailing when an application for a rate change was submitted, and the relatively long time lapsing between that date and when the regulatory commission's resolution was forthcoming. The procedures in place to alleviate this problem included:

- Use of estimated values to cover the time expected to lapse before the new rates would come into effect. This practice was still uncommon at the time, however; regulatory commissions were much more prone to allow slightly higher rates to implicitly compensate for this shortfall [7].
- Clauses on automatic adjustment of rates to accommodate changes in fuel costs. Critics of this procedure sustained that it discouraged companies from seeking less expensive alternatives when fuel prices rose.
- Adjustment of the rate base as new facilities became operational.

The bone of contention in the ratemaking negotiations between electric utilities and regulatory commissions was the determination of the long-term rate of return on the company's financial resources (bonds, debentures, preferred shares and shareholder equity: normal shares plus reserves). Given that the average interest rate on debt (bonds and debentures) and the average rate of dividends on preferred shares were in fact known in advance, the problem was merely a question of determining the rate of return on common equity. The general legislation in effect[6] at the time provided that the rate of return on common equity should (a) be comparable to that of other investments with similar risks and (b) be sufficient to inspire confidence in the soundness of the company's financial position so it could raise new capital when necessary. Much has been written about this important question in ratemaking: see [16, 17] for a detailed discussion and [2, 9] for a briefer review. In practice, regulatory commissions would set this rate after hearing the opinion of different experts and considering issues such as the method adopted to assess the fixed assets, the estimated lag between the period when rates were to be applied and when they were calculated, the inclusion or otherwise of fixed assets in progress and so on. The model used to quantify the cost of a utility's own capital on the grounds of the risk associated with the various types of businesses is known as the Capital Asset Pricing Model (CAPM).

The method most commonly used in the period considered was WACC or Weighted Average Cost of Capital.

[6] Supreme Court, Hope Natural Gas Case, 1944.

A.3 Electric Power Rates: Structure

A.3.1 Introduction

Electric utilities can supply energy to other companies and/or directly to their customers. Both types of sales can be subdivided, in turn, into categories: end user or retail sales may be classed as domestic, commercial or industrial, depending on consumer characteristics, and wholesale or inter-utility energy exchanges as firm power, economic exchange or emergency sales, depending on the type of transaction involved. Rates were established for each of these categories under the supervision of the respective regulatory commission. This section describes American utility ratemaking procedures. The ground rule in this process was coverage of the total costs of supplying electric energy, calculated as described in the preceding section. The emphasis in the present section is on the methods for distributing these total company costs among the various power sales categories (users, distributors, inter-utility exchanges) and, as appropriate, among the various classes of users or customers (domestic, commercial, industrial and so on). Costs were broken down as far as necessary (into the three components—demand, energy and customers—on the basis of the voltage at which power was delivered, as well, perhaps, as by periods—peak, flat and off-peak—, depending on load levels) to obtain all the elements required to design the final structure for each type of rate. The criteria used in inter-utility energy exchanges are introduced in the following chapter.

As noted in Sect. A.1, the determination of rate structures from the average cost of supplying energy (cost of service) involves two essential tasks: the first, called cost allocation, consists in distributing each of the items comprising the total cost of service among the different categories of power sales (and classes of customers), and is described in Sect. A.3.2. The second is the final determination of the rate structure for each type of service, in keeping with the objectives set out in Sect. A.1.4.2, to formulate fair rates. Section A.3.3 is devoted to this second task.

Generally speaking, American electric energy rates for each type of service or class of customer were uniformly structured; i.e. no distinction was made between peak and non-peak times. Rates were based on average embedded costs, which yielded a single price per kWh for the entire year, despite the fact that the cost of supplying electricity varies with the time of day and season of the year.[7] Authors critical of such traditional rates [13, 14, 24, for instance] sustained that prices that do not reflect the actual cost structure lead to an economically inefficient use of electricity. A number of American companies [see 11, 15, 25] had begun to adjust

[7] This notwithstanding, the rate for each class of service was designed to ensure that it contributed proportionally to cover the cost of system peaks through the (explicit or otherwise) demand component of the rate.

their rates to take account of such time-related variations in cost (marginal costs), while keeping average rates equal to average costs (a method is known as "peak load pricing"). In practice, the outcome was the use of two separate schedules distinguishing between summer and winter or peak and non-peak periods. Under a purely marginalist approach, rates would have been established in accordance with actual marginal costs, and that in turn would have upset the balance between average rates and average costs.

The increasing importance attached in the USA to establishing rates that would reflect the real differences in electric power costs at different times of day and/or seasons of the year was addressed in Congress's enactment of the *Public Utilities Regulatory Policies Act* (PURPA) in 1978 and in several utilities' business practices. In addition to introducing the standard traditional approach, this section describes how costs were broken down and rate structures formulated to reflect different demand conditions. This discussion is based on the literature reviewing the practices generally followed across most of the US electricity industry [3, 9–11, 24–26].

A.3.2 Cost Allocation

The traditional procedure for assigning costs to types of services rested on three different blocks of information: (a) the company's accounts, pursuant to the FERC uniform system of accounts [21] to obtain the utility's total cost of service broken down into the usual items; (b) the types of services among which costs were to be distributed, which might fall under the jurisdiction of several different bodies (FERC, state regulatory commissions), types of power sales (consumers, energy exchange) and types of customers; (c) the electricity system operating data needed to partition costs among the different types of service; in particular, the contribution made by each type of service to total system demand, along with certain characteristics of each type of service, load factor and demand factor among them.[8]

The discussion below describes the cost allocation methodology recommended by the National Association of Regulatory Utility Commissioners (NARUC) [3], which consisted in the following: (a) functional cost classification, (b) cost classification by demand, energy and customer components; (c) allocation of customer, energy and demand costs to types of service; (d) allocation of total costs to types of service.

Functional cost allocation

Functional cost allocation involved consolidating the partial costs comprising the total cost of service into a short list of items that could be associated with the chief functions of electric power supply, namely, generation, transmission and

[8] Load factor: average demand in a given interval/maximum demand in the same interval.

distribution. This classification facilitated subsequent cost allocation to the different types of service.

The generating function included all the costs associated with the generation or purchase of electricity at power plant bars and the delivery of this energy at connection points with neighbouring utilities. All costs associated with power transmission within the utility's system as well as to and from other companies were classified under transmission. Distribution, in turn, covered all costs associated with the transfer of energy from the transmission grid to consumers over the distribution grid, as appropriate (for certain types of service and consumers, power was supplied directly from the transmission grid).

Certain costs not directly classifiable in any of the above three functions, such as overhead, other tangible fixed assets, accounting and financial costs and so on were classified by associating them with other more readily classifiable costs and processed following the same method.

The costs attributed to a given function might in turn be subdivided to facilitate their subsequent allocation to one type of service or another (dividing the grid into primary and secondary distribution, for instance).

Cost classification into demand, energy and customer components

In the next step in the final allocation of costs, the items resulting from the breakdown described in the preceding section were reclassified into the three components—demand load, energy consumption and number of customers—that represented quantifiable characteristics of each type of service.

The general classification criteria used were as follows: the demand component depended on the investment in facilities and therefore tended to remain constant in the short term, regardless of the amount of energy actually generated and/or transmitted in the system. The energy component tended to vary directly with the amount of power generated and/or transmitted. The customer component tended to vary with the number of customers serviced. Strictly speaking, this classification and the use subsequently made of it could only be justified if each group of total costs (demand, energy and customer) was assumed to vary linearly with the respective parameter (demand level, energy consumed and number of customers) and independently of the other two. But this was a mere approximation of what actually occurs. The evolution of economic theory in the interim has shown that such assumptions were mistaken.

Following these rules entailed including metering, billing and connection costs, plus a percentage of distribution costs (including capital costs and costs of maintaining a minimum distribution grid [3], in the customer component. The energy component included energy purchases, fuel and generating costs, maintenance costs—which varied with the amount of power generated/transmitted—and even the operating costs (including depreciation) that depended more on use than time. The demand component included the costs of capital and depreciation as well as the taxes associated with generating plants, transmission lines, substations and the part of the distribution grid not included in the customer component.

Again, a detailed description of the criteria underlying this classification may be found in reference [3]. One subject of special interest was the determination of the customer cost component in connection, for instance, with the distribution grid, where no obvious way had been found to separate this component from the demand component.

Allocation of demand, energy and customer costs to types of service

Once the figures on total consumption by type of service and as a whole were found, the total energy costs could be readily allocated to each type of service (assuming that the price per kWh did not depend on the time of day or season of the year when energy was consumed).

Total customer costs were allocated to each type of service on the basis of the number of customers in each type of service, weighted by suitable correction factors to reflect customer differences by type of service [see 3, 10].

The allocation of total demand costs was the most controversial item in cost allocation due to the difficulty in finding a workable procedure that fittingly evaluated the contribution of each type of service to system demand. The three most characteristic methods are defined briefly below: further information may be found in [1, 3, 10, 26]. In the coincident peak load method costs were distributed among the types of service in proportion to their respective demand at annual system peak demand. NARUC [3] suggested this method wherever it could be feasibly applied: average monthly peak demand values had to be used for systems that peaked more than once a year.

In the non-coincident (or more accurately, not necessarily coincident) peak load method, costs were partitioned in proportion to annual peak demand for each type of service. This method yielded stable distribution coefficients and could be implemented with cheaper metering equipment than required in the preceding scheme; this was the procedure of choice in the USA. It met with many an objection, however, for ignoring the effect of coincident use on demand [3, 26]. In a third method, the average and excess demand method, allocation was based on proportionally distributing only the costs incurred to meet the average demand for each type of user. All other costs were distributed in accordance with the second method and in proportion to the difference between the peak and average demand for each type of service.

Regardless of the method employed to allocate demand costs, the demand values used had to be referred to a single reference point on the grid. Consequently, considerations such as voltage and line loss had also to be factored into the equation [3].

Allocation of total costs to types of service

The total costs allocated to each type of service, regulatory jurisdiction and, as appropriate, class of customers, were obtained from the results found as described in the preceding section.

A.3.3 Establishing Rate Structures

While NARUC and other institutions attempted to standardise cost allocation, it seems that there were no guidelines for establishing specific rate structures.

Nonetheless, the cost itemisation described in the preceding section provided a rationale for establishing one rate structure for each type of customer based on the three cost components—demand, energy and customer—, where the unit costs of each component were obtained by dividing the total yearly costs by yearly peak load (kWh), energy supplied yearly (kWh) and number of customers, respectively. Naturally, provision had to be made for the demand factor for the class of customer in question to apply the unit cost of demand. Such a correction could be made [9], because there is a known empirical relationship between the load factor (which can be measured) and the demand factor: the product of peak demand times the demand factor was billed at the unit cost of demand. The total cost of demand was found to be more conveniently expressed in terms of the individual customer's load factor. The result was approximately comparable to a load rate schedule in which costs decrease depending on the load factor [9].

The above rate structure was ill-suited to residential customers, since it would call for individual metres able to record peak demand. Domestic demand was sufficiently even and stable for the other demand characteristics to be deduced from the total energy consumed. In other words, the demand and energy components of cost could be treated jointly as a fictitious "power" component. The result was the typical declining block rate for domestic users [9].

A.4 Electricity Rates: Wholesale Power Sales

A.4.1 Introduction

This section describes the different types of inter-utility power sales then in use in the USA, along with the criteria most widely used by companies to establish the respective rates. It also summarises the formalities and procedures that utilities followed to obtain FERC approval of these rates.

As discussed in Sect. A.1, wholesale energy transactions were divided into two major groups: sales to distributors, primarily municipal government-owned utilities and co-operatives; and co-ordination sales among electric utilities, most of which took place in the context of power pools. The rates for sales to distributors were established in the same way as retail rates, since such sales formed a part of the seller's own load. The rates for co-ordination services ranged from long-term firm power sale rates, which included all a company's fixed costs,

to economic energy exchanges, which took account of incremental variable costs only.

Reference [5] provides an interesting description of co-ordination transactions among electric utilities in the USA, and Ref. [19] an account of the procedures for applying for approval of wholesale rates. All of these issues are discussed below.

A.4.2 Classification of Wholesale Sales

Wholesale sales were divided, first, into sales to distributors and energy co-ordination transactions. Sales to distributors formed a part of the seller's own load, although they could on occasion be assigned lower priority; nonetheless, utilities were required to keep sufficient reserves for such sales and make provision for possible growth when planning their generating and transmission systems. In co-ordination sales account was taken of variations respecting supply reliability, transaction timing, buyer's and seller's reasons for undertaking the transaction, and naturally the rates for each service. The rest of this section is devoted to such co-ordination services.

The FERC [4] classified co-ordination sales as follows:

- Long-medium-and short-term firm power.
- Generating unit capacity.
- System capacity.
- Diversity interchanges.
- Reserves.
- Maintenance.
- Emergency.
- Financial transactions.
- Conservation of fuel.
- Firm transmission.
- Non-firm transmission.
- Run of the river.
- Hydroelectric storage.
- Hydroelectric system co-ordination.

Many of these transactions were associated with the establishment of power pools. Firm power purchases were usually concluded to offset buyer energy shortfalls and inability to meet the obligations inherent in pool membership.

The following is a description of the types of services provided under each of the above items.

Firm power

Under firm power contracts the seller committed to supplying a certain amount of energy during the period specified in the contract. The seller had to maintain the necessary reserves, since this service could only be interrupted under narrowly

restricted conditions. Depending on the duration, these agreements were divided into long term—over 1 year—, medium term—from 1 month to 1 year—, and short term—from 1 day to 1 week—energy sales.

Generating unit capacity
This consisted in the provision of capacity service and the associated energy from a specific generating unit owned by the seller. It granted a contractual right over part of the production of a given unit, but with no share in its ownership. The reliability of this service was contingent upon the availability of the unit specified in the contract.

System capacity
This consisted in a given amount of capacity (without reserves) and/or energy, which was to be supplied, under contract, by the seller's system as a whole or a specific group of generating units. Reliability was higher than under generating unit power arrangements, but lower than in firm power agreements.

Diversity interchange
Capacity and/or energy interchanges between systems whose demands peaked at different times or whose operating costs and/or generating availability were timed differently. These were reciprocal agreements with firm power commitments; therefore, the supplier had to take the necessary measures to ensure availability at the times programmed for interchanges.

Reserves
These were agreements to share the reserves established, explicitly in pools and explicitly or implicitly in other contexts.

Maintenance
Capacity and/or energy supplied to a system to supplement its reserves during programmed maintenance. The terms of such agreements were specific to each contract and services were typically co-ordinated 6–12 months in advance, under firm power sales arrangements.

Emergency
The energy supplied to a system to counter a sudden and unexpected power shortfall. Contracts for such services often contained clauses that established reserve levels for each system, since these services were reciprocal. The duration ranged from 24 to 72 h, after which, if the system experiencing difficulties was still unable to meet demand, service could be continued or otherwise at the seller's discretion. In the event, sales were reclassified as short-term firm power.

Economy transactions
Unconditionally interruptible power supplied for a given period, usually 1 h, in which the seller's incremental costs were lower than the buyer's. The latter had to maintain sufficient reserves, given the interruptible nature of the service.

Conservation of fuel
Energy supplied to refrain from utilising units whose fuel was subject to government-mandated supply constraints. The purpose was to solve a material problem, such as a fuel shortage, not a financial one such as increases in fuel prices.

Transmission services
These services were provided at one of three levels of reliability.

- Firm: non-interruptible, except where system security was at stake.
- Conditionally interruptible: interruptible only under the conditions specified in the contract.
- Unconditionally interruptible: interruptible at the discretion of the utility providing the service.

Wheeling services were and are transmission services provided by a company that neither generates nor consumes the energy sold.

Run of the river
Sales of energy generated by hydroelectric systems in which water cannot be stored.

Hydroelectric storage
Energy interchange agreements that enabled a company with hydroelectric plants to buy energy to store water in reservoirs at times when energy prices were low and run their turbines at times when they rose.

Hydroelectric and conventional steam co-ordination agreements
Agreements under which a steam generation system received surplus hydroelectric power from another system. Energy was fictitiously "stored" in the sense that the hydroelectric system received an equivalent amount of energy at a later time, the next day for instance.

The classification of transactions by type to determine the rate applicable as described in the following section were negotiated by the companies involved prior to application to the FERC for rate approval.

A.4.3 Wholesale Rates

As noted in the introduction, the rates for the sale of power to distributors were computed in the same way as retail rates, following the criteria set out in Sect. A.3.

Co-ordination sales typically comprised three elements: a demand component, an energy component and a surcharge on the latter.

The *demand component* was found in firm power, conditionally interruptible (system energy, maintenance and conservation of fuel) and generating unit power sales. The rationale for this component was the need to recover the seller's fixed

costs (capital costs, depreciation, taxes and fixed maintenance costs) for providing these services.

This component could be readily calculated for generating unit power sales as the yearly fixed cost of the unit, weighted by its availability.

For long-term firm power sales, the demand component was calculated with the system average fixed cost method i.e. by dividing the utility's yearly generating and transmission system fixed costs by system peak load. The underlying assumption was that the service was provided by all the seller's generating units.

For the rest of the services that included the demand component in the rate (medium- and short-term firm power, system energy, maintenance and conservation of fuel), the scheme most commonly used was to determine the share in the weighted average cost of capital. Under these arrangements, the demand component was calculated by multiplying the annual fixed unit costs of the generating units that it was assumed would supply the service, times the number of kilowatts expected to be generated for these services. Under this method the assumption was that energy would be supplied from certain "marginal" units, since these services had lower priority than the seller's own load, and therefore would be provided with less efficient generating units which normally had lower fixed but higher variable costs. The ultimate outcome of this method was a smaller demand component but a larger energy component than would be found for long-term firm power services.

The maintenance services addressed here that included the demand component in the respective rates were services provided outside pools. Inside such pools, as explained in Sect. A.5, the agreements governing such services were designed, among other things, to co-ordinate maintenance programming for each of the systems forming the pool; the rate applicable to the demand and energy supplied while each utility's generating equipment was undergoing maintenance varied depending on whether or not such maintenance was programmed in keeping with pool objectives.

The *energy component* was present in all co-ordination sales and its purpose was to cover the variable costs incurred by the seller to provide the service. Fuel accounted for the bulk of these variable costs and dispatching and administrative costs for the rest. Plant start-up and maintenance costs had also to be included in the energy component in some cases.

The basis for calculating this component was normally the seller's incremental costs, defined by the FERC to be the "costs that would not have been incurred but for the transaction" [5].

Surcharges were also to be found in all co-ordination sales, and were added to the incremental costs in the energy component. Utilities used three types of surcharges: fixed, percentage and "share in savings".

The "share in savings" surcharge was only present in financial transactions. Its purpose, to provide an incentive for such transactions, was completely different from the purpose of percentage and fixed charges, namely to attempt to recover the seller's incremental costs, which were difficult, not to say impossible, to quantify [5].

The share in savings surcharge was calculated as the average difference between the seller's and buyer's incremental costs.

Percentage surcharges met with widespread opposition, because it was believed that the costs generated did not grow at the same pace as the cost of fuel, which constituted the bulk of incremental costs. The FERC subsequently imposed ceilings on percentage surcharges for transmission and conservation of fuel services.

Transmission costs, in turn, were generally included in the overall rate for wholesale service. The method used to calculate these costs is given below:

$$\text{Firm transmission (\$/kW)} = \frac{\text{Total transmission system costs}}{\text{Peak load}}$$

$$= \frac{\text{Conditionally interruptible transmission (\$/kW)}}{\text{Total transmission system costs}}$$

$$\frac{\text{Total transmission system costs}}{\text{Interconnection line capacity} - \text{Transmission system capacity}}$$

$$\text{Unconditionally interruptible transmission (\$/kWh)} = \frac{\text{Total transmission system costs} - \text{Overhead}}{8760 \times \text{Peak load}}$$

Finally, there was no standard rate for hydroelectric storage and hydroelectric system co-ordination services, which depended, rather, on individual agreements and was usually calculated from established formulas.

A.4.4.4 Application for Approval of Wholesale Rates

Rates for new services or modifications of the existing rates were subject to FERC approval. Utilities applied for such approval between 120 and 160 days in advance of the date scheduled for the service to begin. This could give rise to two different situations: (1) claims or objections on the part of stakeholders (purchasing companies, states where the utilities were located, and so on), which were lodged more often than not when sales to distributors were involved and (2), no objections were raised, which was usually the case in connection with co-ordination sales. When claims were filed a hearing was held in which all the parties involved had to submit proof to substantiate their positions. The final decision was made by the FERC. When no claims were filed the commission studied the applications and granted or denied approval based solely on the documents submitted by the seller.

The documentation that the seller had to submit with the application was classified into four groups:

1. General

Scheduled date for implementation of service, sales agreements established, reasons to change or create the rate, customers affected.

2. Effects of the rate

Comparison among past transactions and those envisaged under the new rate schedule, and comparisons with other schedules for similar services.

3. Accounting and cost of service information

Cost data and other factors explaining the rate requested.

4. Rate structure

The volume of documentation required for modifications was much more extensive than for creating a new service, but the commission tended to regard all applications as modifications. There were two procedures for applying for modifications: full and abridged.

Rate applications for co-ordination sales usually took the abridged route, since these services had been negotiated in advance and an agreement already reached on the rate; all the FERC actually did in such cases was to give its consent to these agreements.

A.5 Power Pools

A.5.1 Introduction

A power pool may be defined to be a group of two or more utilities that co-ordinate their operation and planning to minimise operating costs, save on fuel and increase the reliability of the electricity system. Such pools were common in the USA in the 1980s, when there were around 30. They held a substantial share in the American system, accounting for 38 % of the total electricity generated in the country [27].

In order to materialise the benefits of membership in a power pool, utilities established procedures and fostered action that provided for: an equitable distribution of participants' obligations and benefits, shared use of the transmission grid and plants, co-ordinated operation of the power pool and establishment of the prices for energy transactions.

The creation and/or operation of power pools often entailed trade-offs for member utilities in the form of the loss of company independence inherent in pool membership. The obligation incumbent upon pool members to co-operate might also clash with their competitive positions (generally in terms of wholesale energy sold to third parties). Consequently, it was up to each utility to weigh these drawbacks against the advantages of greater co-ordination with other pool members.

The types of agreements in place in pools across the USA varied from informal inter-utility arrangements to formal agreements among all the companies in a group, as well as bilateral and multilateral agreements. The first type required no

legal approval. Bilateral and multilateral agreements involved wholesale energy sales and were subject to approval by the (FERC), which regulated energy exchanges, transmission rights and energy payments. Finally, member companies might also conclude formal agreements that governed the operation of the pool as a whole and each member's responsibilities. These agreements were also subject to FERC jurisdiction. The relationship between each individual utility and the regulatory commission in its respective state was not affected by membership in a power pool. Therefore, the rates for the sale of electric energy to each utility's customers were established in keeping with the criteria described earlier in this document. As far as electricity rates were concerned, then, the presence of power pools added a new and fairly complex framework for negotiating inter-utility wholesale transactions to the scenario described above.

In short, a power pool should be viewed as a series of electric utilities with separate rates for their customers (in all respects: relationships with the respective state regulatory commissions, rate levels and structures and so on) that derived mutual benefit from co-ordinating their operation and planning activities.[9] A list of the possible benefits that such co-ordination should afford [28] in a pool with a maximum degree of integration would include:

- (Shared) saving of operating costs through energy transactions for financial reasons. Savings were maximised when the pool was operated as a single company from the same control centre (with or without satellite facilities in each company).
- Reduction of each company's operating reserves, spinning or otherwise, since the reserves required to deal with contingencies could be shared.
- Benefits deriving from joint operation to exploit the differences in individual companies' load curves (the use of hydroelectric power, for instance).
- More effective response to emergency situations, for co-ordinated handling of such incidents.
- Co-ordinated maintenance programming to minimise the costs of substituting for plants being serviced.
- Lower long-term margin of reserve capacity, increased transmission grid reliability and economies of scale in new facility construction, thanks to better co-ordinated planning across several utilities.

This section begins with a brief description of pool member rights and obligations with regard to the use and maintenance of suitable joint generating and transmission capacity (Sect. A.5.2). This is followed by a discussion of the aspects to be considered when co-ordinating operation and planning (Sect. A.5.3). Finally, a few remarks are devoted to the way prices were established for energy transactions among pool members.

[9] Power pools were created to co-ordinate utility operation and thereby benefit from the savings generated by exchanging energy and sharing reserves. Co-ordinated planning was introduced at a later stage, if at all.

This section has drawn heavily from Ref. [27]. Other references used include [4, 28–31].

Because of the enormous diversity in the degree of integration in the US power pools in and around the 1980s, this section is limited to a discussion of the major issues and general procedures adopted. The specific solutions to each of these issues found by four pools with very different levels of integration can be found in Ref. [27].

A.5.2 Ownership and Use of Facilities: Rights and Obligations

Pool membership rights and obligations in connection with facility ownership and use were concerned, on the one hand, with the use of the transmission grid, and on the other with generating plants and their production in terms of pool needs. Certain aspects of the agreements reached in these areas were subject to FERC review for approval and/or recommendations.

Agreements on the use of the transmission grid could include questions ranging from maintenance specifications, plans for optimum growth and each member's responsibility in the construction of its portion of the grid, to the mere review by a pool committee of each participant's transmission grid planning and cost estimates and their impact on the reliability of the grid as a whole. Member access rights to use the grid for short- and long-term energy exchanges were also defined. In this regard, any of the types of exchanges defined in the preceding section could be the object of such agreements.

The establishment of agreements on the short-and long-term use of the transmission grid was essential to satisfactory pool operation. Short-term conditions were set up in such a way as to guarantee the co-ordinated conducting of pool activities. Long-term grid use rights were defined prior to concluding contracts for firm power sales, shared plant ownership, wheeling and others. One particularly relevant consideration in this context is that when one utility granted another the use of its transmission grid, its own competitive position on the wholesale energy market might be adversely affected. Grid construction, in turn, might prove to be more advantageous for the pool as a whole than the utility concerned. In such cases, compensation or incentives were established so no single company was unduly jeopardised.

Plant-related agreements were reached in the context of a joint capacity expansion plan for the pool, which co-existed with each member's entitlement to implement its own plan to carry its own load, which was merely reviewed by a pool planning committee. Where a joint expansion plan was adopted, it addressed issues such as pool growth forecasts, reliability criteria and planning models that would ensure that overall demand could be met to such criteria.

The right to access the energy produced by plants was defined for both short- and long-term horizons. The transactions cited above would be conducted in the short term and would be open to all pool members. At the same time, arrangements

were made for outages in the event of power shortfalls. Solutions might range from sharing the outage to requiring the member with a power shortage to reduce its load by the amount of the shortfall. Long-term planning involved establishing criteria for each member's share in plant construction and its access to the joint transmission grid.

A.5.3 Co-ordinating Operation and Planning

A.5.3.1 Co-ordinating Operation

Short-term co-ordination among pool members reduced fuel and production costs. Long-term co-ordination allowed for the efficient expansion of pool capacity in terms of the use of capital and energy. System operation-related issues included energy exchanges, reserves, maintenance programming, emergency procedures and hourly dispatching. All these issues are discussed below.

Economic energy transactions

Energy exchanges between utilities with different marginal generating costs were conducted under bilateral agreements with or without a broker or central dispatching, depending on how deeply integrated the pool was.

Bilateral inter-utility agreements facilitated forward plant programming, to save fuel oil, for instance, by using coal-fired or nuclear energy. The transaction price was reached by a method mutually agreed to by the parties and approved by the FERC. Since these exchanges could be terminated when the seller's load increased, the buyer had to have as much energy in reserve as it bought. Such arrangements, therefore, enabled buyers to reduce load but not to shut down plants. Agreements of this nature were readily negotiated and billing was similarly straightforward. They called for no sophisticated control or communications equipment, nor did they entail any loss of autonomy for the parties concerned. The drawback was that it took a fairly large number of such arrangements for all pool members to participate and optimise the benefits of energy exchanges.

Brokering was and is a means for exchanging information that provided all utilities with hour-by-hour energy prices. Energy transactions were conducted in an orderly manner, beginning with the most divergent prices. The broker's role was performed either manually or automatically, i.e. via computer, and ranged from merely displaying prices and facilitating agreements to defining the exchanges each company should conclude for the following year on the grounds of dispatching information. Brokering had relatively low implementation costs and allowed each company to essentially conserve its power of decision. Moreover, the benefits deriving from each transaction were easy to evaluate. It had the same disadvantage as the preceding scheme; however, it would take many bilateral agreements for all pool members to reap all the possible benefits and it could not guarantee optimal operation.

Under central dispatching all the generating plants in the pool were operated as if they formed a part of a single system, to minimise total operating costs. This arrangement was implemented in accordance with the well-known principle of optimum load dispatching: evening off the incremental production costs of plants in operation, after adjusting for line losses.

As a rule, centralised dispatching of all pool plants led to different results than if each company's plants had been dispatched separately. The method for calculating the benefits of economic energy transactions within the pool was based on this difference between the results attained under the two approaches (centralised and individual), which exactly determined the benefits to be derived from all manner of exchanges between pool members. This procedure, which is more complex than the two methods discussed above, was necessary in a context in which pairs of buyers and sellers could not be easily identified. In power pools with centralised dispatching, the comparative analysis of centralised (or actual) and individual (or calculated) dispatching was useful to identify and classify inter-utility exchanges by types (economic transaction, reserves, maintenance and so on) for the intents and purposes of applying the respective rates (see Sect. A.5.4). Under individual dispatching arrangements, firm power purchases and shares in plants were regarded to form a part of a company's self-generation.

Of the three types of energy exchanges discussed, centralised dispatching was the one requiring the most sophisticated control and communications equipment and involving the most complex billing processes and greatest loss of utility autonomy. The advantage was that the benefits generated by energy exchanges were greater under these circumstances, although dispatching investment and operating costs, along with the higher costs deriving from greater administrative complexity, were deducted from such benefits. In other words, investing in a central dispatching facility was only justified where the sum of such costs was smaller than the increase in benefits.

Reserves

For the intents and purposes of operation, a distinction is drawn between spinning (grid-synchronised) and non-spinning (not synchronised with the grid) reserves. The aim of sharing reserves was to reduce the amount of both types of reserves while maintaining the same degree of system security as if the reserves were the exclusive responsibility of each utility. This was the objective regardless of the criterion adopted for establishing reserve capacity requirements.

In pools without centralised dispatching, each company was allocated a share in the total pool reserves in proportion to its maximum demand load. In pools with centralised dispatching, reserves were programmed in accordance with overall criteria of economy and security, regardless of specific plant ownership. The reserves to be contributed by each company were determined by comparing the real results with the (fictitious) results that could have been expected if individual company dispatching arrangements had been in place, as noted above.

Maintenance programming
Another possible area for co-ordination was plant maintenance programming. Appropriate programming would prevent energy reserves from dropping to critical levels when loads peaked. Programming usually covered periods of from 2 to 5 years to take account of nuclear fuel load and supra-annual rainfall cycles, although frequent revisions were required to convert obligatory into programmed outages, as well as changes in facility availability.

One of the most difficult tasks in joint maintenance programming was to persuade pool members to agree to it. Financial incentives could be used to compensate companies that were obligated to overhaul their plants when it was most cost-effective for the group, but not for them, to do so.

Emergency procedures
These procedures included action such as interrupting supply, lowering frequency and reducing voltage. Co-ordinating these measures maximised reliability and minimised the effects of contingencies. Pursuant to North American Power Systems Interconnection Committee (NAPSIC) guidelines, each pool had to establish its own internal rules of procedure and assign each utility specific responsibilities. In highly integrated pools, supply outages were shared.

Hourly plant programming
Plant start-up and shut-down could be programmed by each utility or by the pool. Any pool with both centralised programming and centralised dispatching was in fact operated as if it were a single system. These arrangements, which minimised operating costs, also entailed a greater loss of individual utility independence and control over its own system. Financial compensation was arbitrated for companies connecting plants to the grid above and beyond their own needs to meet pool requirements.

A.5.3.2 Co-ordinated Planning

Pool utility rights and obligations with regard to the ownership and use of production facilities (plants and grid) were discussed in Sect. A.5.2. Here this issue is addressed from the specific standpoint of planning. It was only in highly integrated pools that members acquired specific planning obligations. In less closely-knit pools such obligations were limited to submitting individual plans to a pool committee for the information of other members.

As far as plants were concerned, joint planning helped pool members attempt to attain an optimal generating structure. In the more tightly integrated pools, joint installed capacity needs were determined in accordance with load predictions and established reliability criteria, and then divided among members in accordance with their maximum demand loads and the shape of their load curves. Consequently, each member was obligated to have sufficient installed capacity to satisfy its load curve and its proportional part of the reserves. In the event of

energy shortages, the pool member concerned had to conclude firm power agreements or purchase a share in a plant or otherwise pay a penalty for each kW it was short. Such installed capacity responsibilities were reviewed periodically to take account of variations in demand, both utility by utility and overall.

Companies could plan to build plants or lines that were of no interest to the pool, but only the ones identified to be of joint interest were eligible for pool benefits. Utilities could, and at times were required to, offer other members part of the capacity and energy deriving from the latter types of plants to other members under short- or long-term agreements. Short-term arrangements made it possible to meet a member's power needs or cover shortages before the next review of capacity responsibilities. Under long-term contracts, usually for the life of the plant, the buyer acquired part of the plant's capacity.

A.5.4 Establishing Prices for Energy Exchanges and the Use of the Transmission Grid

As noted above, membership in a power pool had no impact on a utility's relationship with state regulatory commissions for the intents and purposes of ratemaking, or with the FERC in connection with the regulation of energy exchanges with non-pool companies. Moreover, for many of the transactions between pool members (the ones not affected by co-ordinated pool operation in areas such as firm power exchanges, system capacity and so on), the provisions cited in Sect. A.4 on wholesale energy sale characteristics and rates continued to be fully applicable. Power pools, however, were affected by two circumstances that need to be addressed, expanding on the discussion in Sect. A.4.

1. New types of transactions tended to arise in power pools due to their closely co-ordinated memberships. Examples would be reserve exchanges and energy transactions in the event of forced or programmed maintenance outages.
2. In highly integrated power pools, and in particular where centralised dispatching was in place, it was not possible to identify inter-utility transactions a priori to ascertain what type they were or even if they existed at all; in light of this, the ratemaking criteria set out in Sect. A.4 could not be applied.

This section deals primarily with the analysis of these two situations. In any event, a series of principles was established for ratemaking in power pools: the pool identified the transactions that generated profits and costs and designed methods to calculate and distribute both fairly and effectively. Prices were calculated to cover the cost of all transactions, ensure the greatest possible savings and provide for a fair distribution of costs and profits among pool members. The more tightly integrated the pool, the greater was the complexity it faced in establishing equitable prices.

For the intents and purposes of establishing prices, internal pool transactions were classified into one of the following three types:

- Energy transactions.
- Capacity transactions.
- Grid use transactions.

Energy transactions

This category covered the economic energy transactions described in Sect. A.4, which were further subdivided into different categories to establish prices: economic transactions per se, and transactions for programmed outages, forced outages and capacity shortfalls. In power pools with no centralised dispatching these transactions could be readily identified. Where arrangements called for centralised dispatching, the procedure described above was used: i.e. a posteriori comparison of real programming to calculated individual dispatching.

Economic transactions were operations simply designed to take advantage of the differences in the buyer's and the seller's incremental costs. The other three categories of economic transactions were intended to accommodate special circumstances. In programmed or forced outage transactions, energy was exchanged while one of the buyer's plants, which would have otherwise been used, was out of service due to pool-programmed maintenance or forced unavailability, respectively. In capacity shortfall transactions, the buyer experienced a power shortage (installed capacity plus capacity purchases from other plants plus firm energy purchases) and was unable to cover its own load plus its pool-allocated reserves.

In all these cases, the price of the service exchanged was determined by the respective companies' incremental costs. Pools with centralised dispatching applied the costs deriving from the calculation of individual dispatching. This generated a savings fund, since buyers paid more than sellers received. The balance in the fund was subsequently distributed according to pre-established criteria, such as each company's proportional share in strictly economic transactions.

Capacity transactions

These transactions were typically concluded when one utility needed capacity (and not necessarily energy) from another to honour its capacity obligations to the pool, in other words, its own maximum load plus allocated reserves. This situation might overlap with the energy transaction, in which case the respective costs were summed. Several categories could be distinguished: for reserves, programmed outages, forced outages and capacity shortages. The last three were discussed above. Reserve transactions, in turn, were sought by utilities that found that their reserve quotas could be more economically covered by another utility's plant.

The prices for these transactions were, once again, obtained on the basis of the incremental costs incurred by the parties concerned as a result of the transaction. It should be noted that in the case in point these incremental costs did not include the energy component (already taken into account when valuing energy transactions),

but only the component associated with the capacity transaction. For instance, in a reserve transaction which enabled the buyer to refrain from connecting a plant that would only have been used to meet the reserve quota requirement, the price would reflect the costs saved in plant start-up, operation under technical minimum conditions and shut down.

One controversial issue was the establishment of the maximum amount of time that a forced outage could be regarded to be just that, and not a capacity shortage. This sort of situations arose when a plant experienced very long-term forced unavailability. The price of forced outage transactions was usually kept higher than incremental costs as an incentive for utilities to maintain high availability.

A utility with a capacity shortage had to pay a penalty, which typically was very nearly the capital costs of the least expensive plant that could be bought to meet the company's capacity needs. Before reaching such a situation, utilities would buy firm power or capacity from other utilities under the usual terms described in Sect. A.4.

Grid use transactions

As far as the transmission grid was concerned, prices reflected the use of one utility's grid by another. There was no single method for establishing the price for this service, due to the difficulty involved in fairly and effectively compensating a company for the use of its grid.

Prices could be established on the grounds of losses or could be standardised where short-term energy exchanges prevailed. When standardised prices were adopted, it was up to the pool to decide which exchanges were covered and which were not.

A.6 Summary and Discussion

This Annex has presented the most salient features of electricity ratemaking in the USA in the early 1980s. One outstanding characteristic of electricity ratemaking in the United States in the 1980s was the enormous diversity of procedures in place, the outcome of the changes introduced since the industry first began to be regulated by state commissions. New York and Wisconsin created theirs in 1907 and the Federal Power Act of 1920 created the Federal Power Commission, now known as the FERC. Nonetheless, despite all this diversity, the overall structure remained relatively stable over time and was largely shared by systems across the country.

This review would be incomplete, however, without at least a brief discussion of the criticism levelled at various aspects of the traditional American approach to electricity ratemaking and the main changes proposed. This section is devoted to such a discussion. A more detailed treatment of these issues can be found in Ref. [4, 7, 8, 10, 12–15, 22, 24, 32–34], which are only briefly summarised in the paragraphs below.

A.6.1 Discussion

Globally speaking, and judging by the indirect results, it may be sustained that the American electricity ratemaking system worked satisfactorily for a relatively long time [22]: unnecessary duplication was avoided, the cost of electricity was comparatively low, the quality of service was excellent, capital investment was suitably remunerated and clearly unfair discrimination in rate structures was avoided.

Moreover, the traditional regulatory framework provided for regulatory stability. The guarantee that costs would be recovered generated a climate that favoured investment, reduced capital costs and provided for high security of supply. In addition, provision was made for meeting "social obligations" such as special rates for disadvantaged communities of users, R&D activities, protection of autochthonous fuel, diversification of energy sources and environmental protection.

However, in the late 1970s, criticism of and/or proposals for modifying many of the aspects of traditional US electricity ratemaking began to be advanced. The most significant of these are discussed below.

As noted on several occasions in this unit, one particularly striking aspect of American ratemaking was the variety of regulatory bodies involved in the process, which could give rise to the application of very different procedures even within one and the same utility (operating across several state lines, for instance). This led on occasion to paradoxical situations: a certain utility operating in the state of Massachusetts, for instance, unbundled into a generation and transmission company on the one hand and a number of mere distributors on the other, a structure that qualified it for FERC regulation and enabled it to avoid (particularly strict) state rules. In Texas, on the contrary, where the regulatory commission was more prone to favour electric utilities, all the state's electric companies physically disconnected from the rest of the American system to completely avoid FERC jurisdiction. Numerous attempts have been made, generally sponsored by the FERC, the Department of Energy or the National Association of Regulatory Utility Commissioners, to establish more uniform criteria and methods.

Although traditional regulation was based on the recovery of cost of service, under the American regulatory mechanism new rates were often approved considerably after they had been designed (regulatory lag). This lag represented no problem whatsoever for the utilities in an environment of declining electricity prices, such as in the late 1970s. From then on, however, under pressure of a number of factors—including the high cost of money, legal difficulties to build new plants and lines, rising fuel prices and costs in general-price trends reversed. In this new scenario, regulatory lag suddenly became relevant, since the rate of return on capital calculated using last year's costs was insufficient for the current year. The problem was not, then, intrinsic in the revenue requirements approach, which should in fact be regarded more as a systematised method to determine cost of service than an original procedure. It was, rather, a problem of the practical

implementation of the method. Several solutions were proposed [4, 7, 22], the most prominent of which included reducing the time required to process applications for rate changes and using estimated future costs to calculate cost of service.

From the 1980s to date, change in traditional regulation has been driven by a series of considerations of a critical nature. Perhaps the aspect most frequently criticised was the very philosophy on which cost of service and rate structure were based. Economic theory sustains that rates are most economically efficient when they equal the marginal cost of operating the electricity system. Rates based on embedded costs were lower than marginal costs, since they included old plants built very inexpensively. Together with the development of marginalist theory to determine rates, further thought was given to rate structures that would send the right signals to customers [7, 10, 12, 13, 26]. The outcome of these developments has been hourly rates, whose implementation was facilitated by the evolution of computer and communications technology.

Real-time pricing schemes have now been implemented in a number of electricity systems around the world, which are based on real short-term marginal costs at any given time. In a traditional environment, these would naturally have had to be completed with revenue conciliation methods to reach the true cost of service.

Another frontal attack on traditional regulation came with the advent, first, of renewable energies and co-generation or CHP facilities as "external" generators under the incentives provided for in the PURPA act (qualifying facilities). This later gave way to (Build-Operate-Own (BOO) and Build-Operate-Transfer (BOT) arrangements, which in turn broke the ground for the appearance of Independent Power Producers or IPPs. The need to handle transactions with all these new players on the field, together with the ongoing development of power pool transactions, brought changes in the way inter-utility rates were determined, with a tendency to form liberalised organised market places that fuelled competition between companies.

The problems posed by traditional regulation with respect to risk management and investment incentives also contributed to development along these lines. Traditional cost-of-service frameworks encourage overinvestment, with consumers shouldering all the risks. Liberalisation has evened the score somewhat, by passing at least part of the risk on to the generating business.

References

1. D.O.E (1980) The national power grid study. J.S. Department of Energy, Economic Regulatory Administration
2. Electric Council of New England (1981) Electric utility industry in New England statistical bulletin
3. Energy Information Administration (1982) Typical electric bills. Jan 1982

4. Federal Energy Regulatory Commission (1983) Statistics of privately owned electric utilities, 1981, annual (Classes A and B Companies), D.O.E./EIA-0044 (81)
5. Federal Power Commission (1977) National electric rate book. Rate schedules for electric services in communities of 2,500 population or more. Residential, commercial and industrial services. Maine
6. Kahn AE (1970) The economics of regulation: principles and institutions. Wiley, New York
7. Ebasco Services Incorporated (1978) Ratemaking: the transition from costing to rate-design prepared for the electric utility rate design study, EPRI, 12 April 1978
8. Hass JE, Mitchell EJ, Stone BK (1975) Financing the energy industry. Ballinger Publishing Company, Cambridge
9. Ebasco Services Incorporated (1977) Costing for peak load pricing: topic 4. Preparated for the electric utility rate design study, EPRI, 4 May 1977
10. Berlin E, Cichetti CJ, Gillen WJ (1975) Perspective on power: a study of regulation and pricing of electricity power. Ballinger Publishing Company, Cambridge
11. Gordon RC (1981) Reforming the regulation of electric utilities. priorities for the 1980. MIT Energy Laboratory Working Paper MIT-EL-81-033WP. June 1981
12. Sullivan RL (1977) Power system planning. Mc Graw Hill International Book Company, New York
13. Evaluation of Power Facilities. A Reviewer's hand-book, pp 296–338
14. Massachussets Electric Company (1983) Investigation by the departments to the property of proposed tariff changes
15. Ebasco Services Incorporated (1977) Ratemaking: topic 5 and illustrative rates for five utilities. Prepared for the electric utility rated design study EPRI, 6 June 1977
16. New England RegionaL Commission (1976) The new England power pool: descriptions, analysis and implications. Energy Program Technical Report 76–2, Boston
17. Pennsylvania-New Jersey-Maryland Interconnection (PJM) Agreement, 19 Mar 1981
18. Louis I (1984) Two economists look at power regulation. Electrical World Feb 1984
19. Federal Power Commission (1974) National Power Survey. The financial outlook for electric power industry, Dec 1974
20. Federal Power Commission (1967) Federal and state commission jurisdiction and regulation of electric. Gas and Telephone Utilities, Washington
21. Office of The Federal Register National Archives and Records Services (1982) Code of federal regulations. Conservation of power and water resources, 18. Chapter I, 1 Apr, 1982, pp 304–392
22. Schedule of Filing to the FERC by New England Power Company, Dec, 1983
23. Klosowicz PC (1981) FINREG: A Financial/Regulatioy model for utility capacity expansion plan evaluation. MIT energy laboratory report N° MIT-EL-81-022, June 1981
24. Edison Electric Institute (1975) Economic growth in the future. New York
25. Schmalensee R, Joskow PL (1983) Markets for power. An analysis of electric utility deregulation, The MIT Press, Cambridge
26. Turvey R, Anderson D (1977) Electricity economics. The Johns Hopkins University Press, Baltimore
27. Gray, John E (1975) Energy policy: industry perspectives. Ballinger Publishing Company, Cambridge
28. Jeynes PH (1968) Profitability and economic choice. The Iowa State University Press, New York
29. Marsh WD Economic of Electric Utility Power Generation.
30. Federal Energy Regulatory Commission (1981) Federal power commission 1977 Final Annual Report, D.O.E./FERC-0011, Washington

31. Federal Power Commission. The 1970 National Power Survey. Chapters 19 and 20
32. Federal Energy Administration (1976) Study of the electric utility industry demand. Costs and Rates
33. Larley WC (1982) FERC Regulation of bulk power coordination transactions. Unpublished draft staff working paper, federal energy regulatory commission, office of regulatory analysis
34. Federal Energy Regulatory Commission (1979) The florida electric power coordinating group: an evolving power pool. D.O.E./ERA-6385
35. Doran JJ, Hoppe FN, Koger R, Lindsay WW (1973) Cost allocation manual. NARUC, Washington.
36. Ebasco Services Incorporated (1977) Costing for peak load pricing: topic 4. results for virginia electric and power company. Prepared for the electric utility rate design study, EPRI, 6 June, 1977
37. Walters FS (1977) Analysis of various pricing approaches. Topic 1. Prepared for electric utility rate design study. EPRI, 2 Feb 1977
38. Mitchell B, Manning W, Acton JP (1972) Peak load pricing, European lessons for U.S. energy policy.Ballinger Publishing Company, Cambridge
39. Georgia Power Company (1982) Cost of service, allocation procedure. 19 Mar 1982
40. Federal Energy Regulatory Commission (1973) Uniform systems of accounts prescribed for public utilities and licenses classes A, B, C and D. D.O.E./FERC-0028, Washington
41. Vennard E (1979) Management of the electric energy business. McGraw Hill, New York
42. Suelflow JE (1973) Public utility accounting: theory and application. Michigan State University, Public Utilities Studies
43. Bary CW (1963) Operational economics of electric utilities. Columbia J Univ Press 40(1):261–288
44. Resource Planning Associates INC (1980) Power pooling issues and approaches, D.O.E./ERA/6385, Jan 1980
45. Federal Energy Regulatory Commission (1981) Power pooling in the United States. FERC-0049, Washington
46. Gordon RC (1981) Reforming the regulation of the electric power industry: Part II, MIT energy laboratory working paper MIT-EL-81-036WP, June 1981
47. Gordon RC (1984) Reforming the regulation of electric utilities. Lexington Books, Massachusetts
48. Economic Regulatory Administration (1981) Annual report to congress. D.O.E./RG-00034/2, May 1981
49. ICF (1981) Costs and rates workbook. Prepared for the electric utility rate design study, EPRI, Sep 1981
50. Kamat PG (1975) A Financial/Cost of service model of the electric utility in the U.S. MIT, Cambridge

Annex B: Grandma's Inheritance Theorem

I will try here to explain a basic principle in regulatory economics that my experience of nearly 20 years of teaching this subject around the world, and my frequent debates with colleagues in companies and institutions, have made me see how hard it is to assimilate, and that its ignorance or neglect leads frequently to fatally wrong conclusions.[10]

I will explain this principle with the aid of a simple example: an industrialist makes a special chemical product, that cannot be stored for periods of time longer than one month, with a maximum production capacity of 1,000 litres per month and a variable cost of 7 euros per litre. The demand for the product is very volatile, which causes its market price to suffer strong fluctuations that are typically monthly, which is approximately its production time. A buyer offers to our manufacturer the acquisition of the whole monthly production of 1,000 litres at a price of 9 euros per unit to be delivered the following month, to what the industrialist happily agrees.

Let us suppose now that the industrialist is notified that he has inherited from his grandmother a certain amount of money, or a diamond ring. And we ask ourselves whether this fact should make the industrialist change his opinion about the sale of his product. I believe we should all agree that this short-term sale is still as profitable for him as before and that his initial decision should not be modified because he is wealthier now (the improved financial position could affect his strategic decisions in the long term, such as undertaking new investments or even adopting a price "dumping" policy). The short-term decision of the industrialist should not be altered either if the inheritance would consist of shares in the stock market, whose future value is uncertain. Again, whichever is the future price of the shares—which is not under his control—the industrialist improves his financial

[10] The economic principle that is described in this note has much resemblance with Coase's theorem. This theorem, attributed to Ronald Coase, states that when trade in an externality is possible and there are no transaction costs, bargaining will lead to an efficient outcome regardless of the initial allocation of property rights. This finding, along with his 1937 paper on the nature of the firm, which also emphasizes the role of transaction costs, earned Coase the 1991 Nobel Prize in Economics. In October 2005 Ignacio Pérez-Arriaga published a previous version of the present note in the Spanish economic journal "5 Días".

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3, © Springer-Verlag London 2013

position if he sells the 1,000 litres of the product with a margin of benefit of 2 euros per unit.

Until now the rationale is quite obvious. But let us consider the case that the grandmother, rather sophisticated, leaves as inheritance to the industrialist a financial instrument—called contract for differences—that has as underlying or reference value precisely the volatile spot price of the chemical product in question. Let us suppose that the inheritance consists in a contract for differences for a monthly volume of 700 litres during the next 12 months and with a strike price of 8 euros per litre. What this contract—strictly financial, this is, independent of the production activity of the industrialist; it is just like a bid on the spot price—implies is that, if in any given month the spot price for the product were only of 5 euros per litre, the industrialist would receive from the counterpart in this contract—whoever that is—three euros (the difference between 8 and 5) for each of the 700 litres. On the contrary, if the spot price were of 10 euros, the industrial would have to pay the counterpart two euros per litre. We see that, in this case, the inheritance can result in a net balance for the industrialist that can be either positive or negative, depending on the market price that the product ends up having during the next year and on the strike price of the contract. In any case, and that is the conclusion we want to reach here, this new form of the inheritance is no more than a net economic increase or loss for the industrialist, *independent of his activity as manufacturer and seller*, and that therefore the existence of the inheritance should not alter his decisions each month to produce or not depending on the market price of the product. This is Grandma's Inheritance Theorem.

Let us warn—and this is what confuses many—that in this last case the net effect for the industrialist can be considered as a sale contract of 700 litres of the product per month during a year with a fixed price of 8 euros per litre—given that whichever the market price may be this will be the net result for the industrialist—Some may think that this contract conditions the volume of production of the industrialist for the next year, together with its behaviour in the spot market. But this is a deceitful perception. In agreement with the previous "theorem", the industrialist must decide every month to produce for the total of his production or not in accordance with the monthly price of the product, independently of his grandmother's inheritance. The inheritance, in any if its forms, should not interfere on his market behaviour. Let us check this with one of the multiple cases that can occur: if the monthly market price were of nine euros, grandma's inheritance plays him a dirty trick, since it forces him to pay 1 euro for each of the 700 litres of the contract to the counterpart. But, at the same time, the industrialist is interested in producing and selling his total monthly production of 1,000 litres at a price of 9 euros per litre, since he obtains a margin of 2 euros per litre. We see, indeed, that the inheritance does not modify his market behaviour. In the same manner, if the market price was set at 6 euros per litre, the industrialist would have to completely interrupt his production—as he would do if the inheritance didn't exist—and he would simply receive two euros for each of the 100 litres of the contract for differences.

What has this to do with electricity markets? Paying a little attention it can be observed that grandma's inheritance theorem is useful in many situations. Here we shall examine two of them. The first one is the implementation of a kind of special long-term contracts to mitigate the market power that can derive from an excessive horizontal concentration. They are contracts for differences with a volume and strike price determined by the regulator. In Spain, they were proposed in the White Book of 2005 with the name of "virtual contracts", but they have not been implemented. In Ireland they are used since November 2007 under the denomination of "directed contracts". Other countries have considered or applied similar instruments. Some have branded this regulatory instrument as intolerably interventionist, but they happen to be exactly the contracts for differences mentioned before as possible grandma's inheritances. Thus, it can be proved that virtual contracts neither interfere in the normal behaviour of the agents of the electricity market in the short term, nor they "take out of the market" production plants, as it has been said in some cases. But they do have the effect of reducing the market power of the dominant agents, which are those who by their size and characteristics can manipulate the market price in their own benefit.

How is it possible? Let us take up again the case of the industrialist and heir. We shall suppose that with his 1,000 litres of production, he is a dominant agent and that withdrawing production or offering abusively high prices he could manipulate the market price and increase his income. However, grandma's inheritance in the form of a contract for differences for 700 litres, in combination with the normal running of the market, would have the effect that an hypothetical rise in the price achieved by the practice of his dominant position would not affect the 100 litres covered by the contract, but only the remaining 300 (or what the market would accept, after withdrawing production or rising the price of the supply) drastically reducing the interest to manipulate the price. So these "virtual contracts" or "directed contracts" have the interesting property of mitigating the market power without hindering the normal market performance. They could also be used (now with the strike price being fixed by the regulator under the estimated future price of the market) to extract any rents that the regulator sees fit from certain generating companies, once again without interfering in the market. It does not seem a bad regulatory element, if there is an intention to do something to solve the problem of an excessive concentration when the option of divesting assets is not acceptable, or when there is a will to restore to the consumers some windfall profits of generators that the regulator deems inappropriate for the generators to keep.

Grandma's inheritance theorem has other many applications. For instance, it enables the clear establishment of equitable criteria for the assignment of the emission allowances of greenhouse gases under a cap-and-trade system. An emission allowance for one ton of CO_2 and valid for a period of time—the interval from 2008 to 2012, for instance—is just a financial title exchangeable for its monetary value in the market in any moment of its temporary period of validity. For the agent that owns it, it is equivalent to an amount of money of unknown future value, since the market price of the emission allowances fluctuates in time,

and the agent cannot control it. Let us suppose that the criterion to assign the emission allowances to a company that generates electricity was completely independent of its production and emissions, and also of its decisions to retire or modify the existing power stations, or to install new ones—unfortunately this does not seem to be the case in a good part of European countries—In that case, the emission allowances would represent a "grandma's inheritance", that should not affect the behaviour in the market of the power stations to whom the rights were granted, neither the decisions of modification, retirement or new inversions in generation plants. Thus, no agent should claim allowances to be able to run in the market, nor to not retire a facility. Emission allowances can serve as compensation to a company for the loss of income caused by the regulatory change implicated by the introduction of the emissions market. This is about solving an issue of equity. But if the correct assignment of the rights is not made to depend on the behaviour of the agents in the market, its future efficient functioning will not be distorted. Once again grandma has come to show us the correct path.

Index

A
Access tariff, 398, 411
Access to the transmission grid, 309
Accounting approach, 409, 413
Accounting separation, 138
Administratively priced long-term contracts, 392
Advanced meters, 429, 473
AGC or automatic generation control, 39
Agency for the Cooperation of Energy Regulators (ACER), 130, 527
Alberta, 457
Ancillary services, 40, 240, 254, 286, 292, 356, 358, 380, 381, 385, 449, 562, 649, 664
Areas of influence (''areas de influencia''), 304
Argentina, 132, 348
Association of European Energy Exchanges (EuroPEX), 527
Audited costs, 351
Aumann–Shapley, 305
Australian National Electricity Market (NEM), 501
Average participations, 304
Averch-Johnson, 160

B
Balance responsible party (BRP), 376
Balancing markets, 375
Barriers, 107
Barriers to Retail Market Development, 464
Benchmarking, 129, 155, 186
Beneficiary pays, 294, 305, 309, 410
Beneficiary pays principle, 418
Bertrand model, 110
Black-start capability, 385
Book value, 159, 160, 178, 179, 362
Broker, 446

C
California, 229, 230, 348, 432, 451, 452, 594, 656, 668
Capacity charge, 420
Capacity market, 353
Capacity payment, 90
Capital asset pricing model (CAPM), 184
Capital expenditure (CAPEX), 171, 173, 223, 224
CCGT (combined cycle gas turbine) plants, 16
Central American Electricity Market, 501
Centralised planning, 29
Centralised regional expansion planning, 514
Chile, 132, 347, 452
CHP (combined heat and power), 19
Climate change, 539
Coase theorem, 114, 719
Collusion, 107
Colombia, 348, 353, 458
Colombian, 130
Command and control, 546
Commercial or non-technical losses, 233
Commercial quality, 215, 225
Competitive bidding, 345
Competitive tendering process, 287
Complementary charges (CC), 280, 293
Complex auctions, 366
Computation of Nodal Prices, 281
Connection charges, 208, 293, 310, 311, 404, 433, 437
Consumer behaviour, 63
Consumer satisfaction, 154, 172, 230
Consumer surplus, 70
Contestable markets, 99

I. J. Pérez-Arriaga (ed.), *Regulation of the Power Sector*, Power Systems,
DOI: 10.1007/978-1-4471-5034-3, © Springer-Verlag London 2013

Continuity of supply, 10, 100, 215, 221, 226, 233, 238, 484
Contract path, 302
Contracts for Differences (CfDs), 321, 362, 494
Corrected ordinary least squares method (COLS), 186, 188
Cost causality, 294, 410, 418
Cost drivers, 404, 420
Cost-of-service, 39, 129, 133, 152, 155, 156, 158, 160, 161, 166, 174, 193, 213, 348, 480, 575
Costs of producing electricity, 51
Cost reflective network pricing (CRNP), 304
Cost socialisation, 294
Cournot model, 109

D
Data envelopment analysis (DEA), 186, 187, 227, 228
Day-ahead market (DAM), 356, 364
Decentralised market-based, 29
Decoupling, 481
Deep cost tariffs, 433
Deep network charges, 311
Default supplier, 456
Default tariff, 398, 451
Demand-side management (DSM), 9
Depreciation, 157–159, 174–176, 178, 181–183, 192
Distributed generation, 7, 20, 34, 43, 48, 201, 202, 212, 236, 239, 247, 284, 424, 436, 437, 481, 669, 670, 676
Distribution grid, 26, 34, 178, 201–203, 209, 256, 413, 416, 420, 470, 669, 670
Distribution licences, 208
Distribution transformers, 202–204, 207
Dominant position, 102
Double-dividend effect, 555

E
East Australia, 357
Economic instruments, 546
Economics of Transmission, 260
Economies of scale, 20, 72, 82, 97, 262, 295
Elasticity of demand, 49
Electricity Pool model, The, 365
Electromagnetic compatibility (EMC), 219–221, 544
Energy charge, 420
Energy efficiency, 475
Energy losses, 58, 201, 212, 214, 227, 232–236, 238, 247, 266, 271, 435
Energy not supplied, 215, 217, 228, 229, 638, 639
Energy service companies, 481
England, 132
England and Wales, 348, 353, 354
Entry barriers, 98
Environmental adder, 556
ERCOT, 502
European Directive, 450, 454
European Network Transmission System Operators for Electricity (ENTSO-E), 527
European Union's Internal Electricity Market (IEM), 501
EU-wide ten-year network development plan (TYNDP), 530
Explicit transmission capacity auctions, 315
External costs, 52
Externalities, 42, 113, 115, 540

F
FACTS hardware, 259
Federal Electricity Regulatory Commission, 130
Feedin tariffs (FIT), 563
Financial transmission rights, 312, 519
Firm transmission rights, 312
Fixed costs, 52
Flat rate, 494
Flow-gate transmission rights, 324
Format of the transmission charges, 298
Framework Guidelines, 529
France, 369
Frequency control, 381
Frontier methods, 186, 187
Future contracts, 361

G
Gas turbine plants, 16
Generalised network constraints, 311
Generation mix, 19
Generators surplus, 70
Germany, 369, 472
Grandma's inheritance theorem, 364
Grid constraints, 268

H
Hirschman-Herfindahl Index, 103
Hydroelectric stations, 15

I
Imbalance pricing, 378
Impedance of the lines, 257
Implicit auctions, 312
Incentive-based regulation, 134, 135, 155, 161, 164, 165, 171–173, 180, 193, 194, 196, 197, 213, 214, 227, 288, 648
Independence of flows from commercial transactions, 259
Independent power producers (IPP), 345
Independent System Operator (ISO), 342, 502, 510
Information asymmetries, 134, 161, 180, 196, 560
Information Quality Incentive (IQI), 223
Infra-marginal rent, 90
Institutional costs, 425
Integral tariff, 398, 427
Integrated Resource Planning (IRP), 477, 557
Intergovernmental Panel on Climate Change (IPCC), 476
Intermittency, 570
Internal Electricity Market (IEM) of the European Union (EU), 510
International Energy Agency (IEA), 55, 241, 242, 648, 651, 653, 671
Inter-TSO compensations (ITC), 531
Intraday markets, 375
Investment Cost Related Pricing (ICRP), 299
Investments at risk, 515
Ireland, 145
Islanded operation, 238
Italy, 454, 455

J
Joskow, 459

K
Kirchhoff's laws, 2, 257

L
Last resort tariff, 398
Legal separation, 138
Lerner index, 106
Liberalisation, 136
Light-handed regulation, 155
Littlechild, 459
Load curves, 10
Load duration curve, 11
Load–frequency control, 39
Local network constraints, 311
Locational Marginal Prices (LMP), 271, 372
Long-run marginal cost of generation, 80
Long-run marginal costs, 54, 72, 409
Loss factors, 315
Lumpiness, 280, 295

M
Maintenance, 35, 36, 152, 173, 206, 207, 212, 214, 216, 239, 254, 260, 277, 381
Management separation, 138
Marginal cost, 53, 54
Marginalist theory, 409
Marginal participations, 303
Marginal unit, 89, 274
Market-based decentralised approach, The, 30
Market equilibrium in the short term, 66
Market failures, 48, 88, 112, 152, 342, 479, 562, 568, 658, 661, 673
Market Operator (MO), 352
Market player initiative with regulatory supervision, 289
Market power, 96, 385
Market power metrics, 388
Market splitting, 282
Market structure, 97
Massachusetts, 452
MER (Mercado Eléctrico Regional, Regional Electricity Market) in Central America, 510
Merchant lines, 264, 290
Merit order, 89
Merit order effect, 573
Metering, 471
m-firm concentration ratio, 103
MIBEL, 501
Minimum performance standards, 477
Model firm approach, 179
Monopoly, 62, 84, 89, 92, 95, 103, 109, 139, 151–153, 207, 264, 288, 633, 657
MW-mile, 302

N
Natural monopoly, 252, 264
Net metering, 307, 437
Net social benefit (NSB), 70
Netherlands, 464
Network access, 253

Network charges, 207, 210, 289, 298, 301, 306, 310, 311, 426, 434, 437, 506, 525, 670
Network codes, 529
Network cost allocation, 253
Network investment, 253
Network losses, 265
Network remuneration, 282
Network utilisation, 300, 424
New Electricity Trading Arrangements (NETA), 355
New England ISO, 502
New South Wales, 348
New Zealand, 348
Nodal prices, 271
Non-cooperative oligopoly models, 108
NORDEL, 501
Norway, 132, 226–229, 234, 348, 353, 355, 356, 384, 612
Nuclear power plants, 17

O

OFGEM (Office of Gas and Electricity Markets), 222, 223, 225, 226, 668
Ohmic losses, 266
Oligopoly, 107
Operating and maintenance (O&M), 54, 56, 58, 157, 158, 212, 228, 260, 261
Operating constraints, 265
Operating expenses (OPEX), 171, 172, 176, 177, 191, 192, 223, 232, 247
Operation planning, 34
Option contracts, 361
Options versus obligations, 325
Over-the-counter (OTC), 360
Ownership separation, 138

P

Pancaking, 297
Pass-through, 494
Perfectly competitive market, 85
Performance-based ratemaking or regulation (PBR), 172, 229–231, 247
Physical transmission rights, 519
Pigouvian tax, 549
Pivotal supplier indicator, 104
PJM, 502
Planning and Investment, 32
Point-to-point transmission rights, 323
Policy lines, 284
Pool, 354
Postage stamp, 301
Power plants, 15
Power purchase agreements (PPA), 345, 346
Predatory pricing, 112
Price cap, 165–167, 193, 222, 594, 617
Price elasticity of demand, 9
Price quality regulation, 195
Price takers, 67
Primary regulation, 382
Producer behaviour, 64
Property rights, 115
Public good, 116, 543
Public Utilities Regulatory Policies Act (PURPA), 345
PX model, The, 365

Q

Quality of service, 14, 129, 133, 135, 161, 180, 194, 195, 200, 215, 225, 226, 230, 265, 270, 288, 484, 485

R

Ramsey prices, 301, 417
Rate base, 157, 159, 160, 171, 178, 200, 232
Ratemaking principles, 399
Rate-of-return, 133, 152, 155, 156, 160, 169, 191
Reactive power rates, 435
Real-time operation, 38
Real-time pricing (RTP), 429
Re-dispatch of generation, 313
Reference network models, 232, 239, 420
Regional grid cost allocation, 523
Regional grid development, 512
Regional integration, 503
Regional regulatory test, 515
Regional transmission capacity, 518
Regional Transmission Organizations (RTOs), 502
Regulated asset base (RAB), 171, 175, 178–180, 182, 197
Regulated monopolies, 86, 95
Regulatory authority, 141
Regulatory lag, 159
Reliability, 28, 37, 88, 154, 204, 206, 216, 219, 230, 233, 263, 430, 589, 596, 608, 608–610, 614, 617, 619, 638
Renewable DG, 236, 237, 239
Renewable energy auctions, 565
Renewable sources of energy, 18
Replacement cost, 159, 179
Reproduction costs, 159, 178
Requests of connection to the grid, 310

Residual demand, 109
Residual supply index, 105
Responsibility for investment, 306
Restructuring, 136
Retail competition, 346
Retailer, 446
Retail Processes, 482
Revenue cap, 165–169, 172, 175, 193, 194, 201, 222, 228, 231, 232, 234, 288
Revenue reconciliation, 95, 416, 431
Revenue yield, 168, 169, 176, 177, 192, 222
Risk allocation, 140
RPI-X, 134, 166
Rural electrification, 43, 202, 241–243, 671

S
Safety valve, 556
SAIDI (system average interruption duration index), 217, 226, 230, 231, 421
SAIFI (system average interruption frequency index), 217, 226, 230
SCADA, 38
Schweppe, Fred, 85, 95
Screening curves, 79
Secondary regulation, 383
Second-best, 555
Security of supply directive, 533
Shallow cost tariffs, 433
Shallow network charges, 310
Sharing earnings and losses, 169, 229
Short-run, 53
Short-run marginal cost of generation, 81
Short-term marginal cost (STMC), 410
Simple auctions, 367
Single buyer, 456
Single Electricity Market (SEM), 501
Single price areas, 519
Single pricing, 312
Single system paradigm, 297, 507
Siting of transmission facilities, 291
Slack bus, 203
Sliding scale, 169, 170, 180, 223, 225
Smart grid, 207, 667, 670
Smart metres, 207, 420
Smoothing X, 174–176
Spain, 140, 205, 211, 216, 231, 234, 235, 237, 240, 256, 348, 453, 462, 486, 487, 665
Spanish White Paper, 462
Spot market for electricity, 30
Spot prices, 271
Stackelberg model, 109
Stochastic frontier analysis (SFA), 186, 188
Stranded costs, 425
Strategic entry deterrence, 112
Strong sustainability, 540
Substations, 7, 24, 26, 179, 202, 261, 293
Supervised centralised planning, 287
Supplementary charge, 431
Supplier of last resort (SoLR), 454
Switching procedures, 469
System operation, 140
System operator (SO), 31, 39, 352

T
Tariff categories, 424
Tariff periods, 422
Tariff structure, 399, 407
Technical losses, 233–236, 672
Technology policies, 558
Tertiary regulation, 383
Texas, 458, 471
Thermal power stations, 16
Third energy package, 528
Time-of-day, 494
Time-of-use (TOU) rates, 430
Total expenditure (TOTEX), 171, 172, 185
Total factor productivity (TFP), 187
Tradable green certificates, 565
Tradable quotas, 549
Tradable white certificates (TWCs), 477
Trader, 446
Traditional centrally managed approach, The, 30
Traditionally regulated monopoly, 288
Transformers, 24, 202, 204, 261, 266
Transmission and distribution costs, 58
Transmission cost allocation, 292
Transmission network business models, 286
Transmission network investment, 283
Transmission planning, 283
Transmission rights, 320

U
Unbundling, 129, 137–140, 208, 222, 246, 342, 444, 466, 467, 470, 471, 475, 656
Unit commitment, 350
Unit Commitment and Dispatch, 36
United Kingdom, 166, 222, 225, 234, 454, 472, 487
Universal access to electricity, 241, 648, 661, 671
Use-of-system charges, 208
Utilisation factors, 433

V
Value of lost load, 64
Variable costs, 53
Variable Network Revenues (VNR), 292
Vertical integration, 100
Virtual power plants auctions, 392
Voltage, 5, 6
Voltage level, 256
Voltage quality, 154, 201, 215, 218, 238
Voltage-reactive power (V/Q) regulation, 385
Voltage regulator, 40
Volumetric tariffs, 158
Vulnerable consumers, 458

W
Weighted average cost of capital (WACC), 160, 175, 177, 183, 184, 191, 197, 260
Windfall profits, 92, 575

X
X factor, The, 173, 175–177, 197, 222

Y
Yardstick competition, 155

Z
Zonal pricing, 282, 312