Hubert Missbauer
Reha Uzsoy

# Production Planning with Capacitated Resources and Congestion

Springer

# Production Planning with Capacitated Resources and Congestion

Hubert Missbauer • Reha Uzsoy

# Production Planning with Capacitated Resources and Congestion

Springer

Hubert Missbauer
Department of Information Systems
Production and Logistics Management
University of Innsbruck
Innsbruck, Austria

Reha Uzsoy
Edward P. Fitts Department of Industrial
and Systems Engineering
North Carolina State University
Raleigh, NC, USA

ISBN 978-1-0716-0352-9        ISBN 978-1-0716-0354-3    (eBook)
https://doi.org/10.1007/978-1-0716-0354-3

This Springer imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

# Preface

Problems related to planning and control of production facilities and the larger supply chains of which they form an essential part were among the earliest to be addressed using mathematical models, beginning in the early years of the twentieth century and making great advances in the 1950s and 1960s. Lot sizing, workload control, mathematical programming models for capacity allocation and release planning, and performance analysis using queueing or simulation are broadly studied and taught, and many ideas from this work are implemented in practice.

A central concept in the study of production systems, from whatever angle they are approached, is loosely referred to as capacity—their ability to convert a specified set of inputs into a specified set of outputs over time. This capability can be captured to an arbitrary level of accuracy (at least in theory) in simulation models and to different degrees, depending on the mathematical assumptions adopted, by mathematical programming, queueing, and stochastic analysis models. Thus if a given production system is modeled by different mathematical formalisms the results obtained ought to be in agreement, at least to the extent permitted by their different assumptions.

Our work in this volume originated in our realization that the concept of capacity in the widely taught and implemented mathematical programming models for production planning produced results incompatible with those suggested by simulation models and queueing analysis. Their wide use suggests that many researchers and practitioners find them sufficient to their purposes, and these approaches indeed have several advantages, most notably transparency to the user and computational tractability. However, research over the last three decades, in which we have participated, has produced a body of results allowing a more comprehensive view of these issues.

This monograph seeks to bring together in one volume the state of the art in the domain of production planning at the time of writing, beginning with the different frameworks in which this work is deployed in industry, the principal approaches by which these problems have been addressed, and their strengths and limitations. The latter half of the book focuses on recent work that seeks to bridge the gap between deterministic optimization and queueing, by using ideas from the latter to develop metamodels of system behavior that can be incorporated effectively in the former—clearing functions. Our intention is that the interested reader can find all the material they need to understand the basic issues arising from this body of work and the various solution approaches deployed to date. We hope that by the end of the volume the reader will agree with us that there are still many interesting and challenging research problems in the domain of production systems, which big data and increasingly powerful commercial optimization and simulation software open to new approaches.

The two of us arrived at this body of work through quite different paths—Uzsoy through production planning and scheduling in semiconductor manufacturing, and Missbauer from production planning and control concepts and workload control. This collaboration Correct sentence should read "This collaboration that began with a chance meeting at the INFORMS International Conference in Istanbul in 2004 has led us to learn from each other and from the numerous students and collaborators with whom we have interacted along the way. Many have contributed to our understanding of this material and to the body of research as their extensive presence in the bibliography suggests. Special thanks are due to the students of ISE 789 at NC State in the Fall of 2017, who were subjected to an early draft of the book and offered extensive, constructive feedback which improved both content and presentation very significantly.

We would also like to express our deep appreciation for the ongoing support and encouragement we have received from the Springer personnel during this effort, starting with Gary Folven, and later Neil Levine and Matt Amboy, as well as the Series Editor, Professor Fred Hillier.

Reha Uzsoy would like to acknowledge the support of the National Science Foundation and Intel Corporation and by the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. And, as ever, he owes his family—Lucia Mendez, Isabel, and Ana Sofia—and his parents, Nancy and Safak Uzsoy—a debt of gratitude for their love, patience, and support that cannot be measured, only acknowledged.

Hubert Missbauer would like to thank the members of the Production and Logistics Management team at the University of Innsbruck who supported this work. He especially thanks Stefan Häussler for critical reading and helpful suggestions, Alexander Lohr for his help in finishing the manuscript, and the secretaries Güler Ammann, Birgit Baldan, Pia Dialer, Julia Egger, and Simone Kirchner for typing the difficult text in a foreign language. This work is in loving memory of his late parents, Karl and Wilhelmine.

Innsbruck, Austria Hubert Missbauer
Raleigh, NC, USA Reha Uzsoy

# Contents

# Chapter 1
# Introduction

In this chapter, we lay out a generic conceptual framework for the production planning and control (PPC) problems in discrete parts manufacturing systems. We argue that a PPC system must address two distinct problems: coordination of material flow across the different production units making up the supply chain and effective control of the detailed production activities within the different production units. The first of these is the task of the planning level. Since the production units are managed autonomously and their internal operations are not under the direct control of the planning level, order release becomes a crucial function linking the two levels. We illustrate the application of these ideas to structure the PPC problems in two quite different environments: semiconductor wafer fabrication and manufacturing of optical storage media (CDs and DVDs). We conclude the chapter with an outline of the remainder of the volume.

## 1.1 The Production Planning and Control Problem

Any manufacturing firm must have some formal or informal system to decide how much of what product is to be made when, to ensure that all materials needed for production are on hand when needed and that products are delivered to customers on time. This task is referred to in the literature as *production planning and control* (Hopp and Spearman 2008; Buzacott et al. 2013) or *manufacturing planning and control* (Vollmann et al. 2005; Jacobs et al. 2011). Since value-adding processes in industry are increasingly distributed among multiple supply chain partners such as manufacturers, distributors, and third-party logistics providers, efficient production requires coordination not just within the production system itself but across all entities within the supply chain. Planning and control of material flow through the supply chain, in turn, require coordination with functional areas such as investment/finance, strategic planning, human resources, marketing, and sales. The terms *planning* and *control* accurately reflect the nature of these tasks: "planning" that

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_1

coordinated decisions must be made over time and "control" that the progress of work through the production and distribution network must be monitored to ensure it proceeds as planned. Throughout this volume, we shall refer to the generic set of planning and control procedures used by a firm for this purpose as a *production planning and control* (PPC) system.

Even in this age of powerful computing, vast data storage, extremely cheap sensors, and, for all practical purposes, instantaneous data transfer, the amount of information and computing required for centralized control of large manufacturing enterprises remains prohibitive. Hence firms are organized into smaller subunits focused on meaningful subsets of the production and distribution processes, or specific support functions such as sales, marketing, and human resources. The internal processes of these subunits are usually opaque to other units, requiring the specification of protocols to coordinate their activities. Hence in a manufacturing firm of any size, both the authority and the expertise for management decisions are distributed across different organizational units. Another important source of complexity is the need to coordinate decisions made over quite different time scales. Factories operate essentially in continuous time, new products may be introduced several times a year, while sequence-dependent setup times may require weekly or monthly lot sizing computations for critical equipment.

In this volume, we adopt the view of PPC systems for discrete parts manufacturing suggested by Bertrand et al. (1990) and extended to supply chains by de Kok and Fransoo (2003). These authors view a production system or supply chain (henceforth simply the *production system*) as a network of *production units*, which, in turn, are defined as groups of production resources managed as an autonomous unit and responsible for completing a well-defined portion of the production process required by the firm's final products. Each production unit consists of one or more *workcenters*, groups of machines performing a set of related processing steps, and the products may require processing at several production units before delivery to the customer.

This framework poses two fundamental problems to ensure the timely and profitable delivery of finished products to customers: coordinating the flow of material between production units and the efficient execution of activities within the production units themselves. This suggests a two-level, hierarchical approach to production planning and control in which the upper, *planning,* level addresses the *material coordination* task by specifying coordinated production targets leading to the desired combination of inventory levels and output across the entire system. This also requires that the planning level determine *the use and adjustment of capacity over time* based on appropriately aggregated demand information. Hence the planning level performs planning and control of the material flow through the entire production system using only an abstract representation of the production units. We refer to this planning and control problem as the *goods flow problem,* as in the term *goods flow control* introduced in Bertrand et al. (1990: 29 ff.).

The internal operations of each production unit are directed by its local management to achieve the production targets set by the planning level, constituting the lower, *scheduling and control,* level of the planning and control system. This, in its most general definition, allocates the individual unit operations of a particular product to specific machines or workers, determining their sequences and exact timing (Pinedo and Chao 2005; Pinedo 2012). Since it has no control over the internal operations of the production units, the planning level requires a model of how the production units convert their inputs—obtained from suppliers or other production units—into outputs that become available to other production units over time. These predictive models of production unit performance, referred to as *anticipation functions* (Schneeweiss 2003), lie at the heart of the research described in this volume.

The essential interface between the planning and scheduling levels is *order release*. The planning level coordinating the production units issues *production orders* that specify the particular product or component to be produced, the amount to be produced, and the dates by which the order must be completed by each production unit. The amount to be produced for each order (its *lot size*) determines its setup and processing time, and setup times may be sequence-dependent. Control over the production orders then passes to the detailed scheduling level within the production units concerned. Thus the planning level has no direct access to the exact timing of individual processing steps (operations) performed on the production orders, but can only decide their required due dates and when to release them to the scheduling level, i.e., to the control of the local management of the production unit in question. Hence in this view, production planning models are actually order release models since only order releases can be executed by the planning level; at the planning level, the completion time of a production order is not a decision but an estimate.

This separation of the planning task of determining production targets for the different production units from that of executing the plans through detailed scheduling and shop-floor control is the primary motivation for the research in this volume. In order for the firm to perform effectively, the planning function must coordinate the material flows across the different production units effectively; if it fails to do so, some production units may be idle waiting for material from others, substantial inventories may accumulate between some production units, or, frequently, both. Effective coordination of the material flow, in turn, requires that the planning level be cognizant of the capabilities of the production units it seeks to coordinate, at least to the point of ensuring a reasonable probability that they can complete the tasks assigned them in a timely manner.

It is possible, in principle, to build highly detailed models of the capabilities of each production unit and use these to compute detailed schedules, at the level of individual processing steps, for all production units across the entire enterprise. This can, in fact, be done in smaller firms with simple production processes whose capabilities are relatively easy to determine. However, for complex products such as cars or machine tools, where thousands of components, each requiring its own production process that, in turn, requires multiple unit operations, must be assembled into the final product, this approach is impractical due to the volume of information required and the complexity of the resulting planning procedures. Although in principle detailed scheduling can be interpreted as a planning activity, it turns out that, especially in discrete manufacturing, it is so closely linked to rescheduling and other control activities that separating the scheduling and control tasks does not reflect the actual decision structure (McKay and Wiers 2004).

This structuring of the production planning and control system into a planning level and a scheduling and control level is common to virtually all frameworks proposed to address the production planning and control task, especially in discrete manufacturing, indicating a broad consensus across the research and practitioner communities that centralized solution of these problems is neither practical nor desirable. In most large manufacturing firms, the complexity of the planning task itself is such that it must be decomposed into smaller subtasks operating on different time scales with different types of information; thus, the planning level is itself hierarchical.

Process industries such as steelmaking, continuous chemical processing, or paper production often exhibit substantially different characteristics that do not allow a straightforward decomposition into planning and scheduling levels. This is because the technical characteristics of facilities and products, such as sequence-dependent setups and sequencing constraints leading to cyclic production with long production campaigns, or constraints on the amount of time intermediate products can wait between operations, require close coordination between workcenters. In such environments, the scheduling decisions strongly influence the planning level's decisions, requiring specific decision structures. PPC systems for the process industries are discussed by Gunther and van Beek (2003) and Floudas and Lin (2004, 2005), while Tang et al. (2001) and Missbauer et al. (2011) discuss the particular case of steel plants. Local scheduling of groups of production resources organized into semiautonomous units similar to the production units defined above can also be observed in these environments, but the coordination protocol can be different from the planning–scheduling hierarchy described above. For example, Cowling and Rezig (2000) use heterarchical, negotiation-like concepts to coordinate continuous casting and hot rolling in steel plants. Therefore, the following description primarily holds for discrete manufacturing, although similar structures are often observed in process industries as well.

For the firm to operate, the planning level must eventually provide the production units or the manufacturing system as a whole with a *build schedule* specifying what is to be built when in terms of specific items and specific time points—whatever the process used to reach these decisions. This build schedule must be defined at the level of the individual items that constitute the output of each production unit or, at a minimum, at a sufficient level of detail that specific production targets for the production units can be derived and the necessary scheduling activities within the production unit can proceed. The problem is essentially that of matching the supply provided by the production units over time to the demand for final products in the best possible manner. The "best possible manner," of course, can vary widely based on the specific firm and production environment being considered. Substantial complications arise in practice from the need for the build schedule to reconcile the longer-term, often conflicting requirements of different functional groups within the firm. A product development unit may request scarce manufacturing capacity to validate a new product design; sales may insert a small, uneconomic production order that they believe may bring additional business in the future. Trade-offs of this nature are difficult to capture directly in a business process, let alone an optimization model, and involve negotiation between the different functional units.

This discussion has several implications. Firstly, deriving a build schedule that successfully reconciles the requirements of sales, production, and other functional areas is a complex negotiation process that can be supported by optimization models, but in all but the simplest cases cannot be, and probably should not be, fully automated. Secondly, whatever its internal structure, the planning level must translate the build schedule into capacity-feasible production targets—usually communicated in the form of production orders and their required due dates—for the production units. How this is accomplished will largely depend on the complexity of the products (e.g., their bill of material (BOM) structure) and their production processes (e.g., the importance of setup or batching requirements). Finally, detailed scheduling is not just a planning level defined for complexity reduction: it is a distinct decision function performed by the management of the production units, who are autonomous decision makers with their own targets, objectives, and domain knowledge. However, the decision space of the scheduling and control level is largely determined by the output targets that the planning level assigns to the production units. Since, as will be discussed in Chap. 2, the production unit can be usefully viewed as a queueing system, these output targets must be consistent with the production unit state variables like work in process inventory (WIP) levels and output rates. Ensuring that the production targets can be met with the resources available to the production units requires maintaining the state variables describing the production units at their desired values.

This leads to two questions. The first is how the build schedule of final products, whether for the entire production system or for portions of it such as production sites, is formulated. The second is how the production targets for the production units, usually described by a set of production orders with their release dates and required due dates, are derived from this build schedule. The answers to these questions largely determine the structure of the planning level of the PPC system. A still ongoing process of conceptual and software development that started about 1960 has led to two principal proposals for the structure of the planning level: the Manufacturing Planning and Control framework described by Vollmann et al. (2005) and Jacobs et al. (2011) and the Advanced Planning Systems (APS) framework (Stadtler et al. 2015), which are discussed in Chap. 3.

Whichever of these frameworks is adopted, the problem addressed in this volume remains important: the computation of capacity-feasible release schedules for each production unit that meet the output targets set by the planning level in the best possible way. However, the production units may not be able to meet the output targets corresponding to the build schedule for every product in every period due to the necessarily aggregate, incomplete nature of the computations used to obtain the output requirements at the planning level. Hence the output targets derived from the build schedule must often be refined by a subsequent decision level that receives these output targets as input and computes the size and timing of the production orders, specifying the amount of material for each product that is to be released to each production unit over time, together with their required due dates. Thus, the output targets generated by the planning level act as the *demand* for the products or stock-keeping units produced by the production units and will be used in this sense in the order release models described from Chap. 3 onwards.

Since the production units represent autonomous decision-making units, *coordination norms* between the planning and scheduling levels are necessary to determine what output targets can be requested by the planning level with reasonable hope of timely delivery by the scheduling level. A great many firms use *planned lead times* as an estimate of the time that will elapse between the work being released into the production unit and its emergence as finished product that can be used to meet internal or external demand. A maximum capacity loading, on the other hand, specifies an upper limit on the amount of work, usually measured in units of time required on a critical resource, that can be released to a given production unit in a planning period. The most common approach in practice is for the planning level to assume a constant planned lead time for each production order at each production unit, which the production unit commits to meeting as long as the agreed-upon maximum capacity loading is not exceeded. Thus effective management of *cycle times*, the time between the release of the production order to the production unit and its completion by the production unit, and correct specification of the planned lead times and maximum allowable capacity loading, which represent the capabilities of the production unit to the planning level, are of utmost importance. Serious discrepancies between the planned lead times used by the planning level and the cycle times realized in the production units lead to problems, usually the simultaneous presence of high WIP levels and poor on-time delivery performance.

This combination of fixed planned lead times and a maximum capacity loading is only one possible set of coordination norms. Relaxing the assumption of fixed lead times may allow cycle times and capacity loading to be better adapted to the output targets eventually determined by the planning level. If the planning level could accurately predict the consequences of its order release decisions on the amount and timing of the production units' output over time, it could better adjust capacity loading and output targets to their capabilities. This, in turn, requires more sophisticated anticipation of the production units' capabilities as well as optimization models for order release that can exploit this anticipation. *The formulation of mathematical programming models for the computation of capacity-feasible release schedules with enhanced anticipation functions is the central topic of this volume*.

We now illustrate the application of these ideas to the structuring of PPC systems in two different industries: semiconductor manufacturing and the manufacturing of compact discs (CDs).

## 1.2 Applications of Hierarchical PPC Systems

### *1.2.1 Semiconductor Manufacturing*

The process by which very large-scale integrated circuits are manufactured can be divided into four stages as shown in Fig. 1.1: wafer fabrication, wafer probe, assembly or packaging, and final testing (Uzsoy et al. 1992, 1994; Moench et al. 2013).

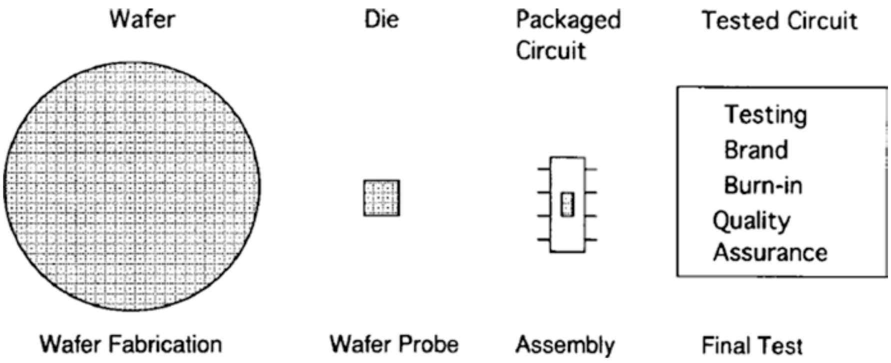


**Fig. 1.1** Basic steps of the semiconductor manufacturing process (Uzsoy et al. 1992)

Wafer fabrication is the most technologically complex and capital-intensive of all four phases. It involves the processing of wafers of silicon (or another semiconductor material such as gallium arsenide) to build up layers of patterned metal and wafer material to produce the required circuitry. The number of unit operations can reach several hundred for a complex component such as a microprocessor. To prevent particulate contamination of the wafers, processing must take place in a clean-room environment referred to as a *wafer fab*, or fab for short. Material moves through the fab in lots, usually of a standard size determined by the material handling system in use. While the specific operations vary with the specific product and technology in use, the basic operations can be described as follows:

*Cleaning:* The removal of particulate matter from the wafer before a layer of circuitry is fabricated.

*Oxidation, deposition, metallization:* A layer of material is grown or deposited on the surface of the cleaned wafer. Extensive setup times are involved in switching from one type of operation to another, resulting in machines being dedicated to a limited number of operations.

*Lithography:* This is the operation requiring greatest precision. A photoresistant liquid (photoresist) is deposited onto the wafer and the circuitry defined using photography. The photoresist is first deposited and baked. It is then exposed to ultraviolet light through a mask that defines the pattern of the circuit. Finally the exposed wafer is developed and baked.

*Etching:* In order to define the circuits, the exposed material is etched away. This may be accomplished through wet etching, where the wafer is immersed in a liquid that removes the exposed material, or dry etching, where the operation is carried out by exposure to gas that produces a chemical reaction resulting in the removal of unprotected material.

*Ion implantation:* Selected impurities are introduced in a controlled fashion to change the electrical properties of the exposed portion of the layer.

*Photoresist strip:* The photoresist remaining on the wafer is removed by a process similar to etching.

*Inspection and measurement:* The layer is inspected and measured to identify defects and guide future operations.

This sequence of operations is repeated, with variations, for each layer of circuitry on the wafer. Detailed descriptions of the technologies used in wafer fabrication can be found in texts on this subject such as Doering and Nishi (2007).

In wafer probe, the individual circuits, of which there may be hundreds on each wafer, are tested electrically, and the locations of circuits that fail to meet specifications are recorded for each wafer. The wafers are then cut into individual circuits and the defective circuits discarded.

In assembly the circuits are placed in plastic or ceramic packages that protect them from the environment, and leads are attached to allow the devices to be mounted on printed circuit boards. Since a given circuit may be packaged in many different ways, there is a great proliferation of product types at this stage. After leads are attached and the package sealed and tested for leaks and other defects, the product is sent to final test.

The goal of the testing process is to ensure that customers receive a defect-free product by using automated testing equipment to determine whether each integrated circuit is operating at the required specifications. An interesting aspect of the final testing process is that of binning, where a device manufactured to a particular specification may fail to meet that specification but may be acceptable for a lower grade of product (Moench et al. 2017). This also allows the possibility of downgrading, where a higher-grade product may be substituted for a lower grade one if necessary. For example, a microprocessor with a clock speed of 4 MHz can be used to meet demand for a 3 MHz device if the manufacturer deems it profitable to do so. Hence a given product may have several alternative bills of material, which the planning level must consider in its decisions. The ongoing evolution of products and manufacturing processes often results in significant differences in production capabilities and costs between different wafer fabs, with more recently built plants capable of processing both older and newer products while older fabs can process only older ones. A similar phenomenon occurs within individual fabs; newer equipment can process both older and newer products, while older equipment cannot produce the newer devices that usually have smaller feature sizes (line widths). Hence a given product can be produced using several alternative bills of material (due to the possibility of substituting different devices raised by binning), alternative processing recipes within an individual fab, and also in alternative fabs, whose production costs may differ substantially.

Several factors make production planning and scheduling in the semiconductor industry particularly difficult (Moench et al. 2017). The presence of alternative processing recipes and alternative equipment for different processing steps within fabs complicates the interface between the planning and scheduling levels, making it difficult for the planning level to gain an accurate picture of the capabilities of the production facilities. The presence of alternative bills of material requires any effective planning system to consider multiple production units (fabs). Uncertainty is pervasive throughout the supply chain, which is subject to significant business cycles and short product life cycles. The overall yield of the supply chain, defined as the fraction of raw material introduced into the fab that completes all four stages of the manufacturing process as salable devices at their original specification, varies over time due to rapid evolution of technology and products that often does not allow time for extensive process and equipment optimization. Finally, the extremely high capital cost of production facilities and skilled personnel, especially for wafer fabs, means that these facilities must operate at high levels of utilization to be economically viable. As we shall see in Chap. 2, this requires them to operate in a regime where the anticipation function consisting of fixed planned lead times and maximum capacity loading becomes increasingly problematic.

The focus of supply chain management in the semiconductor industry is on reducing production costs and cycle times while improving quality and delivery time performance. Major factors affecting costs are yield, labor, materials, inventory, equipment and facility depreciation, and equipment utilization. Historically, the major forces in the industry were the integrated high-volume manufacturers of standard products such as microprocessors and memory devices. In these supply chains, a common approach is to buffer the wafer fabs against fluctuations in external demand by holding inventories of probed die, referred to as die bank inventories, between the front-end and back-end operations. This allows the fabs to operate in a make-to-stock mode at a level production rate, with production lots rarely being associated with a specific customer order or due date. Together with the high capital costs, this results in strong emphasis on maintaining high throughput and equipment utilization while reducing both the mean and the variance of cycle times and inventories.

Recent decades have seen the rapid growth of foundry operations, where products designed by other companies are manufactured to order, due to the ever-increasing capital cost of new fabs (Chou et al. 2007). This prevents many firms that design circuits from building their own production facilities and motivates manufacturers to seek economies of scope by producing devices for other firms using their own excess capacity. Hence these foundries produce a wide range of products for different customers, usually with a common manufacturing process, emphasizing the effective management of individual orders to maintain reliable delivery performance. The make-to-order nature of foundry operations prevents the producing firm from using inventories to smooth production rates at the factory. Many firms address these issues through increasingly complex contracting arrangements between designing and producing firms (Shirodkar and Kempf 2006; Milne et al. 2015). These contracting arrangements impose a number of constraints on the PPC system, since they require the firm to satisfy a variety of clauses such as minimum volumes, maximum change in order from month to month, and significant changes in prices depending on the timing and volume of orders (Knoblich et al. 2011, 2012, 2015; Wu et al. 2014). Extensive discussion of these issues is given in Moench et al. (2017).

#### 1.2.1.1 The PPC Problem

The semiconductor industry has been a global industry from its earliest days, with many firms operating plants in North America, Europe, and Asia. Thus the PPC system must manage the flow of material through this global network to customers. This involves determining which demands will be met from which factories; how probed die will be moved from front-end to back-end operations; and how packaged and tested devices will be routed to their final customers through distribution centers in different countries.

The overall planning cycle in a typical firm begins with the collection of demand forecasts from different sources in the firm. The sales and marketing organization obtains sales estimates from its field sales force, which are then aggregated by geographical regions (e.g., North America, South Asia) and product families (server CPUs, laptop CPUs, auxiliary chipsets such as graphics processors, flash memory). This information is supplemented with input from the research and development organizations as to when new products and manufacturing processes will become available for introduction into the plants, and plans for the construction of new plants and the decommissioning of older ones.

Given these complex and constantly changing inputs, which are subject to considerable uncertainty, the PPC system must consider multiple production units (fabs, probe, assembly, and testing facilities) as well as possibly outsourcing of certain portions of the process. The presence of alternative manufacturing routings, both within facilities and across different facilities, and alternative BOMs due to binning adds considerable complexity. Fordyce et al. (2011) and Leachman et al. (1996) present detailed descriptions of PPC systems implemented at IBM and Harris semiconductor, respectively. Ahmadi et al. (1999) discuss another system implemented at Advanced Micro Devices. Eventually, however it is accomplished, a build schedule of specific wafer types for each manufacturing process within each facility must be obtained, and the quantity of each type of wafer to be released to each facility over time determined. The PPC systems described above all accomplish this, although in somewhat different ways.

The productive capacity of a wafer fab resides in the different types of production equipment located within the facility. This equipment is usually organized by function, according to the basic manufacturing process it performs, such as ion implantation, lithography, dry etch, metal deposition, and so on. A typical fab will generally have between 80 and 100 such equipment groups, which we shall refer to as *workcenters*. The equipment within a workcenter is seldom completely homogeneous; it usually represents several generations of technology, so that not all products can be processed on all machines, while some process steps can be processed on several different machines within the workcenter with different processing times and yields (Leachman and Carmon 1992; Johri 1994; Bermon and Hood 1999). Workcenters often contain additional equipment supporting the basic process being performed; for example, an etching workcenter may contain cleaning equipment to remove particulate contamination from wafers prior to etching and metrology equipment for process monitoring.

Wafers move through the fab in lots whose size is dictated by the material handling interfaces of the automated production and material handling equipment within the fab. The release of a lot into the fab requires considerable processing to prepare the raw material and set up the information systems necessary to track the lot and its associated data throughout the production process. Once a lot is released to the shop floor, its progress is tracked using an automated Manufacturing Execution System. Responsibility for its progress, and control of when and how it is processed at the different workcenters it requires, now rests with the management of the workcenters through which it must pass. This is accomplished by a combination of detailed scheduling, generally involving some form of dispatching rule, with manual intervention and expediting by shop-floor management as necessary (Moench et al. 2013).

Thus the primary control variable exercised at the planning level is the release of lots into the fab; once a lot is released, its progress is controlled by fab management and cannot be directly influenced by the planning level, although shop-floor information systems generally allow monitoring of its progress. A number of alternative approaches for managing the release of lots into the fab have been discussed in the literature (Uzsoy et al. 1992, 1994; Moench et al. 2013). Several of these are based on variations of the workload control concept discussed in Chap. 4, while others use an optimization model. In order to be effective, however, an optimization model for release planning must incorporate a mechanism for assessing the impact of its decisions (in this case, weekly order releases) on the ability of the workcenters to complete processing of the lots by the desired time. The clearing functions we introduce in Chap. 7 form the basis of this mechanism, serving as anticipation functions that permit the fab-level planning problem to estimate the performance of the fab under specific release decisions. This allows us to formulate optimization models that account for these impacts and have been tested on several large data sets drawn from this industry (Kacar et al. 2012, 2013, 2016). We present some of these results in Chap. 10.

### 1.2.2 CD/DVD Manufacturing

Production of optical storage media (CDs, DVDs) has a much simpler material flow structure than semiconductor manufacturing. Hence we use this example to illustrate the PPC problem encountered in a small- to medium-sized company producing discrete products with a simple bill of material (BOM) structure in a make-to-order environment. The central issue is the coordination of the production targets and the resulting state of the production unit with customer orders; material coordination across multiple production units is not necessary.

We consider a customer-driven optical storage media producer making about 90,000,000 CDs and DVDs per year to fill approximately 31,500 customer orders. The basic production process is divided into five stages as follows:

*Premanufacturing:* Each customer provides the data to be reproduced, which is written to a silicon or glass master about 20 cm in diameter, referred to as a stamper, with a laser beam that records the information to be replicated by producing tiny indentations on the polycarbonate known as pits and lands. The stamper is then used to make a disc ingot for further duplication.

*Graphics:* The graphics department produces the covers of the discs according to the design provided by the customer. This involves making print stencils for the different printing technologies (serigraphy or offset printing).

*Manufacturing:* After the stamper is made, the disc duplication using molding machines begins. A syringe injects a heated liquid polycarbonate at approximately 360 °C, producing a disc that contains digital information but cannot be scanned since at this stage it is completely transparent. After a brief cooling, the side of the disc containing the information is covered with a layer of silver, aluminum, or gold, followed by a layer of lacquer that reflects the laser beam, permitting the disc to be read. After an inspection, the discs are collected on a spindle.

*Printing:* Depending on their quality requirements, the disc labels are applied to the discs using either serigraphy or offset printing presses. The principal difference between these techniques is that for serigraphy printing machines, the mesh has to be cleaned after each order and the colors have to be prepared beforehand. The offset printing machines use the CMYK colors (cyan, magenta, yellow, black), requiring only a stencil change after each order.

*Packaging:* All orders are sent to fully- or semiautomated machines where the covers, booklets, inserts, etc. are added to the discs and packaged according to the customer's need (boxes, paper bags, trays, etc.).

The premanufacturing and graphics stages do not pose substantial planning problems, so they are not included in the following discussion.

Figure 1.2 depicts the workflow through the production facility, which is organized as a flexible flow shop. Work flows from left to right in the diagram as each product moves through a subset of the 29 machines on the shop floor. The flow of materials required for printing and packing is depicted by the vertical arrows into the respective departments. The manufacturing stage is organized into two areas, CD and DVD production, consisting of ten and six identical machines, respectively. The printing stage is organized into three areas by printing technology: the serigraphy area (SD) (three machines) and two offset areas (KOD, MOD) consisting of three machines and a single machine, respectively. The packaging stage consists of four areas, V1–V4, for different packaging with usually two machines, which may not be identical, in each area. The system operates 24 h per day, 7 days a week. Bottlenecks generally occur at the printing and packaging stages due to the varying product mix.

The manufacturing system is a flexible flow shop (Pinedo 2012), where all products are processed by exactly one machine in each stage, yielding 2 × 3 × 4 = 24 possible production routes, all of which occur in practice. Since machines are only interchangeable within areas, temporary bottlenecks can occur, so releasing an order mix that leads to balanced workloads across the areas is desirable.

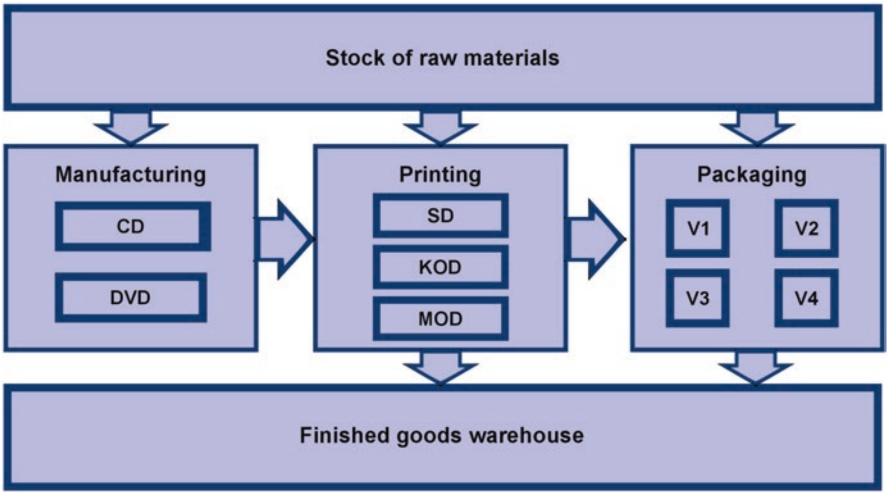


**Fig. 1.2** Structure of optical media production system

#### 1.2.2.1 The PPC Problem

The material flow and BOM structure in this application are both relatively simple. Although packing can be considered an assembly operation since it requires the simultaneous availability of the discs, packing material, and booklets, separate assembly orders are not necessary. Therefore, the entire three-stage production system can be treated as a single production unit.

Production planning is driven by customer orders. The time between the arrival of an order and its latest start date ranges from zero to about 10 days and follows a historically known distribution, largely determining the opportunities for production smoothing and load balancing across machines. The delivery date assigned to a customer order may be delayed if the customer fails to provide necessary material or information (e.g., booklets). Hence matching demand and capacity by controlling order acceptance is difficult because due date changes can lead to situations where demand temporarily exceeds capacity which must be resolved at the planning level.

The number of discs in a customer order varies substantially. In order to reduce variability in the material flow, large customer orders are divided into smaller production orders of about 3500 discs that are processed independently. The processing times of these production orders in the manufacturing and printing departments are of the order of 0.1–0.25 shifts on average, depending on the machine group, and substantially lower on the packing lines. This represents less than 10% of the average cycle time of the overall process.

Since the manufacturing system operates 24 h a day, 7 days a week, capacity cannot be expanded by overtime. Releasing orders based on their planned due dates alone may leave some machines temporarily idle while overloading others, resulting in poor due date performance and low utilization. Hence both load leveling over time and load balancing among workcenters are necessary to avoid releasing an order mix that causes shifting bottlenecks. This is partly addressed by informal dispatching mechanisms: the shop-floor Manufacturing Execution System (MES) provides information on the queue lengths at the machine groups, allowing sequencing decisions at upstream workcenters to prioritize work away from overloaded workcenters.

In the terminology of this chapter, the planning level sets preliminary due dates for the customer orders (orders are accepted whenever possible), derives production orders from the customer orders (which in this case is rather simple), and releases the production orders to the manufacturing system which is organized as a single production unit. While the material coordination task is very simple, controlling the workload at the workcenters is important due to the inflexible working hours and the possibility of temporary bottlenecks. While at the time of this analysis, workload control was performed by the dispatching policy in the informal system, redesigning the production planning level would formalize this logic at the order release level, coordinating order release and order acceptance/due date setting to reduce WIP and cycle time. This can be done by various methods discussed in the following chapters and is possible due to the presence of a significant time interval between order arrival and latest release time.

## 1.3 Contributions and Perspectives of This Volume

The remainder of this volume assumes that production orders specifying the product or item to be manufactured, the quantity to produce and the required due dates, are available and have been generated by the planning level of whatever PPC system the firm in question is using. However, it may not be possible to complete all production orders by their required due dates due to discrepancies between available and required capacity. Hence order release must perform production smoothing and control the system state of the production units simultaneously. This requires models of the production units that can accurately anticipate the impact of release decisions on the time-dependent WIP, cycle times, and output. The volume mainly deals with mathematical programming models for order release that incorporate such anticipation models.

Much of the work in this volume focuses on a particular family of anticipation functions, referred to as clearing functions, which seek to represent the potential output of a production unit in a planning period as a function of a set of variables describing the state of its workcenters at the start of that period. This approach, which originated in the late 1980s, has been the subject of renewed interest in the last two decades. The approach can be viewed as treating each workcenter as a queue or queueing network, and then building a metamodel for this queueing system that is amenable to incorporation in a mathematical programming model. This leads to nonlinear mathematical programming formulations that differ substantially from previous approaches based on exogenous planned lead times. We compare these two formulation approaches in terms of their view of production systems, and the information they can provide to users. We then discuss alternative approaches to the formulation of clearing functions, their incorporation into mathematical programming models, and the strengths and weaknesses of the approach. Our objective is to bring together in a single volume the relevant literature on this approach, which spans a wide range of journals over an extended period of time, so that the interested researcher or practitioner has easy access to this material.

## 1.4 Outline of This Volume

In Chap. 2 we motivate the main ideas of the book by discussing the relation between workload and cycle times in production resources that can be modeled as queueing systems.

Chapter 3 discusses the concepts of structuring the planning level that generates the production orders, beginning with the MPC concept based on MRP/MRPII and the hierarchical planning concept that provides the basis for Advanced Planning Systems (APS). Chapter 4 then describes the workload control (WLC) concept that aims at *controlling*, as opposed to merely predicting, cycle times by releasing the right amount and mix of production orders, and describes workload control systems for a single production unit that do not incorporate optimization models. Chapter 5 then describes optimization models for planning the aggregate material flow and order release for multiple planning periods using the conventional coordination mechanism of planned lead times and maximum allowable capacity loading, assuming that the planned lead times remain constant over time.

Chapter 6 discusses the more complex situation that arises when the planned lead times can vary over time and then examines multi-model production planning approaches that use a detailed simulation or scheduling model to represent the interior workings of the production units while capturing goods flow decisions in a mathematical programming model. These approaches combine well-known mathematical modeling techniques familiar to practitioners (usually linear programming and discrete-event simulation), but the principles governing their convergence and solution quality are not yet well understood, although recent studies are beginning to shed light on these aspects.

Chapters 7 and 8 discuss an alternative anticipation mechanism, the clearing function, proposed by a number of authors (Graves 1986; Srinivasan et al. 1988; Karmarkar 1989) starting in the late 1980s. In its most basic form, a clearing function represents the capabilities of a production resource as a relation between some estimate of the work available to the resource in a planning period and its expected output in that period. Chapter 7 focuses on univariate clearing functions. We first discuss alternative ways to estimate clearing functions, considering ideas from queueing models, traffic modeling, and empirical data analysis. We implement the most common form of a clearing function, a saturating, concave non-decreasing function of the workload, in an optimization model that differs in several ways from the conventional models based on planned lead times and capacity limits. We then identify a major disadvantage of this formulation, which we address by the Allocated Clearing Function formulation. We also analyze the dual of this formulation, which allows the computation of dual prices for resources with utilization below 1, which the LP models of Chap. 5 cannot do.

Chapter 8 examines multivariate clearing functions that have been developed to address the dependence of the output of a production resource on the longer history of the process as well as problems where there are significant interactions between products in terms of their capacity consumption, typically setup times. This chapter examines the benefits of these enriched representations that, however, often lead to non-convex optimization models. Chapter 9 examines the extension of the clearing function approach to lot-sizing problems, while Chap. 10 is devoted to applications of the clearing function concept to a number of different areas including release planning in semiconductor wafer fabrication, dynamic pricing, and modeling of process improvements. Chapter 11 concludes the volume with a summary of the principal insights from the work reported and discusses several directions for future research at length.

## References

Ahmadi J, Benson R, Supernaw-Issen D (1999) Supply chain production planning. In: Ciriani TA, Gliozzi S, Johnson EL, Tadei R (eds) Operational research in industry. Macmillan, Basingstoke

Bermon S, Hood SJ (1999) Capacity optimization planning system (caps). Interfaces 29(5):31–50

Bertrand JWM, Wortmann JC, Wijngaard J (1990) Production control: a structural and design oriented approach. Elsevier, Amsterdam

Buzacott JA, Corsten H, Gössinger R, Schneider HM (2013) Production planning and control: basics and concepts. Munich, Oldenbourg Verlag Munchen

Chou YC, Cheng CT, Yang FC, Liang YY (2007) Evaluating alternative capacity strategies in semiconductor manufacturing under uncertain demand and price scenarios. Int J Prod Econ 105:591–606

Cowling P, Rezig W (2000) Integration of continuous caster and hot strip mill planning for steel production. J Sched 3:185–208

de Kok AG, Fransoo JC (2003) Planning supply chain operations: definition and comparison of planning concepts. In: de Kok AG, Graves SC (eds) Handbooks in operations research and management science, Supply chain management: design, coordination and operation, vol 11. Elsevier, Amsterdam, pp 597–675

Doering R, Nishi Y (eds) (2007) Handbook of semiconductor manufacturing technology. CRC Press, Boca Raton

Floudas CA, Lin X (2004) Continuous-time versus discrete-time approaches for scheduling of chemical processes: a review. Comput Chem Eng 28:2109–2129

Floudas CA, Lin X (2005) Mixed integer linear programming in process scheduling: modeling, algorithms and applications. Ann Oper Res 139:131–162

Fordyce K, Degbotse A, Milne RJ, Waite J, Wang CT, Denton B, Orzell R, Chang CH, Lyon P, Rice R (2011) The ongoing challenge: creating an enterprise-wide detailed supply chain plan for semiconductor and package operations. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning in the extended enterprise: a state of the art handbook, vol 2. Kluwer, New York, pp 313–388

Graves SC (1986) A tactical planning model for a job shop. Oper Res 34(4):522–533

Gunther HO, Van Beek P (2003) Advanced planning and scheduling solutions in process industry. Springer, Heidelberg

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Jacobs FR, Berry WL, Whybark DC, Vollmann TE (2011) Manufacturing planning and control for supply chain management. McGraw-Hill Irwin, New York

Johri P (1994) Overlapping machine groups in semiconductor wafer fabrication. Eur J Oper Res 74:509–518

Kacar NB, Irdem DF, Uzsoy R (2012) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. IEEE Trans Semicond Manuf 25(1):104–117

Kacar NB, Moench L, Uzsoy R (2013) Planning wafer starts using nonlinear clearing functions: a large-scale experiment. IEEE Trans Semicond Manuf 26(4):602–612

Kacar NB, Moench L, Uzsoy R (2016) Modelling cycle times in production planning models for wafer fabrication. IEEE Trans Semicond Manuf 29(2):153–167

Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123

Knoblich K, Ehm H, Heavey C, Williams P (2011) Modeling supply contracts in semiconductor supply chains. In: Jain S, Creasey RR, Himmelspach J, White KP, Fu MC (eds) Proceedings of the 2011 Winter Simulation Conference. IEEE, Phoenix, pp 2108–2118

Knoblich K, Heavey C, Williams P (2012) An evaluation of an option contract in semiconductor supply chains. In: Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM (eds) Proceedings of the 2012 Winter Simulation Conference. IEEE, Berlin

Knoblich K, Heavey C, Williams P (2015) Quantitative analysis of semiconductor supply chain contracts with order flexibility under demand uncertainty: a case study. Comput Ind Eng 87:394–406

Leachman RC, Carmon TF (1992) On capacity modeling for production planning with alternative machine types. IIE Trans 24(4):62–72

Leachman RC, Benson RF, Liu C, Raar DJ (1996) Impress: an automated production planning and delivery quotation system at Harris corporation-semiconductor sector. Interfaces 26(1):6–37

McKay KN, Wiers VCS (2004) Practical production control: a survival guide for planners and schedulers. J. Ross Publishers, Boca Raton

Milne RJ, Wang CT, Denton B, Fordyce K (2015) Incorporating contractual arrangements in production planning. Comput Oper Res 53:353–363

Missbauer H, Hauber W, Stadtler H (2011) Developing a computerized scheduling system for the steelmaking–continuous casting process. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise: a state of the art handbook, vol 2. Springer, New York, pp 461–488

Moench L, Fowler JW, Mason SJ (2013) Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis and systems. Springer, Berlin

Moench L, Uzsoy R, Fowler J (2017) A survey of semiconductor supply chain models part I: semiconductor supply chains and strategic network design. Int J Prod Res 56(13):4524–4545

Pinedo M (2012) Scheduling. Theory, algorithms, and systems. Springer, New York

Pinedo M, Chao X (2005) Planning and scheduling in manufacturing and services. Springer, New York

Schneeweiss C (2003) Distributed decision making. Springer, Berlin

Shirodkar S, Kempf KG (2006) Supply chain collaboration through shared capacity models. Interfaces 36(5):420–432

Srinivasan A, Carey M, Morton TE (1988) Resource pricing and aggregate scheduling in manufacturing systems. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh

Stadtler H, Kilger C, Meyr H (2015) Supply chain management and advanced planning. Concepts, models, software, and case studies. Springer, Berlin

Tang L, Liu J, Rong A, Yang Z (2001) A review of planning and scheduling systems and methods for integrated steel production. Eur J Oper Res 133:1–20

Uzsoy R, Lee CY, Martin-Vega LA (1992) A review of production planning and scheduling models in the semiconductor industry part I: system characteristics, performance evaluation and production planning. IIE Trans 24:47–61

Uzsoy R, Lee CY, Martin-Vega LA (1994) A review of production planning and scheduling models in the semiconductor industry part II: shop-floor control. IIE Trans 26:44–55

Vollmann TE, Berry WL, Whybark DC, Jacobs FR (2005) Manufacturing planning and control for supply chain management. McGraw-Hill, New York

Wu X, Kouvelis P, Matsuo H, Sano H (2014) Horizontal coordinating contracts in the semiconductor industry. Eur J Oper Res 237:887–897

# Chapter 2
# Workload and Cycle Time in the Production Unit

Our description of the PPC problem in Chap. 1 identified the effective management of cycle times as a critical link between the planning level and the realized performance of the production units it seeks to coordinate. Most of the PPC systems prevalent in industry today approach this issue through planned lead times and maximum capacity loading, assuming that as long as the capacity loading does not exceed the agreed-upon maximum level, the production units will be able to complete work within the planned lead time with high probability. This chapter argues that reliance on exogenous planned lead times represents a significant drawback of this approach because cycle times through a production unit are, in fact, an outcome of the work release decisions made by the PPC system. Since this dependence between cycle times and work release decisions lies at the center of the problems addressed in this volume, we now discuss the relationship between a production unit's workload and cycle time in more detail.

## 2.1 Preliminaries

Per Chap. 1, we follow Bertrand et al. (1990) in viewing a production system as a network of production units—groups of production resources such as machines and personnel that must perform specific tasks (e.g., particular operations on particular production orders by a specified due date) and can exhibit different material flow structures such as job shop, flow shop, manufacturing cells, etc. Detailed scheduling and resource allocation decisions within the production unit are not visible to, or subject to the control of, the planning level. Hence the construction of optimization models for planning releases into production units, the primary concern of this volume, must begin with a viable model of an individual production unit that permits anticipation of their behavior by the planning level. Since we seek optimization models that are applicable to a wide variety of manufacturing environments, we must seek general laws describing the behavior of production units. Laws of this

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_2

type have been studied extensively in the field of production economics (Fandel 1991; Hackman 2008), which views production as a transformation process that converts input factors (such as labor and machines) into goods and services for internal or external customers.

Our focus is on the relationship between the work release decisions made at the planning level and the performance measures, particularly work-in-process (WIP) inventory levels, cycle times and output, of the individual production units. Since the planning level seeks to ensure that supply from the production units matches demand for the final products in some "optimal" way, the cycle time, the delay between work being released into the production unit and the completion of its processing at that production unit, takes on a critical role. Hence we are primarily interested in the time dimension of the relationship between factor input, whose timing is determined by work release, and the time the work is completed and output of the finished product occurs.

Since, as argued in Chap. 1, the primary actionable decision of a PPC system is the quantity and timing of work releases into the production units, the evolution of resource workloads over time is determined by decisions at the planning level. Queueing models (Buzacott and Shanthikumar 1993; Curry and Feldman 2000; Hopp and Spearman 2008), which represent production systems as networks of queues, provide useful tools for examining the consequences of planning decisions on the WIP levels, cycle times, and output realized at the production units.

## 2.2 Insights from Queueing Models

A production unit consists of one or more workcenters, groups of possibly nonidentical machines that are managed on the shop floor as a unit. For simplicity of exposition, we shall frame our discussion in terms of a single production resource, such as a machine, whose behavior can be modeled as a queueing system. While production units may have multiple machines and complex structures within themselves, the problem of how to anticipate their behavior at the planning level is the same in its essence, although the resulting queueing models are more complex. Beyond a certain level of complexity, simulation models are required to describe the behavior of many production units as discussed in later chapters.

We consider a single-machine workcenter modeled as a queueing system, closely following the development in Chap. 8 of Hopp and Spearman (2008). Production orders, which we shall refer to as jobs to avoid confusion with lot-sizing models, are released to the production unit and—possibly after being processed at some workcenters that are not modeled explicitly—arrive at the workcenter under consideration according to some stochastic process. The interarrival times between jobs follow a known probability distribution $F_a(.)$ with mean $t_a$ and squared coefficient of variation (SCV) $c_a^2$. The effective processing times of the jobs, which incorporate the effects of disruptions such as setup times, machine failures, and scrap, are independent of their arrival times and follow a known probability distribution $F_e(.)$ with mean $t_e$ and SCV $c_e^2$. Hence the average arrival rate is $\lambda = 1/t_a$ and the average service rate $\mu = 1/t_e$. The cycle time spent by a job in this queueing system consists of the time it spends in the queue and the time to complete its processing (including setup), and is a random variable determined jointly by the two probability distributions $F_a(.)$ and $F_e(.)$. A well-known result (Kingman 1961; Hopp and Spearman 2008) states that the steady-state expected cycle time $T$ of this *G/G/1* queue (Kendall 1953) is approximated by

$$T = \frac{\left(c_a^2 + c_e^2\right)}{2} \frac{u}{1-u} t_e + t_e \tag{2.1}$$

where the average utilization of the resource is given by $u = t_e/t_a$. Equation (2.1) suggests that the expected cycle time is influenced by four quantities: the variabilities of the arrival and service processes, expressed by $c_a^2$ and $c_e^2$, respectively; the mean effective processing time $t_e$; and the average utilization $u$ of the workcenter, which, in turn, is jointly determined by $t_a$ and $t_e$. The effect of the average utilization $u$ is of particular interest for production planning models. The release decisions made by the planning level that specify how much work to release to a given production unit in a planning period determine the mean arrival rate of work $\lambda = 1/t_a$ to the workcenter.

Figure 2.1 shows the behavior of the average cycle time $T$ per Eq. 2.1 as the average utilization $u$ and the variance term $C = \left(c_a^2 + c_e^2\right)/2$ vary. $T$ increases nonlinearly with $u$, eventually tending to infinity as $u$ approaches 1. This behavior shows that the planning level's work release decisions affect the average cycle time; $T$ is endogenous to the planning decision, not an exogenous parameter unaffected by the planning process.

Another important observation from Eq. 2.1 is that $T$ is also affected by the variability $c_a^2$ in the material flow into the workcenter and the variability $c_e^2$ of the production process itself. The influence of $c_a^2$ is particularly important since the arrival

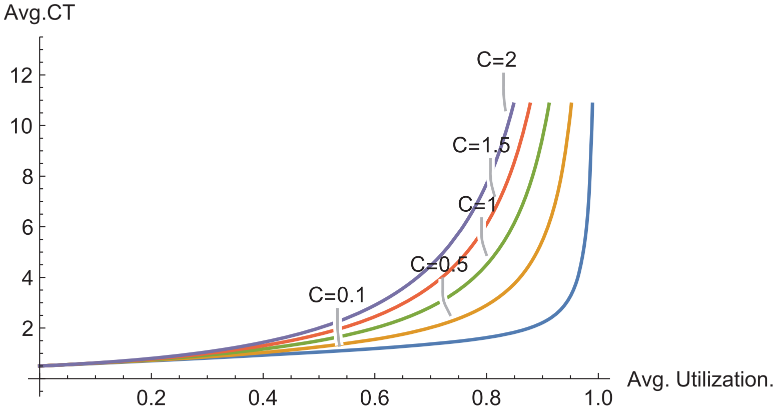


**Fig. 2.1** Behavior of average cycle time $T$ of a *G/G/1* queue

process to a workcenter is determined by the departure processes from the upstream workcenters that provide its inputs. Hence the average cycle time $T$ at a workcenter is affected by how the workcenters upstream of it are managed; variability at upstream workcenters will affect the performance of those downstream, as discussed by Hopp and Spearman (2008) and Godinho Filho and Uzsoy (2014). This functional relationship between average WIP and output, average flow time, and possibly other performance measures of the production unit is referred to as the *characteristic* or *operating curve* in the literature (Aurand and Miller 1997; Schoemig 1999) and is often estimated by simulation (Yang et al. 2006).

Representing our basic workcenter as a *G/G/1* queue allows us to invoke another fundamental queueing result. In a production context, the number of customers in the queueing system (in the queue or at a server) at a given point in time corresponds to the amount of work in process inventory (WIP) at the workcenter, which is a random variable we shall denote by WIP, with $W = E[\text{WIP}]$ denoting the expected WIP level expressed as number of customers or, in our context, jobs. If WIP is measured in units of the product or amount of work (standard hours) the queueing relationships given below must be modified accordingly. Following standard queueing analysis, let us also assume that we observe the system over a long period of time, such that the average rate of arrivals to the production unit is equal to its average processing rate. Thus the system is stable with no unbounded increase in the WIP quantity, and the expected throughput rate $X$ of the system, the average rate at which completed work leaves the workcenter, will be $X = \lambda = 1/t_a$. Under these conditions, Little's Law (Little 1961; Hopp and Spearman 2008) gives

$$W = XT = \frac{T}{t_a} \tag{2.2}$$

This expression has several important implications. For the purposes of managing a production system to achieve a given average throughput rate $X$, the average WIP level $W$ and average cycle time $T$ are directly proportional. A given throughput rate $X$ can be achieved either by controlling the average cycle time $T$ to achieve a desired average WIP level $W$ or by controlling the average WIP level $W$ to achieve an average cycle time of $T = W/X$. Loosely speaking, the former approach is associated with "push" systems such as MRP, where work is released into the production unit to meet due dates derived from customer orders or from forecasts of future demand. The latter is associated with "pull" systems such as the kanban system used in the Toyota Production System (Sugimori et al. 1977; Liker 2004). An excellent discussion of the distinctions between, and relative merits of, push and pull systems is given by Hopp and Spearman (2004).

Combining Eqs. (2.1) and (2.2), the expected WIP level of the steady-state queue is

$$W = \frac{T}{t_a} = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)\frac{t_e}{t_a} + \frac{t_e}{t_a} = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u^2}{1-u}\right) + u \tag{2.3}$$

The average utilization $u$ can be interpreted as the long-run fraction of time the resource will be busy, and thus producing output. Using the average WIP level $W$ as a measure of the resource's workload, i.e., the amount of work available for it to process, and solving for $u$ in terms of $W$ yields a quadratic in $W$ whose nonnegative solution is given by

$$u = \frac{-(W+1)+\sqrt{(W+1)^2+4(C-1)W}}{2(C-1)} \text{ for } C \neq 1 \tag{2.4}$$

where $C = \left(c_a^2 + c_e^2\right)/2$. When $C = 1$, representing an *M/M/1* queue with exponentially distributed interarrival and service times, Eq. (2.4) takes the simpler form

$$W = \frac{u^2}{1-u} + u = \frac{u}{1-u} \tag{2.5}$$

yielding

$$u = \frac{W}{(W+1)}. \tag{2.6}$$

As shown in Fig. 2.2, for given values of $t_e$ and $C$, $u$ is a monotonically nondecreasing concave function of $W$; as the average WIP level $W$ increases, $u$ increases at a decreasing rate. Intuitively, the higher the average WIP level $W$ in the system, the lower the probability $(1-u)$ that the resource will be idle due to lack of work; maintaining a given average throughput requires maintaining a certain average WIP level in the production unit.

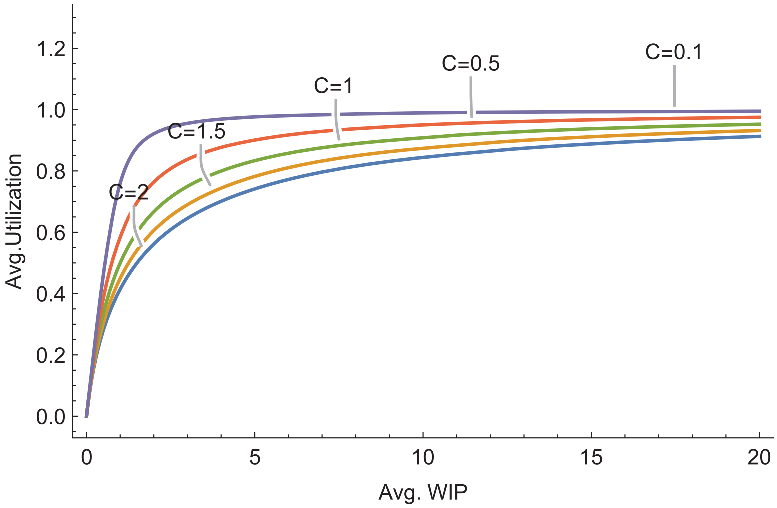


**Fig. 2.2** Average utilization as a function of average WIP

Relationships similar to Eq. (2.4) between the expected throughput of a queueing system and its expected WIP level can be derived analytically for a variety of queueing models, under steady-state or transient behavior (Selçuk et al. 2007; Asmundsson et al. 2009; Missbauer 2009). When closed-form analytical expressions are not available, empirical relations can be postulated by fitting an appropriate functional form to data obtained from either industrial observations (Häussler and Missbauer 2014) or a simulation model (Kacar et al. 2012; Kacar and Uzsoy 2015). We shall refer to these functions as *clearing functions*, since they represent the ability of the workcenter to process, or clear, some fraction of its workload in a planning period. They are the central construct of interest to this volume, discussed in Chaps. 7 and 8.

Equations (2.1) and (2.2) together determine the relationship between average WIP and average cycle time. Substituting Eq. (2.4) or Eq. (2.6) for $u$ into Eq. (2.1) yields

$$T = \begin{cases} \dfrac{t_e(C-1)\left(1-W+\sqrt{(W+1)^2+4(C-1)W}\right)}{2C-1+W-\sqrt{(W+1)^2+4(C-1)W}} & \text{for } C \neq 1 \\ t_e(W+1) & \text{for } C = 1 \end{cases} \tag{2.7}$$

The average cycle time increases linearly for the *M/M/1* case where $C = 1$. For $C < 1$ the slope is smaller for low WIP levels in both single- and multiple-server systems since there will be (almost) no queueing delay for low WIP levels.

Together, Eqs. (2.4) and (2.7) imply that given the queueing characteristics of the production unit, once the average WIP level is determined, the average utilization and cycle time are determined as well. This observation motivates the Workload Control framework presented in Chap. 4.

A number of caveats are, however, in order. The discussion above assumes that the given input rate $\lambda$ unambiguously determines the utilization of the workcenter. This is not the case in the presence of sequence-dependent setup times, since in this case the distribution of the effective processing time depends on the sequence in which the jobs are processed. If jobs are released without considering this issue, some form of batching and sequencing must be performed within the production unit to manage setup times. More WIP in the production unit gives its management more options to optimize the job sequence with respect to setups, reducing the total setup time for the given production quantities as average WIP increases. Several papers have examined the relationship between average WIP level and total setup time per period (Kekre 1984; Kim and Bobrowski 1995; Missbauer 1997; Thürer et al. 2012). Since these savings in setup time reduce the utilization required to produce a given output, the relationship between WIP and output illustrated in Fig. 2.2 is also affected. Informal production control rules applied at the shop-floor level, such as those that adapt the processing rate to the WIP level (e.g., Agnew (1976)), might also affect the operating curve.

The queueing analysis above suggests that the cycle time of a job through a production unit is a random variable whose distribution is affected by the utilization $u$ of the resources. The planned lead times used for order release planning by the PPC system are based on estimates of the cycle times through the production units making up the production system, so it is important to understand the structure of these cycle times. We now turn to this discussion.

## 2.3 Structure of Cycle Times in Production Units

The cycle time of a production order (job) through a production unit is the time elapsing between its release and its completion and is the sum of the cycle times of all operations performed on this order, accounting appropriately for any overlaps in time. The cycle time of the $k$'th operation of a job (a *throughput element* in Wiendahl 1995: 41 ff.) is usually defined as the time from the completion of the previous operation $k-1$ to the completion of operation $k$ and consists of any necessary delay between the completion of operation $k-1$ and the start of operation $k$ (such as curing time for a painting operation, or transportation time between locations), queueing time and the setup and processing time of operation $k$. In discrete manufacturing, the interoperation time, defined as the time from the completion of the previous operation $k-1$ to the start of operation $k$, consists mainly of waiting time due to queueing at capacitated resources and is often substantially higher than the operation time. Empirical studies report the ratio of operation time (raw process time in the terminology of Hopp and Spearman) to cycle time as about 0.1 in mechanical engineering (Wiendahl 1995: 37f.), and about 10% in the CD/DVD manufacturing system in Sect. 1.2.2. This is consistent with queuing-theoretical results where at high utilization the queuing time constitutes by far the greater part of the average cycle time in (Eq. 2.1).

Hence the variance of the cycle times is mainly determined by the variance of the waiting times, which is often fairly high in queueing systems. In the *M/M/1* queue, the conditional waiting time given that the server is busy on arrival is exponentially distributed. For the *G/G/1* queue, the waiting time distribution depends on the distributions of the interarrival and service times (Shortle et al. 2018: 320 ff.). In line with these analytical results, the empirical distribution of the cycle times at a workcenter often exhibits high variance, as illustrated in Fig. 2.3. The positive skewness due to very long cycle times experienced by a small fraction of the orders is typical of many production environments, and can be caused both by time-varying WIP levels at a workcenter and by expediting or delaying orders by dispatching; Ehteshami et al. (1992) illustrate the effect of expediting in the context of semiconductor wafer fabrication. Four priority classes can be distinguished in the figure; some orders are deliberately delayed for the reasons given in the legend. However, even the cycle times of the normal orders exhibit high variance, making it difficult to derive planned lead times from observed cycle times.

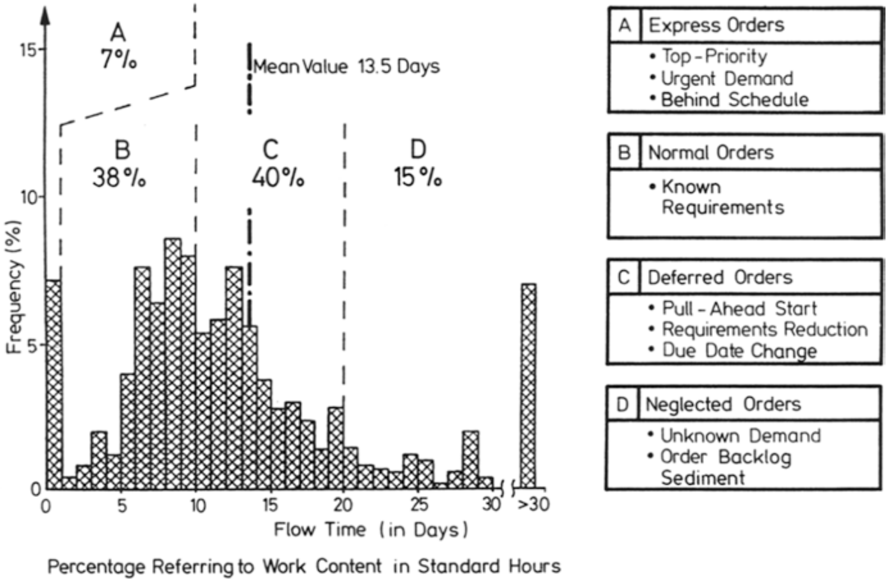


**Fig. 2.3** Distribution of the weighted cycle time of the orders processed at a lathe workcenter over 16 weeks; Wiendahl (1995): 30

The mean cycle time at a workcenter, which is of major concern in this book, is usually defined as the mean of the distribution of the individual job cycle times, which is also the standard definition in the scheduling literature (Pinedo 2012). Wiendahl (1995: 55ff) recommends using the weighted mean cycle time

$$\overline{T}_m^W = \frac{\sum_{j \in \Im} T_{jm} a_{jm}}{\sum_{j \in \Im} a_{jm}} \tag{2.8}$$

since it is less sensitive to priorities at the dispatching level than the unweighted mean cycle time. In Eq. (2.8) $T_{jm}$ denotes the observed cycle time of order $j$ at workcenter $m$, $a_{jm}$ the processing time (including setup time) of order $j$ at that workcenter, and $\Im$ the set of all operations represented in the observed sample of orders. This quantity represents an estimate of the average cycle time of each hour's worth of work processed at the workcenter in a certain time interval, the definition used in Fig. 2.3.

The importance of realistic lead times for the planning level and the large contribution of waiting time to the observed cycle times, at least at bottleneck workcenters, makes management of the waiting times an essential task for shop-floor management, and the derivation of accurate planned lead times from them crucial to effective operation of the planning level.

## 2.4 From the Production Unit to the Goods Flow Problem

Having described the behavior of a generic production unit, we are now in a position to relate the conceptual model of PPC systems developed in Chap. 1 to the vital statistics of our production unit: WIP, throughput, and cycle time. Since the production units are managed autonomously to meet the output targets determined by the planning level, the planning level must be able to estimate the impact of its requests, i.e., planned releases and output, on the ability of the production unit to meet them in a timely and cost-effective manner. Per Sect. 2.1, such a model must recognize the nonlinear relationship between average throughput $X$ and average cycle time $T$ as approximated by Eqs. (2.3), (2.7), or some similar relation. The task of the planning level is to release production orders into the production units such that they can carry out the processing necessary to meet demand in time. This requires coordination of activities across multiple production units across time. This, in turn, requires both effective management of the cycle times within each production unit to coordinate the timing of production with demand, and planning and control of the production-inventory system according to the product structure, including the determination of desired stock levels at the various stock points over time. Much of the complexity of the PPC task results from the interference of these two modeling and control tasks, and it is not surprising that PPC systems in practice emphasize one or the other of these tasks in order to keep complexity manageable. We now turn to the PPC frameworks that provide the basis for the developments presented in this book.

## References

Agnew C (1976) Dynamic modeling and control of some congestion prone systems. Oper Res 24(3):400–419

Asmundsson JM, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning models with resources subject to congestion. Nav Res Logist 56(2):142–157

Aurand S, Miller P (1997) The operating curve: a method to measure and benchmark manufacturing line productivity. In: 1997 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop ASMC 97 Proceedings, IEEE, Cambridge, pp 391–397

Bertrand JWM, Wortmann JC, Wijngaard J (1990) Production control: a structural and design oriented approach. Elsevier, Amsterdam

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs

Curry GL, Feldman RM (2000) Manufacturing systems modelling and analysis. Springer, Berlin

Ehteshami B, Petrakian R, Shabe P (1992) Trade-offs in cycle time management: hot lots. IEEE Trans Semicond Manuf 5(2):101–106

Fandel G (1991) Theory of production and cost. Springer, Berlin

Godinho Filho M, Uzsoy R (2014) Assessing the impact of alternative continuous improvement programmes in a flow shop using system dynamics. Int J Prod Res 52(10):3014–3031

Hackman S (2008) Production economics. Springer, Berlin

Häussler S, Missbauer H (2014) Empirical validation of meta-models of work centres in order release planning. Int J Prod Econ 149:102–116

Hopp WJ, Spearman ML (2004) To pull or not to pull: what is the question? Manuf Serv Oper Manag 6(2):133–148

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Kacar NB, Uzsoy R (2015) Estimating clearing functions for production resources using simulation optimization. IEEE Trans Autom Sci Eng 12(2):539–552

Kacar NB, Irdem DF, Uzsoy R (2012) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. IEEE Trans Semicond Manuf 25(1):104–117

Kekre S (1984) The effect of number of items processed at a facility on manufacturing lead time. Working Paper Series. University of Rochester, Rochester

Kendall DG (1953) Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann Math Stat 24(3):338–354

Kim S-C, Bobrowski PM (1995) Evaluating order release mechanisms in a job shop with sequence-dependent setup times. Prod Oper Manag 4(2):163–180

Kingman JFC (1961) The single server queue in heavy traffic. Math Proc Camb Philos Soc 57(4):902–904

Liker J (2004) The Toyota way: 14 management principles from the world's greatest manufacturer. McGraw-Hill, New York

Little JDC (1961) A proof of the queueing formula L =λw. Oper Res 9:383–387

Missbauer H (1997) Order release and sequence-dependent setup times. Int J Prod Econ 49:131–143

Missbauer H (2009) Models of the transient behaviour of production units to optimize the aggregate material flow. Int J Prod Econ 118(2):387–397

Pinedo M (2012) Scheduling. Theory, algorithms, and systems. Springer, New York

Schoemig AK (1999) On the corrupting influence of variability in semiconductor manufacturing. In: Proceedings of the Winter Simulation Conference, P. A. Farrington, H. B. Nembhard, D. T. Sturrock, G. W. Evans (eds), 1999, pp. 837–842, IEEE

Selçuk B, Fransoo JC, de Kok AG (2007) Work in process clearing in supply chain operations planning. IIE Trans 40(3):206–220

Shortle JF, Thompson JM, Gross D, Harris CM (2018) Fundamentals of queueing theory. Wiley, Hoboken

Sugimori Y, Kusunoki K, Cho F, Uchikawa S (1977) Toyota production system and Kanban system: materialization of just-in-time and respect for human system. Int J Prod Res 15(6):553–564

Thürer M, Silva C, Stevenson M, Land M (2012) Improving the applicability of workload control (Wlc): the influence of sequence dependent setup times on workload controlled job shops. Int J Prod Res 50(22):6419–6430

Wiendahl HP (1995) Load oriented manufacturing control. Springer, Heidelberg

Yang F, Ankenman B, Nelson BL (2006) Efficient generation of cycle time-throughput curves through simulation and metamodeling. Nav Res Logist 54(1):78–93

# Chapter 3
# Production Planning and Control Frameworks

The work described in this volume lies at the intersection of two streams of literature. The first of these addresses the structuring of the planning problem as a sequence of decisions made at different levels of the organization with different levels of information and different time frames. The second is related to the mathematical modeling techniques used to describe and solve planning problems formulated at different levels of the organization. We begin by reviewing the different ways in which the problem of planning and coordinating production in complex organizations has been addressed by presenting the two most widely used decision frameworks for designing and implementing PPC systems in this chapter.

## 3.1 Material Requirements Planning (MRP) and the Manufacturing Planning and Control (MPC) Framework

The steady increase in the scale of industrial operations over the course of the nineteenth and twentieth centuries brought the need for more sophisticated organizational structures and management tools to support the effective coordination of complex activities over wide geographical areas (Chandler 1962, 1980). These developments, together with the increasing complexity and technical sophistication of industrial products, rapidly rendered it impossible for any individual, or body of individuals, to have complete command of all the information needed to manage the entire organization effectively. The unsuccessful attempts at centralized planning in the totalitarian economies, despite the very high levels of resources dedicated to these exercises, serve only to underscore the difficulty of this undertaking. As a result, most firms decompose the planning function into a sequence of steps carried out by different groups, with each group's decisions defining the range of possibilities for those made by the next and successively adding detail until a workable

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_3

solution has been obtained. Initially such systems were almost completely manual, but various computational and informational aids were developed over time. An overview of the early development of such systems, which has been centered on the widely used Material Requirements Planning (MRP) approach systematized by Orlicky (1975), is given by McKay (2010). The resulting Manufacturing Planning and Control (MPC) framework, although subject to variation, has been widely adopted and represents the state of industrial practice across a wide range of industries. Hence we begin by discussing this framework in detail, which will allow us to identify potential improvements that may be obtained using the more elaborate approaches discussed in this volume.

In many discrete manufacturing industries, final products are assembled from a large number of components, each of which is itself manufactured using a multi-stage process. For example, in the mechanical engineering industries, the production of complex products such as construction machines or machine tools requires the coordination of tens of thousands of purchased, manufactured, and intermediate items, referred to as stock-keeping units (SKUs). In these environments, the PPC system has to coordinate thousands of parallel material flows for the components in order to guarantee the availability of *all* required components at the time of assembly. Demand for components is mostly a *dependent* demand required to meet the build schedule stated in the Master Production Schedule (MPS). The computation of this dependent demand through the well-known *bill of material (BOM) explosion* logic of Material Requirements Planning (MRP) (Vollmann et al. 2005) requires substantial computational power and, not surprisingly, was one of the first planning tasks to be automated when computers became available for business applications in the early 1960s.

Combining the current inventory levels and planned lead times of the production units (which are, of course, estimates of cycle times) with the BOM explosion yields time-phased net requirements for each SKU in the BOM. Since the gross requirements of each SKU are calculated from the lot sizes of the SKUs for whose production the SKU under consideration is required, lot sizing is an integral part of this calculation. Once time-phased net requirements and lot sizes for both production and purchase orders are calculated, the planned lead times of the production orders can be decomposed into planned lead times for the individual manufacturing operations at the workcenters performing them. This process also assigns the capacity requirements of these operations to the planning periods in which their performance is planned, permitting the calculation of time-phased capacity requirements using the Capacity Requirements Planning (CRP) procedure (Vollmann et al. 2005). Software systems performing these functions, termed MRP systems, represented the state of the art in the mid-1970s and constituted a tremendous advance over the independent demand inventory control systems or, in some cases, manual calculations of material requirements (Wight 1983: 44) used previously. We do not describe the MRP computations in detail here; a concise description with illustrative examples is given by Hopp and Spearman (2008). More extensive descriptions are presented by Baker (1993) and Vollmann et al. (2005), while Tardif and Spearman

(1997) and Voss and Woodruff (2003) describe MRP in terms of mathematical programming. The original book by Orlicky (1975) remains an interesting and useful reference.

Several characteristics of MRP are of interest to the discussion in this volume:

1. The procedure uses fixed, exogenous planned lead times, usually derived from historical observations of realized cycle times. This facilitates both lead time setting and coordination of the material flow through multiple production stages.
2. MRP is uncapacitated. The capacity requirements over time result directly from the MPS, the lot-sizing rules, the planned lead times, and the current inventory levels of SKUs at the various levels of the BOM. Hence substantial imbalances between required and available capacity can occur.
3. The Master Production Schedule (MPS), which determines the medium-term capacity requirements, is an exogenous input to the system.

MRP does largely what its name implies—material requirements planning with very limited support for capacity planning. The lack of support for Master Production Scheduling is particularly critical; the MPS is treated as a fixed, exogenous input that may be infeasible with respect to capacities. Any infeasibility must first be identified, often a challenging task in itself, and then "repaired" by adjusting either the available or required capacity in the planning periods. The latter can be accomplished by adjusting the MPS, modifying the lot-sizing rules after the CRP process has been completed, or by detailed scheduling at the order or operation level within the production units, which requires a fairly high amount of released work and WIP to be effective. In order to overcome these limitations, a PPC system should support integrated planning of both material and capacities for all resources, as well as the creation of the build schedule (the MPS) that determines the resource requirements.

The serious nature of these limitations raises the question of why a production planning approach with such deficiencies is so widely employed. A confluence of several factors has led to this situation. Firstly, in environments where production capacity is relatively cheap, plentifully available, or both, it is relatively easy to address delays in production plans by adding capacity through overtime, subcontracting, or additional machines. In production environments that maintain relatively constant capacity utilization, cycle times will also remain relatively stable, allowing suitable planned lead times to be learned over time. We have observed several cases of firms deliberately maintaining a constant utilization level, to the extent of temporarily deactivating production equipment in periods of low demand. Another factor in favor of MRP is the transparency of its logic to the end users, in contrast to optimization models that frequently produce solutions that are difficult to explain. Historically, MRP was a tremendous advance over independent demand inventory control systems since it derives the material requirements from the MPS, which is a statement of *future* production as opposed to a forecast. Finally, the wide adoption of MRP in industry has provided an extensive ecosystem of software, consultants, and corporate knowledge supporting its use.

### *3.1.1 Role of the Master Production Schedule*

Master production scheduling has proven more difficult to standardize than the MRP calculation itself. A reference structure for a master production scheduling system derived mainly from empirical observations that covers many practical cases in discrete manufacturing was proposed by Berry et al. (1979) and is described in Vollmann et al. (2005). This MPS reference structure considers the general case where the production and purchasing activities upstream of a specified *customer order decoupling point* (CODP) are based on demand forecasts and the completion of final products downstream of the CODP on customer orders. The CODP is located at the point in the supply chain where material is committed to a particular order and cannot be used to fill any other. A CODP located at the assembly stage results in assemble-to-order (ATO) production, as is common with PCs, and allows short delivery times for complex products, e.g., in mechanical engineering. Make to stock and make to order are special cases with the CODP at the final product or at the raw material level, respectively.

In this MPC structure, the MPS incorporates information from both demand forecasts and confirmed customer orders to provide a time-phased build schedule for all independent demand items, usually based on weekly time buckets. This build schedule provides the input to the MRP system that generates the orders for the production units that are forecast driven (e.g., purchasing and component production). The time-phased capacity requirements that result from an MPS can be calculated by estimating the capacity requirements induced at all relevant workcenters by each unit of the final product to be completed in a given period based on the BOM and process routings for the individual BOM items (Vollmann et al. 2005: 339 ff.). Such Rough-Cut Capacity Planning (RCCP) procedures consider neither existing inventories nor the effects of lot sizing and can thus be quite inaccurate. More advanced methods such as Capacity Requirements Planning (Vollmann et al. 2005) are applied after the MRP computation, considering planned lead times, component inventories, and lot sizes.

In an ATO environment, the forecast-based MPS drives only the production of the components upstream of the CODP. Final Assembly Scheduling is driven by customer orders and controls the production of customer-specific products by the manufacturing stages downstream of the CODP. If any components are not yet available at assembly (due to inaccurate demand forecasts or production issues), exception orders are generated and their production expedited to minimize the delay at assembly. This structure leads to two interrelated control loops within the PPC system that are controlled based on the MPS and the final assembly schedule, respectively, as shown in Fig. 3.1.

This MPC structure is complemented by an aggregate planning level above the two control loops that performs seasonal planning of production and sales quantities, capacities and inventories over time, usually for product families and longer time buckets over a planning horizon of 12–18 months. Since the aggregate "products" representing product families cannot actually be produced, the planned

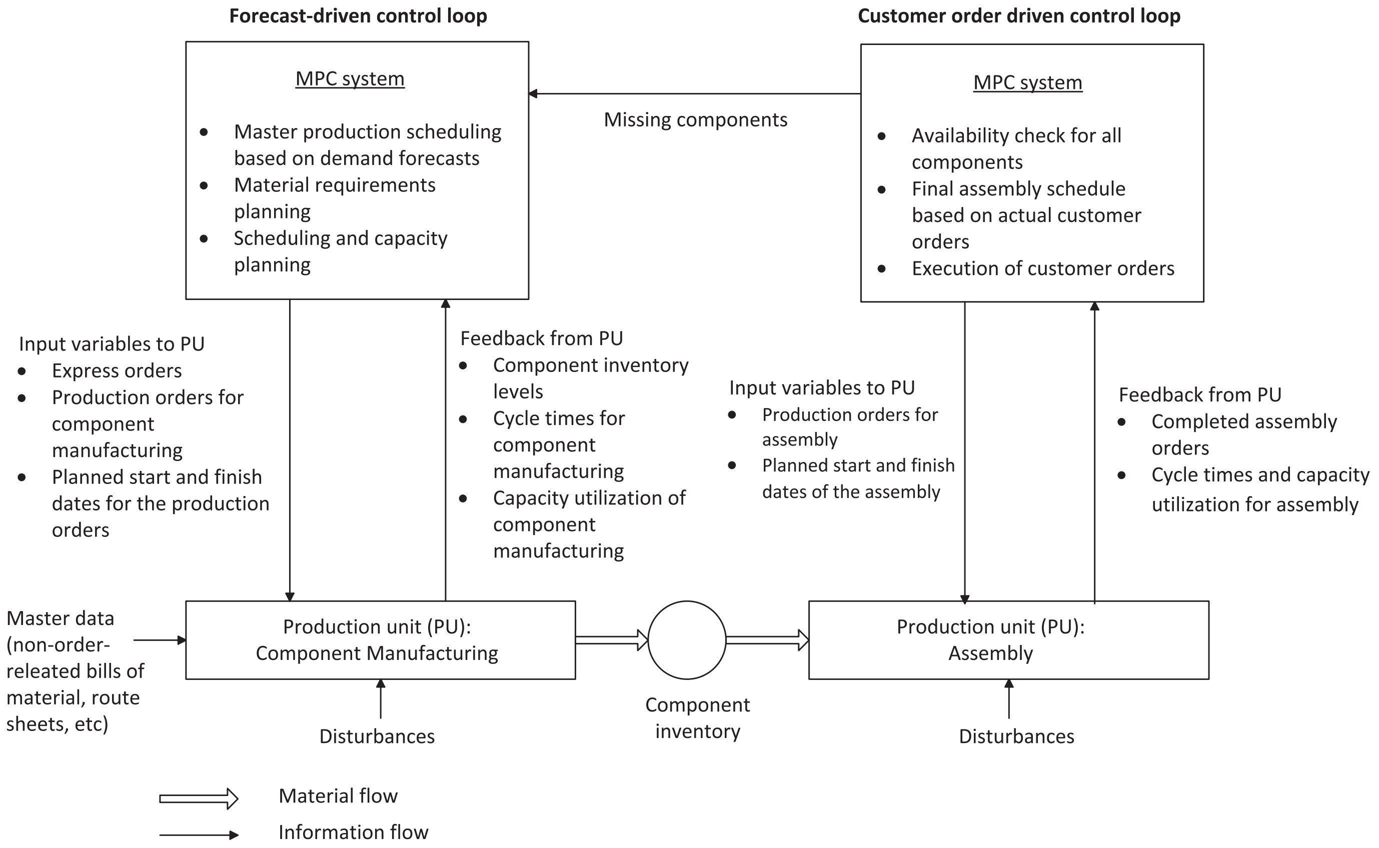

Forecast-driven control loop
Customer order driven control loop
MPC system
• Master production scheduling based on demand forecasts
• Material requirements planning
• Scheduling and capacity planning
Missing components
MPC system
• Availability check for all components
• Final assembly schedule based on actual customer orders
• Execution of customer orders
Input variables to PU
• Express orders
• Production orders for component manufacturing
• Planned start and finish dates for the production orders
Feedback from PU
• Component inventory levels
• Cycle times for component manufacturing
• Capacity utilization of component manufacturing
Input variables to PU
• Production orders for assembly
• Planned start and finish dates of the assembly
Feedback from PU
• Completed assembly orders
• Cycle times and capacity utilization for assembly
Master data (non-order-releated bills of material, route sheets, etc)
Production unit (PU): Component Manufacturing
Component inventory
Production unit (PU): Assembly
Disturbances
Disturbances
Material flow
Information flow

**Fig. 3.1** Dual control loop structure for assemble-to-order production (from Zäpfel 2000: 215, authors' translation)

production quantities at this level represent capacity reservations for each product family that serve to coordinate production and capacity planning. Aggregate production quantities also largely determine possible sales, inventory levels, and cash inflows and outflows, etc., and thus are crucial to coordinating the functional areas of the company. They also serve as a coordination instrument with strategic planning since planned changes in sales quantities in different markets are reflected there. This task, termed *Sales & Operation Planning* (S&OP), links production planning and the larger corporate planning process, forming an important input to the MPS.

This structure leads to a hierarchical PPC system that, at least conceptually, simultaneously considers all resources necessary for production (primarily material and capacities) at each level, which are

- Sales and operations planning—resource planning
- Master production scheduling—rough-cut capacity planning
- MRP—capacity requirements planning and load levelling
- Shop-floor control—detailed scheduling

with increasing levels of detail as one moves down the list. This type of PPC system, often termed MRP II (Wight 1983; Landvater and Gray 1995) or the Manufacturing Planning and Control framework, is depicted in Fig. 3.2.

PPC systems of this type allow seasonal inventories only for MPS items, which are generally final products but may also include important subassemblies. All estimated capacity requirements are derived from the MPS. However, this approach does not allow integrated planning of the material flow across the supply chain when inventory levels at each stage must be considered. If coordination across the supply chain is necessary, the production quantities at each production unit, the transportation quantities between the production units and their inventory levels in each planning period must be defined as separate decision variables whose values determine the MPS, requiring a high level of detail in the MPS. This type of master planning, described in Chap. 1 for semiconductor manufacturing, is a standard function of today's Advanced Planning Systems (APS). Voss and Woodruff (2003) discuss its formulation as a mathematical program and its relationship to MRP II. The resulting planning and control structure is described in Sect. 3.2.

In the MRP II framework, sales and operations planning is performed for aggregate product families, and the MPS is obtained by disaggregating this aggregate production plan. To accomplish this effectively, products in the same family should share similar seasonal demand patterns and resource requirements, even if the strong assumptions of *perfect aggregation* (Axsäter 1981) do not hold. Similarly, master planning for product families requires aggregate bills of material and determination of safety stock levels that allow feasible disaggregation even if the mix of individual product demands within the aggregate demand varies (disaggregation slack). These issues, raised in Bitran et al. (1981), remain critical today.

MRP/MRP II thus has its origins in the material coordination task addressed by the planning level. Several important issues related to this task have attracted extensive research, such as MRP nervousness (Blackburn et al. 1985, 1986; Sahin et al.

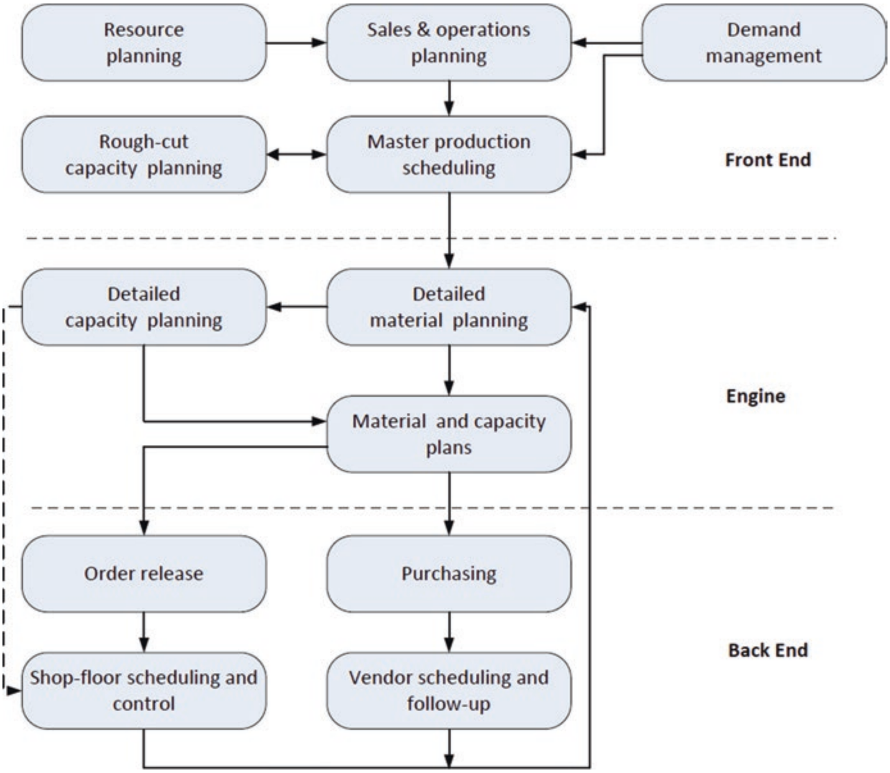

Resource planning
Sales & operations planning
Demand management
Rough-cut capacity planning
Master production scheduling
Front End
Detailed capacity planning
Detailed material planning
Material and capacity plans
Engine
Order release
Purchasing
Back End
Shop-floor scheduling and control
Vendor scheduling and follow-up

**Fig. 3.2** Hierarchical structure of a MRP II/MPC system (Vollmann et al. 2005: 371, modified)

2013; Lin and Uzsoy 2016), multilevel lot sizing (Kimms 1997) and determination of safety stocks and safety times (Meal 1979; Miller 1979; de Bodt and Van Wassenhove 1983; Grubbstrom 1999). As expected, when a complex stochastic production—inventory system operating in a rolling horizon environment is controlled by a simple procedure like MRP, the complexity that is not addressed by MRP emerges elsewhere, and the resulting control system will be as complex as required by the planning problem according to the Law of Requisite Variety (Ashby 1956: 202 ff.). However, our focus in this volume is on the ability of MRP/MRP II to effectively control the system state within the production units in order to manage the cycle times and other performance measures of the manufacturing system.

### 3.1.2 *Lead Time Management in MRP/MRP II Systems*

The observed cycle times of production orders through the production units and the planned lead times estimated from them play a crucial role in the performance of PPC systems, and hence of the production systems they control. As discussed in

Chap. 2, Little's Law (Little 1961; Hopp and Spearman 2008) implies that average cycle times determine the average WIP level at a given throughput rate, while their variability determines how consistently the production system is able to meet the planned lead times, influencing due date performance and safety stock levels. The lead times also constrain the location of the CODP since the total lead time of the make to order portions of the system cannot exceed the customer's requested delivery time. Thus the planned lead times strongly influence essential elements of the MPC problem, making lead time management an important issue.

Planned lead times in MRP/MRP II are treated as *forecast* variables to be estimated from observations of realized cycle times. It is assumed that, as long as some maximum capacity loading is not exceeded, historical cycle times will provide a reasonable estimate of the cycle times of production orders released in the current time frame; the past is representative of the future. This use of planned lead times and maximum capacity loads to coordinate the production planning and detailed scheduling levels for the production unit poses substantial problems. First of all, it requires accurate time-phased load projections and sufficient planning capability to avoid the unduly long cycle times that arise when resources are temporarily overloaded. Capacity planning methods are provided in MRP II both at the MPS level (RCCP) and after the MRP run (CRP). However, since RCCP can only approximate the time-phased capacity requirements with no information on lot sizes or component inventories, and load leveling after the MRP run is based on predetermined lot sizes and lead times (and is a very complex task in its own right), the result can be far from optimal. Integrating MRP and capacity planning by solving multilevel capacitated lot-sizing models remains challenging for practical applications despite substantial progress in recent years (Tempelmeier and Buschkühl 2009; Helber and Sahling 2010). Thus there is always a substantial possibility that capacities are overloaded in certain periods or that overloading is avoided by suboptimal measures.

If realized cycle times deviate from the planned lead times, the latter are often updated to maintain high due date performance, and the release schedule is adapted accordingly. As discussed in detail in Chap. 2, however, the workload in the production unit—controlled by the order release function—determines the cycle times. This inconsistency—treating a control variable as a forecast variable—can lead to a vicious cycle called the *lead time syndrome* illustrated in Fig. 3.3: planners respond to long and unreliable cycle times by specifying longer planned lead times, causing orders to be released earlier in order to meet their required due dates. This increases the number of orders in the production unit (i.e., the WIP level), leading to longer queues at the workcenters, which, in turn, increases the average cycle time. Planners often react by increasing the planned lead times still further, causing the next batch of orders to be released even earlier. This effect is often further exacerbated in practice by uncontrolled releases of urgent orders (usually for missing parts that are delaying assembly of an order).

The lead time syndrome was first described in the 1970s (Wight 1974; Mather and Plossl 1978). Although rigorous studies are quite recent (Selcuk et al. 2006, 2009), anecdotal evidence suggests that it can inflate planned lead times beyond any defensible level (Wight 1974: 108 ff.). Whether the lead time syndrome is reversible

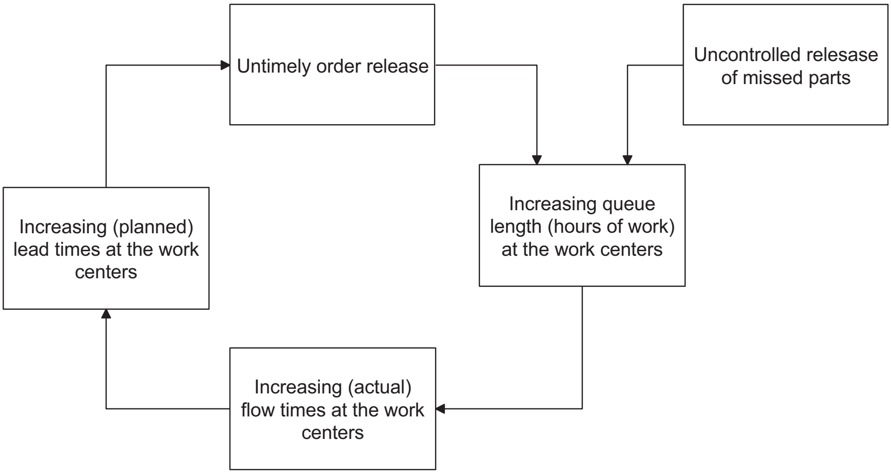

Untimely order release
Uncontrolled relesase of missed parts
Increasing (planned) lead times at the work centers
Increasing queue length (hours of work) at the work centers
Increasing (actual) flow times at the work centers

**Fig. 3.3** Lead time syndrome

and the circumstances under which the system can become “locked” in a long lead time regime are still not well understood. Selcuk et al. (2009) show that the variability of planned lead times increases with their update frequency, suggesting a trade-off between lead time accuracy and system stability when lead times are treated as forecast variables.

Overcoming the lead time syndrome requires a fundamental change of perspective: instead of treating lead times as an exogenous parameter to be forecast, they should be treated as a *control* variable whose value can be influenced by order release and capacity decisions. This requires replacing the forecasting task of MRP/MRP II by an anticipation task—that of understanding the relationship between order release and capacity adjustment decisions and the cycle times that will be realized when these decisions are implemented. This view of lead times as endogenous to the planning process lies at the heart of this volume and will be discussed in more detail in later chapters.

## 3.2 Hierarchical Production Planning (HPP) and Advanced Planning Systems (APS)

Developments in information technologies over the second half of the twentieth century, most notably the development of ever more powerful computers, relational database systems capable of organizing the massive amounts of data involved, and the evolution of client-server computing, brought the possibility of Hierarchical Production Planning (HPP) systems where material flows and capacities are planned simultaneously at multiple time frames from medium-term aggregate planning to very short-term dispatching. Conceptually, this is a vertical decomposition of the overall PPC problem into a series of (hopefully!) tractable planning subproblems that avoids the well-known problems of solving and implementing a single monolithic model of the overall production planning problem as a single planning task. The advantages of hierarchical planning in companies are obvious, and the observation that hierarchical planning systems fit the organizational structure better than monolithic models may well be due to the fact that the organizational structure is an adaptation to the same factors that make hierarchical planning systems desirable. Thus ideas for Hierarchical Production Planning (HPP) systems were expressed very early in the literature on production planning and management (Holt et al. 1960; Anthony 1966).

Mathematical models have been developed to support a range of planning tasks within this hierarchy. However, due to the complexity of the planning problem, especially in multistage production systems with complex BOM structures, deriving this decision hierarchy and the respective planning models by mathematical decomposition of a monolithic model has not proved possible, although it remains an interesting theoretical goal.

For simpler production planning problems a theoretically sound hierarchical production planning system should be within reach, and a body of research addressing this problem has emerged alongside the MRP approach. We now describe the essence of this work on Hierarchical Production Planning, using this term not in the general sense that each PPC system exhibits a hierarchical structure (although this is usually the case), but to refer to specific PPC systems within this research tradition, although the boundary is often ambiguous. We then describe the structure of Advanced Planning Systems (APS) based on this hierarchical concept and have a different focus compared to the MRP/MRP II framework.

### *3.2.1 Hierarchical Production Planning*

The seminal paper in this research tradition is that of Hax and Meal (1975), who model a tire manufacturer as a single-stage production system. The number of products is high, and the planning horizon must cover at least one entire seasonal cycle due to substantial demand seasonality. A centralized PPC approach must determine the production, sales, and inventory quantities of each product in each period of the planning horizon using a single monolithic model. This requires medium-term demand forecasts for each product and period, including forecast updating before each planning cycle, and makes medium-term decisions (e.g., how to handle seasonal demand) and short-term production decisions (production quantities for the next production run) simultaneously. Such an approach, although feasible from a modeling and algorithmic perspective, is very likely to fail; Meal (1984) describes the failure of such a centralized approach. The hierarchical approach provides a way out of this dilemma. Products that share setups constitute natural product families with negligible setup times between products of the same family and hence can be aggregated. Product families with similar seasonal demand patterns, capacity requirements, revenues, and unit costs (or inventory investment produced per unit time; see Graves (1982)) can be further aggregated into product types. This three-level structure is specific to the particular case of the tire manufacturer, but has proven to be viable in many batch manufacturing environments (Hax 2013: 709).

Once this aggregation hierarchy is identified, planning tasks can be assigned to the aggregation levels as follows:

*Seasonal planning* can be performed at the product type level since this level determines capacities and their usage or reservation, and the parameters that determine the seasonal plan are similar for products of the same type. The decision model is usually a linear program.

*Lot sizes* are determined at the level of product families since setups only occur with a change of product family. The decision model is specific to the case of the tire manufacturer and is solved by a heuristic.

*Production quantities* for individual products within the product families are determined in the short term to approximately equalize the projected run-out times of the products, when inventory will be exhausted and must be replenished by another production run. Since all costs are determined at the product type and product family levels, this allows products of the same family to share a family setup.

Only the seasonal planning performed at the product type level considers multiple planning periods. Product family and item-level planning are only performed for the first planning period, and the entire process is repeated at the start of the next planning period.

The key issue in HPP is that of aggregation, primarily of products in this case, but also of capacities (machines to workcenters to production units) and time. The higher level decisions constrain the lower level ones; only if these constraints are satisfied are the decisions at the different levels consistent. The ability to aggregate products depends on the specific situation, although common structures such as aggregate products that allow capacity-oriented seasonal production planning can be identified in many cases.

The vertical decomposition and strict top-down approach of the Hax/Meal approach impose some important limitations. Although the planning models are specified at all levels and the production quantities of product types, product families, and individual products are consistent, overall optimality is not guaranteed, for two primary reasons:

1. The production plan obtained from the optimal aggregate plan is only equal to the optimal production plan obtained from a model formulated at the item level under the strong assumptions of perfect aggregation (Axsäter 1981, 1986). In practice the data of individual products differ to some extent, making only approximate aggregation possible.
2. The decision models at higher levels often cannot accurately anticipate the impact of their decisions on the costs of the base-level decisions. For instance, in the Hax/Meal framework, the seasonal planning carried out at the product-type level does not accurately represent the impact of its decisions on the total setup

> costs determined by the product family subproblem at the next lower level (Graves 1982: 263 ff.). This information can only be obtained by feedback from the product family level. Graves (1982) extends the Hax/Meal approach with a feedback mechanism based on Lagrangian techniques that modifies the holding cost coefficients used in the product type problem, dividing the holding costs between the product type and product family subproblems (Graves 1982: 265).

The Hax/Meal case study considers only one production stage. Extending the approach to a two-stage system as in Bitran et al. (1982) raises additional issues. Product aggregation now requires aggregation of multistage material flows, requiring the definition of aggregate bills of material (Axsäter 1986). Secondly, minimum inventory levels must be defined for aggregate planning in order to guarantee SKU availability at the item level. Determining these minimum inventory levels is a complex research topic in its own right (Axsäter 1986; Lasserre and Mercé 1990; Gfrerer and Zäpfel 1995).

In the 1980s and 1990s, the HPP research tradition was largely pursued through case studies, with some conceptual work (Bitran and Tirupati 1993). McKay et al. (1995) present a review and critique of the approach, while Leachman (1993, 2001) presents an extensive case study in the semiconductor industry. Conceptual issues are discussed in Schneeweiss (2003).

Since HPP emphasizes the capacity aspects of the PPC problem that are the principal weak point of MRP, whose focus is material planning, integrating the two frameworks seems reasonable. Meal et al. (1987) attempt this integration for a manufacturer of computer peripherals, noting that HPP encompasses the allocation of production among plants that is not considered in MRP. At the plant level, although "both MRP and HPP deal with capacity and material plans" (p. 952), HPP tends to focus on the capacity side of the MPC hierarchy (Fig. 3.2) "communicating the constraints from the front end to the engine to the back end," whereas MRP focuses on the material side communicating the material requirements from production planning to Master Production Schedule to detailed material requirements. The distinction between "capacity oriented" and "product oriented" planning approaches (Bertrand et al. 1990: 57 ff.) expresses this difference. Hence capacity requirements can be derived from MRP, while estimates of available capacity can come from HPP (Meal et al. 1987: 953). MRP determines material and capacity requirements, while HPP "starts with capacity available and schedules the jobs to fill the capacity" (p. 954).

This capacity-oriented view of HPP raises the question of how much the maximum possible output the system can produce is affected by the aggregate capacity loading. High capacity loading may allow more effective optimization of lot sizes than is possible when there is less work available to the resources. A large amount of work available to a machine reduces the probability of its idling due to lack of material. The clearing function models discussed in Chaps. 7 and 8 formulate several different models of this relationship between workload and output. We now discuss the Advanced Planning Systems framework that has its roots in the HPP research we have just briefly reviewed.

### *3.2.2 Advanced Planning Systems (APS)*

Today's Advanced Planning Systems (APS) (Stadtler et al. 2015) seek to implement essential PPC functions, emphasizing planning and coordination of the material flow between companies or manufacturing plants using the data collection and organization capabilities of the Enterprise Resource Planning (ERP) and Manufacturing Execution Systems (MES) used by many companies today. The Supply Chain Planning Matrix (Fleischmann et al. 2015), shown in Fig. 3.4, provides a basic framework for the development of these systems. In the figure, which is modified somewhat from the original to avoid additional terminology, each planning function, represented by a rectangle, produces decisions that may form inputs for other planning functions. The horizontal axis represents material flow across business functions (procurement, production, distribution, and sales), and the vertical axis the time frame associated with those decisions (long-, mid-, and short-term).

Strategic Network Design is an ongoing long-term process across all business functions, determining the products to be produced, the markets to be served, and the locations and sizes of the facilities to produce and distribute them. As in the MPC framework in Sect. 3.1, Demand Management involves developing demand forecasts at different levels of aggregation: long-term aggregate forecasts at the level of product families, large time buckets and regional geographies required for Strategic Network Planning, and the disaggregated, shorter-term forecasts used for Master Planning. Master Planning takes as inputs the long-term Strategic Network Design decisions and determines a time-phased plan specifying how much of each

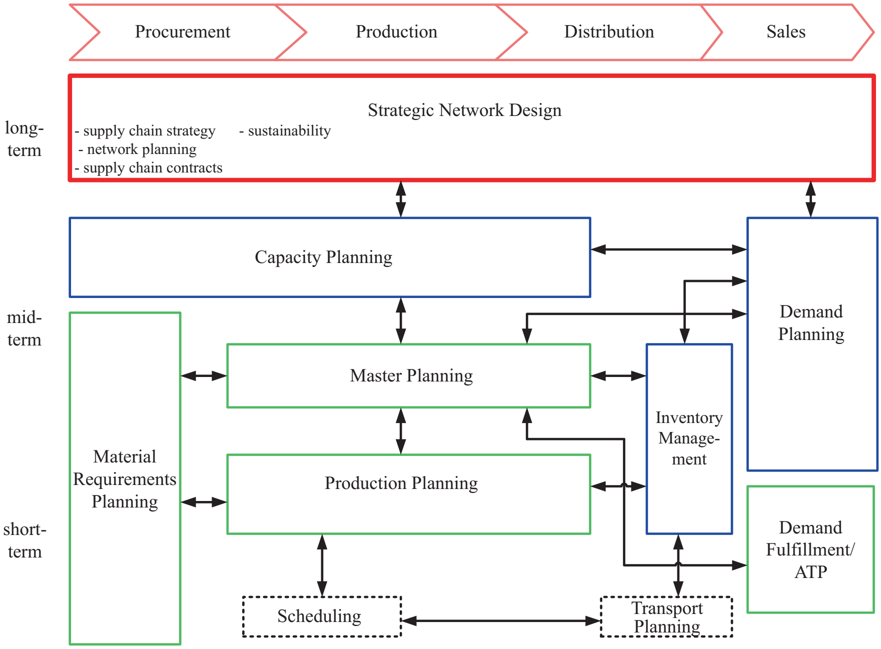


**Fig. 3.4** Supply chain planning matrix (Fleischmann et al. 2015, modified by Moench et al. 2017)

product or product family will be produced in what facilities in order to coordinate material flow through the supply chain. Since the management of seasonal demand fluctuations by building inventories ahead of demand peaks, outsourcing, or delaying demand is an important consideration, the time frame for Master Planning must consider an entire seasonal cycle. The level of aggregation in the Master Planning activity can vary; it is usually focused on potentially constraining resources and product families, but can also be performed at the level of individual products. Note that the Master Plan of the Supply Chain Planning Matrix is not necessarily the same thing as the Master Production Schedule of the MPC framework; the Master Plan is not necessarily computed at the level of specific items and usually considers multiple production units and capacity constraints at potentially limiting workcenters. The Master Production Schedule, on the other hand, does not consider the bill of material explosion necessary to synchronize material flow across multiple production units; in the MPC framework, this is performed by the MRP logic.

After Master Planning is complete, Production Planning seeks a capacity-feasible release plan that will allow each facility in the supply chain to meet the production targets set for it by Master Planning. Again, the Supply Chain Planning Matrix uses the term “Production Planning” in a different meaning than that in the MPC framework (and this volume); in the latter it encompasses all planning activities leading up to the computation of the order releases, while under the Supply Chain Planning Matrix, it is limited to computing capacity-feasible order releases that will meet the production goals set by Master Planning for the individual production units. Once work is released into a production unit, its progress towards completion is controlled by that unit’s internal scheduling function.

The structures of the mathematical models for Master Planning and Aggregate Production Planning are quite similar; in fact, the term “Sales and Operations Planning” is used in both the frameworks (Vollmann et al. 2005, Chap. 3 and Stadtler et al. 2015: 173 f.). The principal decision variables are either releases or production quantities of each product (or product family) in each period in the planning horizon at each facility considered; we show in Chap. 5 that under the assumption of fixed, workload-independent lead times, these two quantities are equivalent. The models must include material balance constraints for all inventory locations considered, capacity constraints for critical resources, and domain-specific constraints representing technological and business policy constraints specific to the application of interest. Models for Aggregate Production Planning or Master Production Scheduling are usually formulated for one level in the product structure, mostly final products or—more generally—MPS items, whereas Master Planning explicitly models flows and inventories for all facilities considered at the specified level of detail. As the level of detail in Master Planning models is increased to model the process more precisely, at least some portions of a Master Planning model can easily acquire the level of detail usually associated with the Production Planning function of APS. Hence the authors of both MPC and APS frameworks emphasize that they need to be adapted to different situations. The primary function of the combined problems is to coordinate the flow of material through the supply chain to best meet the firm’s objectives.

Both PPC frameworks described here—the MPC framework based on MRP/MRP II and APS—eventually yield production orders for the production units: from MRP and lot sizing in the MPC framework, or from the master planning and production planning functions under APS. Production orders can also be generated from independent demand inventory control systems (e.g., for spare parts), and in MTO companies, production orders can result directly from customer orders. A (hopefully small) fraction of the production orders might be unplanned, resulting, e.g., from specific material requirements of customer-specific product variants in assemble-to-order production as described in Sect. 3.1. All these orders must be released to the production units in a way that guarantees that the planned due dates are satisfied, which requires keeping the cycle times under control.

Mechanisms for managing cycle times within PPC systems fall into two basic camps: those that treat cycle time as an exogenous variable to be forecast and those that view it as a variable to be controlled (Tatsiopoulos and Kingsman 1983). The former contradicts the queueing perspective developed in Chap. 2, which makes it quite clear that the average cycle time $T$ is determined by the planning level's release decisions through their effect on resource utilization and the variability of material flows. The other camp, motivated by Little's Law discussed in Sect. 2.2, attempts to maintain stable mean cycle times $T$ by regulating the short-term release of work into production units over time to maintain a constant workload $W$. We now turn to a discussion of these latter approaches.

## References

Anthony RN (1966) Planning and control systems: a framework for analysis. Harvard University Press, Cambridge

Ashby WR (1956) An introduction to cybernetics. Chapman and Hall, London

Axsäter S (1981) Aggregation of product data for hierarchical production planning. Oper Res 29(4):744–756

Axsäter S (1986) On the feasibility of aggregate production plans. Oper Res 34(5):796–800

Baker KR (1993) Requirements planning. In: Graves SC, Rinnooy Kan AHG, Zipkin PH (eds) Handbooks in operations research and management science, Logistics of production and inventory, vol 3. Elsevier Science, Amsterdam, pp 571–627

Berry W, Vollmann T, Whybark D (1979) Master production scheduling: principles and practice. American Production and Inventory Control Society, Chicago

Bertrand JWM, Wortmann JC, Wijngaard J (1990). Production Control: A Structural and Design Oriented Approach. Amsterdam, Elsevier

Bitran GR, Tirupati D (1993) Hierarchical production planning. In: Graves SC, Rinnooy Kan AHG, Zipkin P (eds) Logistics of production and inventory, vol 4. Elsevier, Amsterdam, pp 523–568

Bitran GR, Haas EA, Hax AC (1981) Hierarchical production planning: a single stage system. Oper Res 29(4):717–743

Bitran GR, Haas EA, Hax AC (1982) Hierarchical production planning: a two-stage system. Oper Res 30(2):232–251

Blackburn JD, Kropp DH, Millen RA (1985) MRP system nervousness: causes and cures. Eng Costs Prod Econ 9(1–3):141–146

Blackburn JD, Kropp DH, Millen RA (1986) A comparison of strategies to dampen nervousness in MRP systems. Manag Sci 32(4):412–439

Chandler AD (1962) Strategy and structure: chapters in the history of the American industrial enterprise. MIT Press, Cambridge

Chandler AD (1980) The visible hand: the managerial revolution in American business. Belknap Press, Cambridge

de Bodt MA, Van Wassenhove LN (1983) Lot sizes and safety stocks in MRP: a case study. Prod Inventory Manag (First Quarter):1–15

Fleischmann B, Meyr H, Wagner M (2015) Advanced planning. In: Stadtler H, Kilger C, Meyr H (eds) Supply chain management and advanced planning. Springer, Berlin, pp 71–95

Gfrerer H, Zäpfel G (1995) Hierarchical model for production planning in the case of uncertain demand. Eur J Oper Res 86:142–161

Graves SC (1982) Using Lagrangian techniques to solve hierarchical production planning problems. Manag Sci 28(3):260–275

Grubbstrom RW (1999) A net present value approach to safety stocks in a multi-level MRP system. Int J Prod Econ 59:361–375

Hax AC (2013) Hierarchical production planning. In: Gass SI, Fu MC (eds) Encyclopedia of operations research and management science. Springer, Boston

Hax AC, Meal HC (1975) Hierarchical integration of production planning and scheduling. In: Geisler MA (ed) Logistics. MIT Press, Amsterdam, pp 53–69

Helber S, Sahling F (2010) A fix-and-optimize approach for the multi-level capacitated lot sizing problem. Int J Prod Econ 123(2):247–256

Holt CC, Modigliani F, Muth JF, Simon HA (1960) Planning production, inventories and work force. Prentice Hall, Englewood Cliffs

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Kimms A (1997) Multi-level lot sizing and scheduling: methods for capacitated, dynamic, and deterministic models. Physica-Verlag, Heidelberg

Landvater DV, Gray CD (1995) MRP II standard system: a handbook for manufacturing software survival. Wiley, New York

Lasserre JB, Mercé C (1990) Robust hierarchical production planning under uncertainty. Ann Oper Res 26(1):73–87

Leachman RC (1993) Modeling techniques for automated production planning in the semiconductor industry. In: Cinani TA, Leachman RC (eds) Optimization in industry. Wiley, Chichester, pp 1–30

Leachman RC (2001) Semiconductor production planning. In: Pardalos PM, Resende MGC (eds) Handbook of applied optimization. Oxford University Press, New York, pp 746–762

Lin PC, Uzsoy R (2016) Chance constrained formulations in rolling horizon production planning: an experimental study. Int J Prod Res 54(13):3927–3942

Little JDC (1961) A proof of the queueing formula $L = \lambda W$. Oper Res 9:383–387

Mather H, Plossl GW (1978) Priority fixation versus throughput planning. Prod Invent Manag J 19(3):27–51

McKay KN (2010) The historical foundations of manufacturing planning and control practices. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise, vol 12. Springer, New York, pp 21–32

McKay KN, Safayeni FR, Buzacott JA (1995) A review of hierarchical production planning and its applicability for modern manufacturing. Prod Plan Control 6(5):384–394

Meal H (1979) Safety stocks in MRP systems. Operations Research Center, Massachusetts Institute of Technology, Cambridge

Meal HC (1984) Putting production decisions where they belong. Harv Bus Rev 62:102–111

Meal HC, Wachter MH, Whybark DC (1987) Material requirements planning in hierarchical production planning systems. Int J Prod Res 25(7):947–956

Miller JG (1979) Hedging the master schedule. In: Ritzman LP (ed) Disaggregation problems in manufacturing and service organizations. Martinus Nijhoff, Boston

Moench L, Uzsoy R, Fowler J (2017) A survey of semiconductor supply chain models part I: semiconductor supply chains and strategic network design. Int J Prod Res 56(13):4524–4545

Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York

Sahin F, Narayanan A, Robinson EP (2013) Rolling horizon planning in supply chains: review, implications and directions for future research. Int J Prod Res 51(18):5413–5436

Schneeweiss C (2003) Distributed decision making. Springer, Berlin

Selcuk B, Fransoo JC, De Kok AG (2006) The effect of updating lead times on the performance of hierarchical planning systems. Int J Prod Econ 104(2):427–440

Selcuk B, Adan I, de Kok AG, Fransoo JC (2009) An explicit analysis of the lead time syndrome: stability condition and performance evaluation. Int J Prod Res 47(9):2507–2529

Stadtler H, Kilger C, Meyr H (2015) Supply chain management and advanced planning. Concepts, models, software, and case studies. Springer, Berlin

Tardif V, Spearman ML (1997) Diagnostic scheduling in finite-capacity production environments. Comput Ind Eng 32:867–878

Tatsiopoulos IP, Kingsman BP (1983) Lead time management. Eur J Oper Res 14:351–358

Tempelmeier H, Buschkühl L (2009) A heuristic for the dynamic multi-level capacitated lotsizing problem with linked lotsizes for general product structures. OR Spectr 31(2):385–404

Vollmann TE, Berry WL, Whybark DC, Jacobs FR (2005) Manufacturing planning and control for supply chain management. McGraw-Hill, New York

Voss S, Woodruff DL (2003) Introduction to computational optimization models for production planning in a supply chain. Springer, Berlin

Wight O (1974) Production and inventory management in the computer age. Wiley, Boston

Wight O (1983) MRPII: unlocking America's productivity potential. Oliver Wight Limited Publications, Williston

Zäpfel G (2000) Strategisches Produktionsmanagement. Oldenbourg, Munich

# Chapter 4
# The Workload Control Approach

The relation between workload and cycle time revealed by queueing models suggests that regulating the workload of the production unit can maintain its average cycle time at a desired level while still producing the desired level of output. Since workload is determined by the order release process linking the planning and scheduling levels, a wide range of order release procedures based on maintaining a stable workload level in the production units have been suggested. In this chapter, we discuss the extensive body of work that approaches this problem without explicit models of material flow over time. We then discuss the advantages and shortcomings of this family of approaches, setting the stage for the model-driven release planning approaches presented in the subsequent chapters.

## 4.1 Basic Concepts

MRP/MRP II approaches lead time management by focusing on the left side of the vicious cycle in Fig. 3.3, treating lead times as *forecast* variables to be estimated from historical observations instead of endogenous variables determined by the state of the production system. The computation of coordinated production plans for manufacturing stages or plants by master planning, which is a standard function of today's APS systems and widely applied, e.g., in semiconductor manufacturing (see Sect. 1.2.1), often makes the same assumptions, although their sophisticated load leveling procedures usually mitigate the negative consequences of this assumption to some degree. Viewed from the queueing perspective developed in Chap. 2, the mechanism that produces the lead times—the nonlinear relation between workload and cycle time—is not captured at the planning level in either of these PPC frameworks.

The *Workload Control* (WLC) approach is motivated by the insight that, at least if the system is approximately in steady state, its mean cycle times and throughput are determined by the average WIP level. Thus the behavior of any production unit

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_4

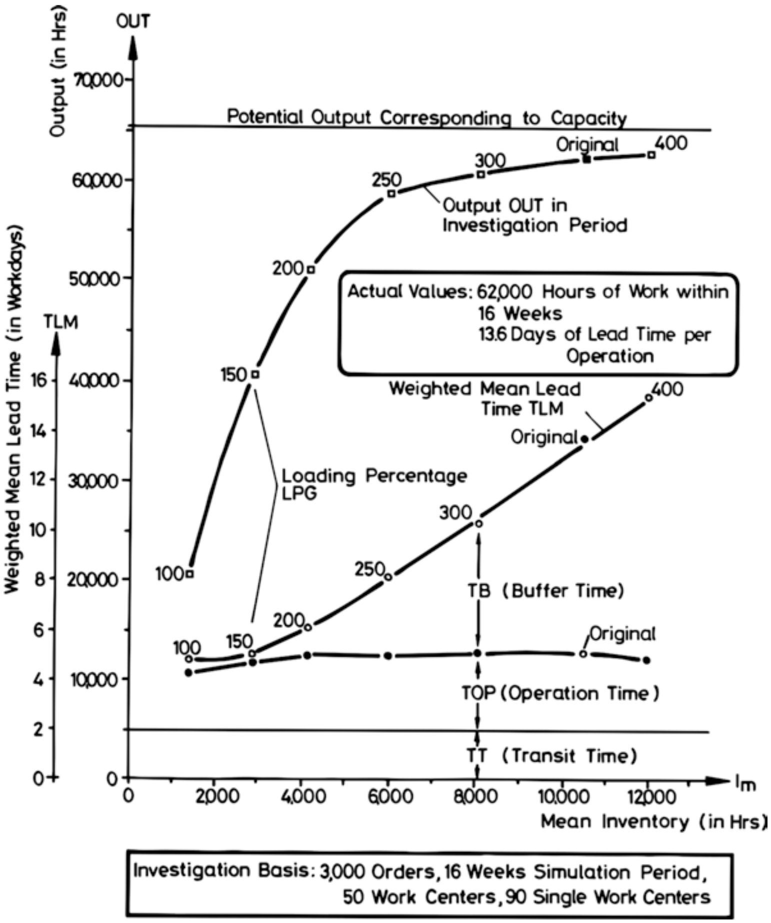


Fig. 4.1 Curves for output and mean flow time per operation with change in the work in process (simulation result; source: Wiendahl 1995: 246)

can be described, at least in principle, by the operating or characteristic curves (Aurand and Miller 1997) linking these variables shown in Fig. 2.1 and, in more detail, in Fig. 4.1. Cycle times, and hence the planned lead times derived from them, are treated as control variables, whose target values can be related to a target WIP level yielding an acceptable compromise between the goals of maintaining low WIP and short cycle times on the one hand and high capacity utilization and output on the other. Hence in the short term when capacities are fixed, order release becomes the principal decision function of the WLC approach.

Under the WLC framework, all production orders arriving at a production unit, whether they originate as external customer orders or production orders generated internally by the planning level, are held in a pre-shop pool until their release to the production unit. The time a particular order is released depends on its due date and some measure of the WIP level in the production unit. An order can be released when the WIP measure falls below a specified threshold, which is thus an important parameter of any WLC approach.

This idea has important implications for the architecture of the broader PPC system. The characteristic curves depicted in Fig. 4.1 depend not only on the technical properties of products and the production system but also on decision logic such as sequencing rules, lot sizes, capacity flexibility and order release frequency. Workload Control assumes that these decision procedures are either fixed over time or functionally related to the average WIP level. Collectively, they determine the stochastic processes describing job arrivals at and departures from the workcenters, and thus the operational characteristics of the production units, as suggested in Chap. 2. Hence, a PPC system using the WLC approach at some planning level must define consistent target WIP values across different production units and requires decision procedures both at the planning level (e.g., lot sizing) and within the production units themselves (e.g., dispatching) to be consistent with these targets (Bertrand and Wortmann 1981; de Kok and Fransoo 2003). Adoption of the WLC framework thus implies an architecture for the entire PPC system. Important conceptual discussions of the WLC approach, which was developed in the 1980s and early 1990s, are given by Bertrand et al. (1990), Zäpfel and Missbauer (1993b), and Missbauer (1998). A PPC system following this architecture must release production orders to the production units to maintain desired (not necessarily constant) WIP levels while coordinating the material flow across the production units to the customers. Implementing WLC thus requires the solution of two subproblems: (1) determining the target WIP levels at all workcenters along the process routing and (2) determining release times for specific orders, i.e., how work will be released into the shop over time to maintain these target WIP levels. We first describe the general logic of the former task and then possible approaches to the latter.

### *4.1.1 Determining Target WIP Levels Under WLC*

The target WIP level in a WLC-based order release procedure must represent an acceptable compromise between the goals of maintaining short, stable cycle times (which requires low WIP levels) on the one hand and high output (which requires low probability of bottleneck starvation, and hence high WIP levels) on the other. The characteristic curves show which combinations of these performance measures are mutually consistent. A desirable target WIP level in Fig. 4.1 would be around 6000 h; higher WIP yields very little additional output, while lower WIP leads to a sharp decrease in output, and hence serious loss of revenue.

Unless extensive simulations are performed to map them (Yang et al. 2006; Ankenman et al. 2010), the characteristic curves are not known in practice since available empirical data represent only the limited range of operating conditions, especially WIP levels, under which the production unit has historically operated. Thus an initial target WIP level is frequently obtained by applying Little's Law to the target average cycle time and the desired output level that, in turn, is derived from the workcenter capacity. The target cycle time should be as short as possible while allowing the production unit to reliably maintain the desired output level. Note, however, that the target WIP and output levels determine only the *average* cycle time; individual order or operation cycle times are random variables whose distribution is affected by dispatching and scheduling decisions within the production unit, as seen in Fig. 2.3. Thus safety stocks or safety lead times obtained by appropriate due date setting (Land 2004: 13 ff.) will be necessary, although the controlled WIP level should result in reduced cycle time variance, and hence lower safety stock levels.

Several additional factors can complicate the situation. The characteristic curves assume that specifying the average WIP level is sufficient to determine the long-term average output; effectively, the average WIP level is used as the argument of a deterministic function that estimates the corresponding average output, assuming that the production system remains unchanged over time. This implies that reducing the average WIP level below the value suggested by the characteristic curves will lead to a sharp decrease in utilization and output. However, in practice the WIP reduction may eventually lead to improvements in the production unit that allow high output to be maintained at a lower WIP level than that suggested by the characteristic curve. This is a common argument in the literature on Just in Time or lean production (Krajewski et al. 2013): inventories, including WIP, hide inefficiencies. Reducing WIP reveals imbalanced capacities, process uncertainty, etc., as targets for improvement efforts, eventually allowing high levels of output to be achieved at a lower WIP level. This process of *continuous improvement*, generally accomplished by eliminating sources of variability, is widely used in industry under a variety of labels such as lean manufacturing (Womack et al. 1990), the Toyota Production System (Liker 2004), Six Sigma (Pande et al. 2000), and Theory of Constraints (Goldratt and Fox 1986), among others. These improvements will change the shape of the output (and the related cycle time) functions in Fig. 4.1 towards a lower WIP level for given output. Hence the characteristic curves obtained by simulation always reflect the structure and operating rules of the production unit assumed in the simulation model, and it is difficult to determine with certainty the minimum WIP level required to maintain high utilization given the possibility of improvements that can be realized without substantial investment. Our discussion of WLC in this chapter will thus assume, with most of the literature on this subject, that all target values and decision rules are part of the definition of the system under study, and hence not subject to change over the time frame of the study.

### 4.1.2 Order Release Mechanisms Under WLC

The problem of determining how to release work over time to maintain a predetermined target WIP level has been studied since the 1970s. Two main approaches can be distinguished:

(a) *Rule-based* order release mechanisms that select orders to release over a short planning horizon, typically one planning period (e.g., 1 day or 1 week), usually without explicit reference to an objective function. These constitute the majority of WLC research since the early 1980s (Land 2004; Stevenson and Hendry 2006).

(b) *Model-based* order release mechanisms that determine the release quantities or the orders to release over a multiple period planning horizon based on an explicit model of the material flow and the resulting WIP levels over this planning horizon. If, as is common, the model only specifies the release quantities, additional logic may be needed to define the orders to release in a manner consistent with these release quantities, especially when lot sizing issues are significant.

These approaches differ substantially with respect to their integration into the overall PPC system. The rule-based approach seeks to maintain a specified WIP level in the short term, so it must be supported by a medium-term planning system that balances load and capacity over a time horizon beyond the current period. Model-based order release, on the other hand, simultaneously determines order releases and material flow over a specified planning horizon, performing both medium-term planning (e.g., load leveling for time-varying demand) and release planning in the same model. Order release planning can be performed at the level of individual products, but also in aggregate terms, based on product families or hours of work moving through the system along different production routings. In the latter case, short-term order release must determine which individual orders to release within these targets, which is usually a straightforward task.

We now discuss rule-based order release mechanisms for WLC in more detail, since the limitations of these approaches motivate the optimization models that are the main topic of this volume.

## 4.2 Rule-Based WLC Approaches

### *4.2.1 Overview*

The most common approach to designing rule-based WLC algorithms assumes that the production orders, their required due dates, and the available capacities are balanced in the medium term by some medium-term planning function such as capacitated Master Planning (Stadtler et al. 2015), or an MRP II procedure with Rough-Cut Capacity Planning (RCCP) or Capacity Requirements Planning (CRP) capability as discussed in Chap. 3. The order release function seeks to manage short-term order releases to maintain the target WIP level over the current planning period, typically 1 day to 1 week. Since no order release plan is made for future periods, there is no reference to an objective function such as minimizing costs over a certain planning horizon. Therefore we refer to these algorithms collectively as *rule-based WLC* although in a few cases optimization models play a limited role (Irastorza and Deane 1974; Yan et al. 2016).

Several rule-based order release mechanisms of this type were proposed in the 1980s and 1990s. These include CONWIP (Spearman et al. 1990), Load-Oriented Order Release (Wiendahl 1995), the LUMS approach (Hendry and Kingsman 1991a, b), and the method of Bertrand and Wortmann (1981). Bergamaschi et al. (1997), Land (2004), Fredendall et al. (2010), and Thürer et al. (2011) review developments in this very active research area over the last 30 years. Kanban (Sugimori et al. 1977; Liker 2004) can also be considered a Workload Control technique, but is based on more restrictive assumptions. Drum-Buffer-Rope (Goldratt and Fox 1986; Cohen 1988; Gupta 2005) is based on a detailed schedule of the bottleneck (the "drum beat") and thus differs from the hierarchical MPC concept underlying this chapter (Zäpfel and Missbauer 1993a), although the determination of the constraint buffer size and the control of order releases into the constraint buffer via the "rope" are related to WLC ideas.

The input to most rule-based order release mechanisms is a list of production orders, each specifying a product or component type, quantity, and required due date, generated by the production planning system applied in the company. These orders are initially held in a *pre-shop pool* of unreleased orders, which may be a physical area where raw materials are staged and documentation prepared, or simply a list of unreleased orders maintained by the firm's information systems. A planned start date derived from the required due date and a planned lead time consistent with the target WIP level used in the WLC logic are usually available for each order. Orders are released from the pool based on their planned start dates and the load situation in the shop. Upon order release, control over the order is transferred to the management of the production unit whose task is to meet the required due date. The goal is to complete all orders on time while maintaining the target WIP levels.

In its most common version (Land 2004), the procedure is invoked periodically (e.g., daily or weekly) to select a subset of the candidate orders in the order pool for release in the current period. The orders in the pool are first sequenced in some priority order, usually by planned start date (due date minus planned lead time). A *time limit* specifying how far in advance of its planned start date an order can be released may also be considered. All orders with planned start dates within the time limit are considered for release in increasing order of their planned start date and released if their release does not violate the target WIP level at any workcenter along its routing. Unreleased orders are reconsidered in the next period.

Order release mechanisms of this type perform three functions: load leveling, load balancing, and timing (Land 2004: 36). *Load leveling* refers to the smoothing of the capacity loading over time by advancing (within the time limit) or delaying the release of particular orders. *Load balancing* requires releasing a mix of orders that maintain a balanced capacity loading across workcenters over time, avoiding both temporary bottlenecks and idleness. These aspects of order release seek to maximize throughput, but also support the *timing* function that seeks to ensure high due date performance by releasing urgent orders before less urgent ones.

A wide variety of order release mechanisms can be obtained by specifying several *design options* left open by this generic procedure in different ways. Following Bergamaschi et al. (1997), these are summarized in Table 4.1 using the terminology common to this literature.

*Order release mechanism*: “Under load limited OR [order release], orders are released to the shop based upon their distinctive features and the existing workload in the shop” (Bergamaschi et al. 1997). The workload in the shop can be defined and measured in various ways as discussed below. Workload limits may be determined for each workcenter, the entire production unit, or both based on the characteristic curves of the respective unit (Fig. 4.1), and the order release mechanism prevents these load limits from being exceeded. Time-phased release, in contrast, computes a planned release time for each order without explicitly considering the workload.

*Timing convention*: Order release is usually performed at the beginning of each period. The time between two consecutive invocations of the order release logic, referred to as the *check period*, need not be equal to the planning period (Perona and Portioli 1998), although this is often the case; event-driven order release in continuous time is also possible as in CONWIP (Spearman et al. 1989, 1990). Both options can be used simultaneously. For instance, LUMS uses an “intermediate pull release” option that can release orders within a check period to prevent workcenters from starving (Stevenson and Hendry 2006).

*Workload measure*: Workload can be measured in number of orders or in amount of work (e.g., standard hours). Under work-conserving dispatching rules, the orders at the workcenter completely determine the amount of work; under sequence-dependent setup times, this is no longer the case (see Sect. 2.1).

*Aggregation of workload measure*: The workload limit can be defined for the entire production unit (total shop load), only for bottleneck workcenters, or for each workcenter.

*Workload accounting over time*: If target WIP levels are defined for individual workcenters, WLC can control the *direct load* at each workcenter, defined as work that has arrived at the workcenter and requires processing there. The direct load in the future—even its evolution during the period under consideration—has to be estimated at the time of the release decision, since the input to a workcenter is controlled by scheduling decisions at upstream workcenters and hence is not known with certainty.

**Table 4.1** Design options in traditional order release mechanisms (Bergamaschi et al. 1997)

| Design option | Choices |
|---|---|
| Order release mechanism | Load limited, time phased |
| Timing convention | Discrete, continuous |
| Workload measure | Number of jobs, workload quantity |
| Aggregation of workload measure | Total shop load, bottleneck load, workcenter load |
| Workload control | Upper bound, lower bound, upper and lower bounds, workload balancing |
| Capacity planning | Passive, active |
| Schedule visibility | Limited, extended |

This forecasting of detailed order arrival patterns at specific workcenters can be avoided by using the aggregate load as the relevant workload measure. The *aggregate load* of a workcenter is the sum of the direct load and the *load in transit*, work already released to the shop that will require processing by that workcenter but has not yet arrived there. This requires that the release function have access to information on the progress of jobs through the shop when it performs its calculations. In small or medium businesses, such feedback may not be available at all points on the routing or at all points in time. If feedback from the shop floor is only available on order completion, the *extended aggregate load*, given by the sum of the operation times at the workcenter for all orders in the shop, can be used (Oosterman et al. 2000; Henrich et al. 2004). The *corrected aggregate load* (Oosterman et al. 2000: 112) is calculated by multiplying the aggregate load by a factor that can be interpreted as the steady-state ratio of aggregate load to average direct load, providing an estimate of the direct load (Missbauer 2009).

These load measures are defined for each point in time, most importantly for the start of the current planning period for which the order releases are being determined, and allow limited predictions of the future load situation. A workcenter's aggregate load and its capacity jointly determine the time span for which the released work can keep the workcenter busy, as long as its direct load is high enough to avoid excessive idle time. The *time bucketing approach* divides the time horizon into discrete time buckets. Based on cycle time estimates for each operation the order must undergo, a forward finite loading technique is used to check whether releasing an order will violate a load limit at any workcenter along its routing. The operations of the candidate order are assigned to the required workcenter in the period determined by the estimated cycle time if capacity is available. Otherwise the operation is loaded in the earliest future period with available capacity. The decision whether or not to release the order is based on the match between calculated and required due date (Bobrowski 1989). Although release decisions are only made for the current period, explicitly considering multiple periods opens the possibility (at least in principle) of extending the method to order release *planning*, the second order release approach in Sect. 4.1.2.

*Workload control*: This option specifies whether upper and/or lower bounds for the workload are used.

*Workload balancing*: This refers to the presence or absence of logic that compensates for the overloading of certain workcenters by adjusting the loading of others (Bergamaschi et al. 1997).

*Capacity planning and schedule visibility*: In most rule-based order release mechanisms, the workcenter capacities are fixed (passive capacity planning). Schedule visibility refers to the look-ahead capability of the release mechanism that defines its ability to perform load leveling (smoothing) over time and is controlled by the time limit (Zäpfel and Missbauer 1993a).

We now discuss some important release mechanisms of this type to illustrate the variety of WLC approaches possible.

### 4.2.2 Load-Oriented Order Release

Load-Oriented Order Release was developed at the University of Hannover (Bechte 1980, 1988) and is presented in detail in Wiendahl (1995). The approach is largely based on the characteristic curves shown in Fig. 4.1 and is linked to flow diagrams and their extensions that are recommended as diagnosis tools (Nyhuis and Wiendahl 2009).

In Load-Oriented Order Release, orders are released in each period (in practice 1 day to 1 week) for the current planning period only. A time limit is used to prevent premature release of orders. A target WIP level, measured in hours of work, is defined for each workcenter and represents an upper bound on the amount of WIP that can accumulate at that workcenter. It is expressed as a limit on the *direct load* $\Lambda_t$ of the workcenter in the planning period *t,* given by the sum of the WIP available at the start of the period and the estimated work arriving during the period. The release algorithm follows the generic release procedure described in Sect. 4.2.1. The limit on the direct load of a workcenter, referred to as the *load limit*, is used such that the first order that exceeds the load limit of a workcenter in the release run can still be released. After release of this last eligible order, the workcenter is blocked and all other orders requiring this workcenter are rejected in the current release run.

The value $\Lambda^{\max}$ of the load limit is derived from the target WIP level assuming an idealized material flow at a workcenter whose input and output rates are held constant and equal to its capacity $C$ (stated in units of time) during the planning period. WIP is assumed to be held constant at the target WIP level. In this situation, the load limit for the planning period is given by the sum of the target WIP level $\hat{W}$ and the available capacity $C$ in the planning period, giving

$$\Lambda^{\max} = \hat{W} + C \tag{4.1}$$

It is often convenient to express the load limit as a percentage of the capacity per period. This *load percentage* $\Lambda^{\text{percent}}$ is defined as

$$\Lambda^{\text{percent}} = \frac{\Lambda^{\max}}{C}\left(100\%\right) \tag{4.2}$$

By Little's Law, the average cycle time $T$ is

$$T = \frac{\hat{W}}{C} \tag{4.3}$$

Substituting (4.1) into (4.2) yields

$$\Lambda^{\text{percent}} = \frac{\hat{W} + C}{C}\left(100\%\right) \tag{4.4}$$

Finally, substituting (4.3) into (4.4) yields

$$\Lambda^{\text{percent}} = \left(1+T\right)\left(100\%\right) \tag{4.5}$$

Thus a target flow time of one planning period requires a load percentage of 200% of capacity, a target flow time norm of two periods a load percentage of 300%, and so on.

The precise arrival time of an order at any workcenter that is not the first in its routing is uncertain at the time of order release. Load-Oriented Order Release approaches this issue by estimating the probability that the order arrives at the workcenters on its route during the planning period under consideration. The order then contributes its expected work content, given by the product of its processing time (including setup time) and the probability of its arrival during the period, to the planned load at these workcenters during the planning period. This probability is derived from the load percentage $\Lambda_m^{\text{percent}}$ at workcenter *m,* as defined in (4.2). $\frac{100}{\Lambda^{\text{percent}}}$ then represents the fraction of the direct load at workcenter $m$ that leaves that workcenter during the period. This can be interpreted as the probability that an order currently contributing to the direct load of workcenter $m$ will arrive at the subsequent workcenter during the planning period. If these probabilities are assumed to be statistically independent, the probability that an order will pass through the first $m$ workcenters of its routing and arrive at workcenter $m$+1 within the planning period is

$$p\left(m\right) = \prod_{k=1}^{m} \frac{100}{\Lambda_k^{\text{percent}}} \tag{4.6}$$

and its expected contribution to the direct load of workcenter $m$+1 is the product of the probability (4.6) and the order's operation time at workcenter $m$+1. These orders contribute only the fraction (4.6) of their operation time to the direct load of a downstream workcenter, which increases the closer they are to workcenter *m.* Hence the direct load at a workcenter can increase in a period even if no orders are released in that period.

Equation (4.6) is a particular solution to a general problem that occurs in order release models that seek to control the length of individual queues in production units with multiple workcenters. These models require an estimate of the fraction of releases in a period $\tau$ that contributes to output (or to the load of the downstream workcenters) in periods $t = \tau, \tau + 1, \ldots$. The probability (4.6) is treated as a deterministic fraction. Thus, if we consider an operation that is to be performed at workcenter $m$+1 on an order $j$ released in period *t,* the fraction $w_{j,t,t}^{m+1}$ of the time of this operation that reaches workcenter $m$+1 in the release period $t$ is assumed to be

$$w_{j,t,t}^{m+1} = \prod_{k=1}^{m} \frac{100}{\Lambda_k^{\text{percent}}} \tag{4.7}$$

Since only one planning period is considered, the $w_{j,\tau,t}^{m+1}$ values for $\tau < t$ need not be specified. We shall encounter these *loading factors* again in Chap. 6.

Load-Oriented Order Release was developed from extensive work on monitoring and diagnosing the material flow in manufacturing systems. Standard software that implements both the monitoring/diagnosis aspects and the order release algorithm has been available since the 1980s (Bechte 1988), and good results have been reported. A description of a typical software functionality can be found in Wiendahl (1995: 292ff.), and Yan et al. (2016) present several improvements to the method.

### *4.2.3 The LUMS Order Release Mechanism*

The Lancaster University Management School (LUMS) approach (Stevenson and Hendry 2006) is a comprehensive framework for a PPC system based on WLC specifically for make-to-order (MTO) companies. It encompasses order release as well as customer enquiry and order entry and, at least conceptually, priority dispatching.

The LUMS order release mechanism follows the generic procedure described above. Orders are released periodically, but mechanisms for event-oriented release are included to prevent workcenter starvation and late orders. Work is measured in time units (e.g., standard hours), and the workload measure is the aggregate load (direct load plus load in transit) or—more recently—the corrected aggregate load (Thürer et al. 2012). A limit is set on the Released Backlog Length per workcenter, the time required to process the aggregate load given the current capacity constraints. A non-enforced lower bound is provided to foremen as a decision support, e.g., for intermediate pull release.

### *4.2.4 CONWIP*

CONWIP (CONstant Work In Process) (Spearman et al. 1989, 1990; Hopp and Spearman 2008) is designed for controlling serial production lines as an alternative to Kanban with less restrictive requirements. In the form described in Hopp and Spearman (2008), "a new job is introduced to the line each time a job departs," which "results in a WIP level that is very nearly constant" (Hopp and Spearman 2008: 363). The WIP level is measured in units of product. Measuring WIP in work content, e.g., hours of work at the bottleneck, has also been proposed (Spearman et al. 1989); this is largely equivalent to the number of orders if, as assumed in Spearman et al. (1990), "that parts are moved in standard containers, each of which contains roughly the same amount of 'work'." Order release is driven by events (departure of jobs) in continuous time, not on a periodic basis. Unlike most other order release mechanisms of this type, CONWIP is not designed for job shop control, which must cope with substantially more complex material flows than serial production lines. The approach is less restrictive than Kanban, which assumes a serial production-inventory system with low demand variation over time for each product, whereas the product mix in the CONWIP line can vary. Therefore, CONWIP can be viewed as a generalized form of Kanban (Spearman et al. 1990).

## 4.3 Design and Parameter Setting of Order Release Mechanisms

Designing a rule-based order release mechanism for a specific system requires:

- Determining the design options that specify the release mechanism
- Determining the values of the parameters required by the selected design options (time limit, target WIP level or load limit, length of the planning period and of the check period)

The extensive body of research in this area over the last three decades, mainly using simulation, seeks to provide a *rule base* suggesting which design options from the menu in Table 4.1 and what values of the associated parameters are most appropriate under which circumstances. To the best of our knowledge, a unified summary of the design rules studied so far does not yet exist, and providing this is beyond the scope of this volume. Establishing rules for setting the parameters of order release mechanisms is easier and has been explored extensively in the WLC literature, suggesting the following guidelines:

The *time limit* determines how far in advance of their planned start dates orders can be released, determining the ability to perform load leveling over time. The demand pattern strongly influences the optimal time limit: under low demand, and thus low resource utilization, there is no reason to release orders before their planned start date since this would just increase finished goods inventory. Under time-varying demand, a longer time limit allows load leveling and is advantageous unless load leveling is done at the medium-term planning level, before the order release logic is invoked (Zäpfel et al. 1992). Thus any general guideline on the optimal length of the time limit (for instance, Wiendahl (1995) recommends three planning periods as a reasonable starting value) represents a situation-specific compromise.

The WIP targets at the workcenters are closely related to the target value of the average cycle time (see Sect. 4.1). They can be expressed in different ways depending on the order release mechanism (e.g., load limit or load percentage in Load-Oriented Order Release) and define the trade-off between output, WIP and cycle time. Since WIP acts as a buffer against various types of variability, including demand variability, a higher WIP level allows the system to operate under higher demand variability without starving bottleneck workcenters. Thus the optimal values of the target WIP level and the time limit may be related, but to our knowledge this issue has not been studied.

Other parameters that we do not discuss in detail include the length of the planning period and the order release frequency if orders are released periodically (Perona and Portioli 1998). Event-driven release in continuous time as in CONWIP represents a limiting case where the check period approaches zero.

Until now we have considered the order release decision assuming that the volume of orders arriving in the pre-shop pool is compatible with available capacity in the medium term. For the WLC system to work properly, the order release mechanisms must be integrated into a PPC system that generates the orders and their required due dates, thus controlling the input of new work to the pool. Since this order input is usually based on the Master Production Schedule (in the MPC framework) or the Master Plan (in the APS framework), unless independent demand inventory control systems are applied, the Master Production Scheduling or Master Planning functions must be structured to support the order release mechanism from this perspective.

## 4.4 Rule-Based Workload Control Within the MPC system

The relationship of the Master Production Schedule (MPS) to the inflow of orders to the pre-shop pool depends on the architecture of the PPC system. We must distinguish between PPC systems based on a mid-term MPS derived from demand forecasts and/or customer orders whose production orders are generated by the MRP and lot-sizing logic, and MTO systems whose production orders are generated directly from accepted customer orders. We now describe the integration of order release mechanisms for these two structures and the shortcomings of the resulting system architectures.

Much WLC research has focused on MTO companies where the acceptance of customer orders leads directly to the addition of production orders to the pre-shop pool. For MTO companies, the released orders in the shop and the as yet unreleased orders in the pool are treated as a *hierarchy of workloads* (Kingsman 2000) that are controlled by order input and capacity decisions at different stages of an order's progress through the production system. Three such stages are usually distinguished (Land 2004):

*Order entry*, which determines the input to the pre-shop pool by order acceptance/rejection and due date assignment. At this stage, medium-term capacity adjustments based on the volume of work accepted can be recommended.

*Order release*, which determines the input to the production unit through order release decisions. At this stage, short-term capacity adjustments can be recommended to ensure their timely completion.

*Priority dispatching*, which determines the sequencing and timing of the operations for the released orders. Capacity adjustments in the very short term, such as expediting or unplanned overtime, constitute the corresponding output control. Note that the dispatching rule or shop-floor scheduling policy in use can influence both the total amount of work to be performed, if setup times are sequence dependent, and the workload balance among workcenters if it considers the load at downstream workcenters.

LUMS divides the order entry stage into two stages (Kingsman 2000; Stevenson and Hendry 2006):

*Customer enquiry*: A proposal to deliver the requested work at a specified time is made in response to a customer enquiry and the customer's decision is awaited.

*Order entry*: An order enters the company if the customer accepts the proposal. Order entry is considered as a separate decision since the company's situation may

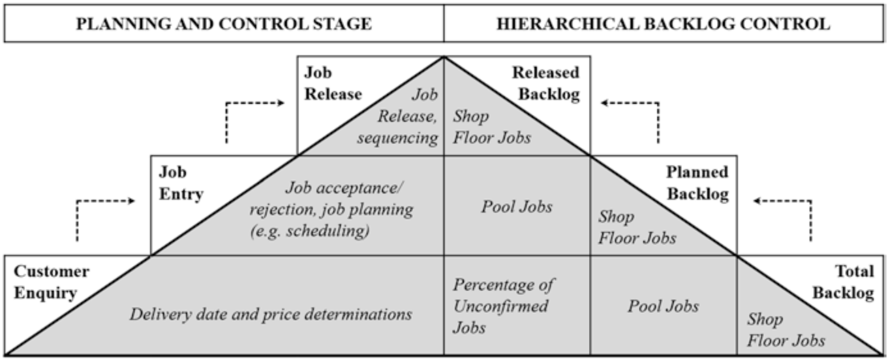

PLANNING AND CONTROL STAGE
HIERARCHICAL BACKLOG CONTROL
Job Release
Job Release, sequencing
Released Backlog
Shop Floor Jobs
Job Entry
Job acceptance/ rejection, job planning (e.g. scheduling)
Pool Jobs
Planned Backlog
Shop Floor Jobs
Customer Enquiry
Delivery date and price determinations
Percentage of Unconfirmed Jobs
Pool Jobs
Total Backlog
Shop Floor Jobs

**Fig. 4.2** Planning and control levels and hierarchy of workloads (Stevenson and Hendry 2006, modified)

change between the proposal being made and its acceptance by the customer (Kingsman 2000).

Hence an order must pass through four stages before exiting the system: customer enquiry, order entry, order release, and dispatching. This sequence of events on each order induces a hierarchy of workloads (backlogs) and an associated hierarchy of lead times. Figure 4.2 depicts the relationship between the planning and control stages and their associated backlogs, while Fig. 4.3 depicts the timeline and the associated hierarchy of lead times.

The hierarchy of workloads implied by the stages of an order's progress through the system can be described as follows: The *released workload* consists of all released orders in the production unit and is controlled by order release, based on the planned shop cycle time. The *planned workload* combines the orders in the pool and the released workload and is controlled by the order entry level. The total cycle time for an order is the sum of pool waiting time and shop cycle time. The *total workload* further includes orders that have not yet entered the pre-shop pool. Controlling the total workload requires an estimate of the probability that a proposal will be accepted by the customer (Kingsman and Mercer 1997). The delivery lead time to the customer is related to this total workload.

Keeping these cycle times at the specified target levels requires holding the respective workloads at their target levels, which must be derived from the target cycle times and the workcenter capacities. This is done by a combination of input control at the respective decision points and capacity adjustments (Kingsman 2000). We define the following notation:

$PB_{mt}$: planned workload for workcenter $m$ at the end of period $t$, defined as the total work content of all jobs in the job pool requiring this workcenter and all jobs on the shop floor requiring this workcenter that have not yet completed their processing there. This represents the total work required to complete all jobs currently in the pool and the released workload.

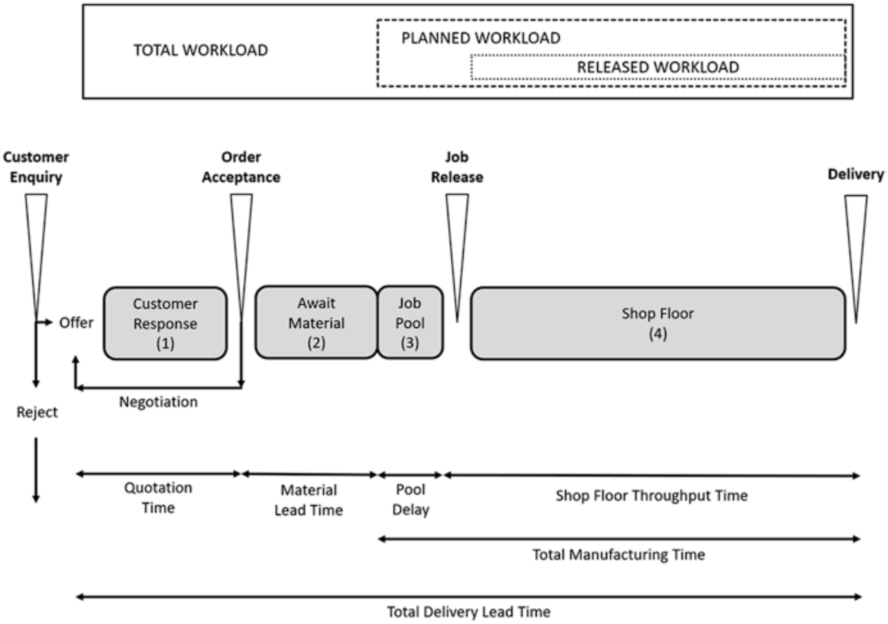


**Fig. 4.3** Lead time hierarchy in WLC systems for MTO companies (Kingsman 2000, modified)

$\mathrm{TB}_{mt}$: total workload of workcenter $m$ at the end of period $t$, defined as the total work in the firm that has to be carried out on workcenter $m$.
$\tau_{\mathrm{p}}$: planned total cycle time from entering the order pool until completion of the last operation.
$\tau_{\mathrm{D}}$: target (maximum) delivery time.
$\tau_{\mathrm{s}}$: target (maximum) shop cycle time.
$X_{mt}$: output of workcenter $m$ in period $t$.

In order to maintain the total cycle time $\tau_{\mathrm{p}}$ and the delivery time $\tau_{\mathrm{D}}$ at their planned values, the workload at each workcenter must be limited to the maximum value that can be cleared within the target lead time, implying the constraints

$$\mathrm{PB}_{mt} \le \sum_{k=t+1}^{t+\tau_{\mathrm{p}}} X_{mk}, \quad \text{for all} \quad m, t = 1, \ldots, \tau_{\mathrm{D}} - \tau_{\mathrm{P}} \tag{4.8}$$

$$\mathrm{TB}_{m0} \le \sum_{k=1}^{\tau_{\mathrm{D}}} X_{mk}, \quad \text{for all} \quad m \tag{4.9}$$

Equation (4.9) states that the total workload of each workcenter at the current time, which is the start of period 1 and end of period 0, must be processed within the maximum delivery time $\tau_{\mathrm{D}}$. If (4.8) is to be formulated for periods $t = \tau_{\mathrm{D}} - \tau_{\mathrm{P}} + 1, \ldots, \tau_{\mathrm{D}}$ as well, end of horizon effects must be considered appropriately. The material balance constraints for $\mathrm{PB}_{mt}$ are given in Kingsman (2000).

In (4.8) and (4.9) the output is limited by the available capacity

$$X_{mt} < C_{mt}, \quad \text{for all} \quad m, t = 1, \ldots, T \tag{4.10}$$

As in (4.8) and (4.9), the aggregate load for each workcenter *m*, given by its released workload *RB*$_{mt}$, cannot exceed the relevant upper bound. The resulting constraint equivalent to (4.8) would be

$$\mathrm{RB}_{mt} \leq \sum_{k=t+1}^{t+\tau_\mathrm{S}} X_{mk} \quad \text{for all} \quad m,\ t = 1, \ldots, \tau_\mathrm{D} - \tau_\mathrm{S} \tag{4.11}$$

Additional elaboration of the constraints may be necessary to incorporate additional factors that must be considered at this level, such as the fraction of high-priority orders and the order's operation times at the workcenters (Hendry and Kingsman 1991a).

Constraints (4.8)–(4.11), with suitable modifications, ensure the feasibility of the cycle times provided that the workload norms (the specified upper bounds) used allow output close to the maximum capacity. The complete mathematical formulation, including additional constraints that guarantee due date feasibility of the orders, can be found in Kingsman (2000). The model is dynamic and relates the workload norms to the future output and operation due dates within the delivery time $t_\mathrm{D}$. Hence the LUMS approach exceeds our narrow definition of rule-based workload control by incorporating some characteristics of the multi-period optimization models for order release models discussed in Chap. 5. Thürer et al. (2014) describe the integration of customer enquiry management, including due date setting, with order release based on LUMS.

The lowest level in the WLC decision hierarchy is priority dispatching. Early WLC literature (Ragatz and Mabert 1988) states that limiting the WIP level reduces the differences in performance between dispatching rules due to shorter queues and makes simple FCFS dispatching preferable provided that there is no unplanned delay of the orders once they have entered the shop due to factors such as machine breakdowns or quality problems (Bechte 1988). This shifts complexity from the dispatching level to the release level, suggesting that appropriate release control can largely obviate the need for delaying or expediting of orders on the shop floor. Due-date-oriented rules can be beneficial since they reduce the lateness dispersion and can correct progress disruptions of orders (Kayton et al. 1997; Land 2004). The case of sequence-dependent setup times is more complex since setup time reduction and due date performance are conflicting goals and the optimal dispatching rule depends on the importance of these goals and on the significance of the setup times (Thürer et al. 2012). An interesting study of precisely this issue in the context of semiconductor manufacturing was conducted by Lu et al. (1994), who concluded that in their experiments, order release was the primary determinant of mean performance measures, but dispatching rules can still have significant impact on their variance.

The LUMS decision structure was developed specifically for MTO companies. If rule-based WLC is applied, e.g., to component manufacturing shops that receive their orders from an MRP system, order release must be integrated into the MPC architecture. In this case, order release is based on the results of the MRP/lot-sizing calculation and the associated infinite capacity loading that determines the planned start and finish dates of the orders (Vollmann et al. 2005). Ideally, the WLC functionality replaces the CRP/load-leveling logic provided that the orders from the MRP system are balanced with the capacities in the medium term. The extent to which load leveling is necessary after the MRP run depends on the degree to which this condition is satisfied and on the ability of the release mechanism to cope with time-varying demand. This interface can be complex and is discussed in the next section. Wiendahl (1995: 279ff.) and Zäpfel and Missbauer (1993a) discuss the integration of order release into the MPC system in this situation. The situation is largely analogous under an APS system if the orders result from Master Planning and a subsequent lot-sizing stage where capacity requirements and availability are computed for all production units in the master plan. This is the case, e.g., in semiconductor manufacturing for which tailored order release approaches have been developed (Moench et al. 2013, Chap. 6).

## 4.5 Critical Assessment of Rule-Based Workload Control

Rule-based workload control, defined in the broad sense as WLC order release mechanisms supported by medium-term planning to balance workload with available capacity and appropriate dispatching rules to maintain the flow of orders through the production unit, is an important way to achieve short, predictable shop cycle times and improved shop-floor transparency. Unlike model-based WLC approaches, it decomposes the production planning problem into a short-term order release subproblem and a medium-term load-leveling and capacity adjustment subproblem. This allows relatively straightforward release mechanisms that do not require accurate time-phased load projections that are difficult to obtain (Bechte 1980; Tatsiopoulos and Kingsman 1983). The complexity of the overall planning problem is largely shifted to the medium-term planning level, which must be effectively integrated with short-term order release. Hence the overall planning system can be very complex, especially when the system faces substantial demand variability and thus deviates from idealized steady-state assumptions.

An example of this limitation is the dynamic behavior of Load-Oriented Order Release (Sect. 4.2.2) in production systems with directed material flow, such as a flow line, with its bottleneck at the end of the line. If the system starts with no WIP and a large number of orders in the pool whose planned start dates lie within the time limit, Load-Oriented Order Release cannot prevent overloading of the bottleneck workcenter in future periods even if the load limit is not exceeded in the current planning period. This is because the load of the bottleneck workcenter in period 1 is mostly still at upstream workcenters, and only the *expected* direct load, a fraction of the actual operation time, contributes to the load in period 1 per (4.7). Several periods later, the work actually arrives at the bottleneck workcenter, requiring its full operation time and leading to an overload situation.

We now discuss the limitations of rule-based WLC, namely their concentration on the cycle times in the production unit and their integration into the overall PPC system. The latter issue motivates the multi-period order release models with fixed lead times described in Chap. 5.

The central idea of most rule-based WLC methods is to maintain a target WIP level, and thus planned cycle times, in the production unit by controlled release of orders, decoupling short-term release decisions from medium-term planning and material coordination. However, the unreleased orders that do not contribute to the WIP (and the workload) within the production unit do not disappear; they remain in the order pool outside the production unit. Thus it is critical to assess the performance of rule-based order release when the system boundary is extended beyond the production unit. If the production unit is modeled as a queueing network, controlled order release changes an open queueing system to a controlled one (Kushner 2001) where the orders that arrive from "outside" first enter an admission queue (order pool) and are then admitted to the queueing network according to the release algorithm. Thus the effects of controlled order release can be analyzed by means of queueing models or simulation.

Compared to an open queueing system with an exogenous arrival process, controlled order release reduces WIP and shop cycle times (from release to completion) not only because of the WIP reduction for a *given* functional relationship between WIP and output, but because it induces negative correlation between the WIP level at the workcenters and the probability of additional input in the near future, altering the functional relationship between WIP and output in a favorable direction. For CONWIP, which in its basic form is a serial, single-class closed queueing network, Hopp and Spearman (2008) state that "for a given level of throughput, a push system will have more WIP on average than an equivalent CONWIP system." However, this assumes an order pool sufficiently large to prevent unnecessary idle time and does not count the unreleased orders waiting in the pool as WIP. Kanet (1988) argues that for a single-stage queueing system with an autonomous external arrival process, controlled release merely partitions the system into two subsystems: the pool of unreleased orders and the server with the queue containing the released, but as yet unfinished, orders. Thus the total cycle time from arrival to completion remains unchanged. For production systems with complex material flows, like job shops, the situation is different since the order release mechanism has no direct control over the input to the individual queues. In this case, limiting the WIP level by controlled releases can lead to idle time at some workcenters, reducing available capacity and actually *increasing* total cycle time since the admission queue faces a system with reduced capacity. This highlights the importance of the master production scheduling function in decoupling the external demand from the order input to the pre-shop pool (Spearman and Zazanis 1992).

This argument can be evaluated by modeling a workload control system as a semi-open queueing network (Jia and Heragu 2009) in which orders arrive at an admission queue from where they enter the queueing network, which admits at most $N$ orders. Each time a completed order leaves the system, a new order is introduced to the queueing network if it is available in the admission queue; otherwise the number of orders in the queueing network is reduced by 1 until the next order arrives. Unfortunately, analytical tools for analyzing semi-open queueing networks are as yet very limited (Heragu and Srinivasan 2011).

Jansen (2012) shows by means of an analytical model that the capacity reduction effect described by Kanet (1988) can occur even in a single-stage queueing model if orders are released periodically and there is a maximum WIP level of *N* orders at the server. If the total operation time of the *N* orders at the server after release is less than the capacity during the release period, the server may idle even with many orders in the admission queue, reducing the available capacity and hence the maximum arrival rate that allows a stable system. The average length of the admission queue approaches infinity as the arrival rate approaches this threshold value (Jansen 2012). In a single-stage system, this phenomenon can only occur if the load limit is expressed in number of orders (not in amount of work, which is more usual for periodic release) and if there is no possibility of releasing orders during the period to prevent idleness, such as the intermediate pull release option in LUMS. Nevertheless, the model clearly supports the argument in Kanet (1988).

The arguments presented so far do not consider systematic changes in the product mix and therefore cannot fully account for the complex material flow structure in production units. Simulation can overcome this limitation and hence has been the most prevalent research tool for this type of order release mechanism. The impact of limiting workload under a specified release mechanism is usually depicted as parametric curves with the target WIP level (measured appropriately) as control variable, shop cycle time on the *x*-axis and other performance measures, especially the total cycle time (in the shop and the pool), on the *y*-axis. Figure 4.4 shows a typical graph obtained from simulation.

The highest value of the shop cycle time (*x*-axis) and the corresponding total (gross) cycle time (*y*-axis) are obtained with unconstrained WIP, i.e., immediate release of arriving orders. Limiting the WIP level reduces the shop cycle time, which in the graph also *reduces* the total cycle time and the percent tardy for FCFS dispatching, contradicting the arguments presented above. However, this is not the case for Operation Due Date (ODD) dispatching (Land 2004: 71). Hence in at least some cases controlling the workload in the shop can reduce both shop cycle time and total cycle time as long as the target WIP level is not too low.

An important explanation for this phenomenon is the load balancing capability of the release mechanisms, which is especially important in multi-product situations (Van Ooijen 1998). Orders that require certain workcenters may arrive at a lower or higher rate at different times, so releasing all orders immediately can lead to temporary overloading of certain workcenters and idling of others unless the dispatching policy can prevent this. WLC generally seeks to prevent the release of orders processed on overloaded workcenters by releasing instead orders that require underloaded workcenters. Ideally, a balanced mix of orders is released by changing the release sequence of the orders relative to the sequence of their planned start dates, avoiding temporary idleness and increasing throughput.

Even if limiting WIP and shop cycle time leads to a modest increase in total cycle time, it may be economically preferable depending on the relative importance of

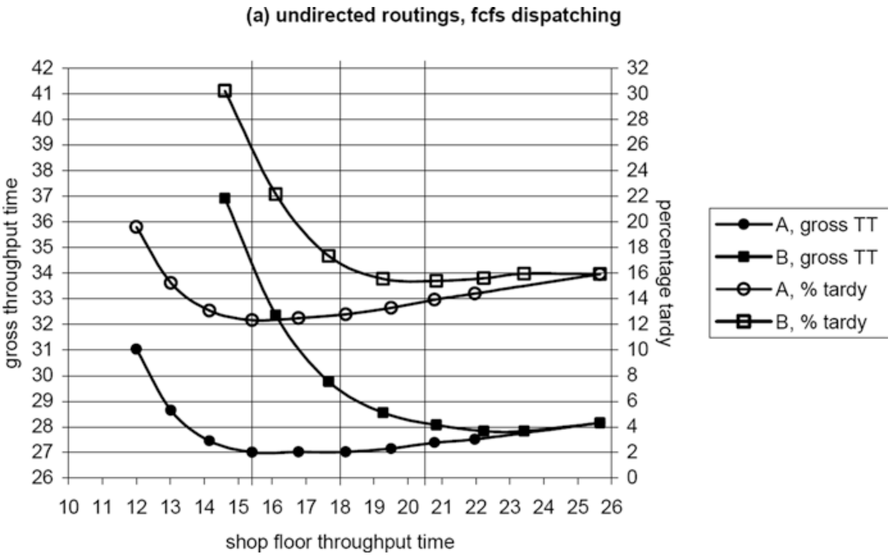


**Fig. 4.4** Average shop cycle time (shop floor throughput time) and average total cycle time (gross throughput time) for different target WIP levels and FCFS dispatching (A: direct load norms, B: aggregate load norms) (Land 2004: 70)

these objectives, specifically the ratio of WIP costs to lead time costs (Bertrand and Van Ooijen 2008). There are several reasons to emphasize the importance of the shop cycle time and the advantages of delayed order release (Land and Gaalman 1998; Thürer et al. 2010). Inventory holding costs for raw materials are usually lower than those for WIP inventory due to the lower value of the items. Delaying the operations performed on an order also delays the associated cash outflows relative to given cash inflow by the customer which implies reduced investment in inventory and lower interest payments (Grubbstrom 1980).

Although in principle most WLC methods require that the material for an order in the pool be available, the same material can be used for several products, but once the first operation is started, the material cannot be used for anything else. Thus postponing order release can realize pooling effects through delayed ordering of raw materials and reduce the risk of obsolescence. Whether the constraint of material availability for the orders in the pool can be relaxed, and if so to what degree, is difficult to say in general.

From a planning perspective, under WLC the time at which control over the material flow is passed from the production planning level to the production unit is delayed. Ideally this leads to predictable finish dates for the planning level since the release dates are known and the shop flow times are under control. The high variability of cycle time distributions observed in practice was an important motivation for developing the WLC concept, and delaying released orders due to changes in demand or due date can substantially contribute to this variability (Fig. 2.3). It can also be argued that the perspective of the planning level goes beyond the production unit as an "internal supplier," and retaining control over orders at the planning level as long as possible may provide benefits. Reduced queue lengths tend to reduce both the importance of dispatching and its complexity.

The effects of rule-based order release mechanisms on flow times and other performance measures like due date performance have been investigated extensively by simulation, often examining the effects of different design options and parameter settings for the release mechanisms. *Empirical* field research (McKay 2010a, b) on the effects of implementing WLC release mechanisms has been relatively rare, although positive effects were reported in the literature relatively early (Bechte 1988; Wiendahl 1992; Thürer et al. 2011). Some researchers (Melnyk and Ragatz 1989; Bertrand and Van Ooijen 2002) state that these positive experiences cannot be explained by the limited evidence for improvements obtained from simulation studies. It is often concluded that effects of WLC that are usually not captured in simulation models might explain this. Land (2004: 171) distinguishes between non-modeled aspects of the shop and non-modeled aspects of the planning system but concludes that "…our simulation results showed that strongly restricting the quantity of work on the shop floor should not necessarily jeopardize other performance aspects" by some of the reasoning given above. However, integration of rule-based WLC into the larger PPC system is complex, and this complexity may well limit our ability to predict the impact of implementing these release mechanisms on the overall system performance.

Starting around 2005, several empirical studies have sought to address this research gap by using case studies to identify relevant research issues that have not received sufficient attention in WLC theory and by using WLC implementation projects to gain experience and insights into the implementation process, necessary refinements of the concept and methods, and their practical performance (Henrich et al. 2004; Hendry et al. 2008, 2013; Stevenson and Silva 2008; Stevenson et al. 2011; Soepenberg et al. 2012; Silva et al. 2015). Thürer et al. (2011) give an overview of WLC implementations, while Hendry et al. (2013) and Silva et al. (2015) give a detailed report on a particular WLC implementation. A recent report on a WLC implementation as well as a review of the empirical WLC literature, structuring the research into work on implementation results, applicability/implementation process, and implementation strategies, can be found in Hutter et al. (2018).

The relative importance of shop cycle time and total cycle time is related to the process governing order input to the pool, and hence to the degree of integration of the release algorithm into the overall planning and control system. The myopic nature of rule-based release mechanisms requires a medium-term planning level that balances load and capacity for the next several periods. Keeping the hierarchy of workloads at their target levels per Sect. 4.4 is consistent with WLC-based order release for MTO companies. If the order release mechanism is integrated into an MPC system based on MRP/MRPII (Sect. 3.1) or the APS logic (Sect. 3.2.2), capacity planning usually balances the required and available capacity in each period, which is not necessarily consistent with keeping WIP at a desired level. Conventional capacity planning techniques and order release mechanisms differ in their modeling assumptions. For instance, assuming that each order is completed within its planned operation lead times is not the same as assuming the loading factors (4.7) used in Load-Oriented Order Release. Hence the interface between mid-term planning and order release is critical.

It is also difficult to determine the extent to which load variations must be smoothed out by mid-term production and capacity planning. Appropriate parameter values allow order release mechanisms to cope with varying demand patterns to a certain extent. The values of the time limit and target WIP level that provide the desired smoothing behavior and lead to the desired compromise between throughput and cycle time depend strongly on the demand pattern and also, quite possibly, on each other. Since the demand pattern may change over time, these values will generally be time-dependent. Even if reliable demand forecasts are not available, the choice of these parameters implies certain assumptions about future demand. Zäpfel and Missbauer (1993a) formulate aggregate models of a production unit that is aggregated to a single workcenter and derive the optimal time and load limits for Load-Oriented Order Release based on the optimal material flow obtained from these models. This provides an alternative to a rule-based parameter setting of the form *if* underutilization *then* time limit close to zero, etc. Apparently this line of research has not been pursued, but it is a step towards optimization of order releases by means of multi-period optimization models: The optimization model estimates the aggregate material flow over all products and workcenters, and the release mechanism disaggregates the resulting release quantities, measured in hours of work, to individual orders. If the production unit is modeled in more detail, disaggregation is easier, eventually leading to multi-period planning models that jointly determine the material flow and the order releases. There appears to be little research on the relationship between rule-based workload control and planning models that optimize the material flow over a multi-period planning horizon.

By limiting their scope to individual production units and short-term order release, rule-based workload control mechanisms treat an isolated subproblem whose integration into the overall PPC system is not straightforward. This motivates the development of planning models that embed order release into mid-term planning and thus avoid these shortcomings, which are treated extensively in this volume. It is based on mathematical models for mid-term production planning, and this stream of literature is reviewed next.

## 4.6 Mathematical Models for Mid-Term Production Planning

While production planning has obviously been executed in some form since the beginning of even craft production, quantitative methods for these problems are of surprisingly recent origin. McKay (2010a, b) reviews the historical development of production planning since the beginning of the Industrial Revolution. While the well-known work of Harris (1915) launched the area of inventory modeling, it was not until the 1950s that this became a major research area (Arrow et al. 1958). Optimization models for planning production over time have their origins in the work of Modigliani and Hohn (1955), although its roots reach back to the activity-based economic models of economists such as Leontief (1951), Koopmans (1951), and Schneider (Whitin 1954). The area of sequencing and scheduling also had its pioneering papers in this period, notably those of Jackson (1955) on single machine scheduling problems and Manne (1960) on job shop scheduling. Thus the basic formulations of most classical production planning and scheduling problems were essentially in place by 1960, when the book by Holt et al. (1960) collecting their previous work appeared. It is worthwhile examining these early papers in some detail to understand why these formulations developed the way they did, and their implications for the current situation.

The work of Modigliani and Hohn (1955) views the problem of production planning over time as that of trading off production costs against inventory holding costs. Production costs are assumed to be convex with increasing marginal production costs, while inventory holding costs are approximated by the time average of the ending period inventories, leading to a linear cost function. The problem is formulated on discrete time periods, the cost function is assumed to be stationary over time, demand in each period is known with certainty, and no backlogging is allowed. The monotone increasing marginal production cost makes it more economical to meet periods of high demand by producing in prior periods of low demand and holding inventory, defining the basic trade-off in the problem. The authors develop an optimal solution based on calculus that essentially identifies planning horizons, allowing the problem to be decomposed along the time horizon into subproblems consisting of a certain number of consecutive periods that can be solved independently. This approach forms the basis for their later work (Holt et al. 1955, 1956), which subsequently led to the Holt, Modigliani, Muth and Simon (HMMS) book (Holt et al. 1960). In Chap. 6 of their book, they explicitly address the extension of their decision rules to an environment with uncertain demand and show that under the specific quadratic objective function they assume, the deterministic equivalent of the stochastic problem is achieved by using expected demand values in their deterministic rule, which is equivalent to assuming an unbiased demand forecasting procedure. This insight appears to have contributed to the heavy focus on deterministic models in the production domain, although the proof they provide is only valid for the specific case of a quadratic objective function. An interesting discussion of this body of work is given by Singhal and Singhal (2007).

It is interesting that capacity constraints are not modeled; the implicit assumption appears to be that capacity can be varied in the short run, and the costs of doing this are captured by the increasing marginal cost of production. This discussion is made more explicit in the context of labor costs by Charnes et al. (1955). It is also interesting that while the cost function is explicitly built up from holding, production and fixed costs that are independent of production volume, there is no discussion of how one might actually estimate these costs from existing business records. Finally, the basic paradigm is that of modeling the physical flows in the problem—production and inventories—and assigning costs to these, rather than modeling the cash flows explicitly as a means of capturing the financial impact. The Modigliani and Hohn (1955) paper also seems to have motivated the idea that in problems over time, perfect information for the entire planning horizon is not necessary, but rather just planning the first few periods on a rolling horizon basis is quite close to optimality. This has led to a long stream of papers using these and related ideas, including the well-known dynamic lot-sizing model of Wagner and Whitin (1958).

In the mid-1950s, researchers realized that the models of Modigliani and Hohn (1955) and Holt et al. (1956) could be formulated as linear programs. The two principal papers that appear to have accomplished this independently of each other are Hanssmann and Hess (1960), whose title very much resonates with the HMMS work, and Manne (1957). Another notable early paper is that of Bowman (1956), which appears to be one of the earliest to identify the extensive network structure present in production planning models.

The principal characteristics of the mathematical programming models used for most production planning problems were now in place. The decisions cover a planning horizon that is divided into discrete time periods, each of which has an associated set of decision variables reflecting the decisions made in that period. The decision variables represent the physical flows of materials through the different production resources and inventory points; the objective function is generally that of minimizing the variable costs of production, inventories, and backlogs over the planning horizon, while satisfying aggregate capacity constraints on the production resources in each period. In Chap. 5 we shall investigate the basic assumptions of these models in more detail, focusing on how they model the dynamics of capacitated production resources. As noted earlier, the MRP procedure for exploding a bill of materials and computing time-phased releases can be formulated as a mixed-integer linear program (Voss and Woodruff 2006). In order to overcome the well-known limitations of MRP, Billington et al. (1983) introduced the multilevel capacitated lot-sizing problem (MLCLSP) that performs capacitated lot sizing for a multistage production system and all SKUs simultaneously, integrating lot sizing into mathematical programming models for production planning.

Most deterministic production planning models establish optimal production, inventory, and release levels over a finite planning horizon to meet the total demand (Holt et al. 1960; Buffa and Taubert 1972; Hax and Candea 1984). The planning horizon is divided into discrete periods during which production and demand rates are assumed to be constant; the capacity of the system is represented by the number of hours available on key resources in a planning period; and the production, inventory, WIP, and demand associated with a period are treated as continuous quantities. These models allocate capacity to products to minimize a specified objective and satisfy aggregate constraints representing system capacity and dynamics. The need to match output to demand requires some estimate of the delay between work being released into the production units and its emergence as finished products that can be used to meet demand.

The most common approach to handle this problem in both the research literature and industrial practice is to use planned lead times that are fixed, exogenous quantities independent of resource load, as long as a maximum capacity loading is not exceeded. As we have seen, the Material Requirements Planning (MRP) approach (Orlicky 1975) uses fixed planned lead times in its backward scheduling step to determine job releases. Several authors have suggested ways of adapting MRP to uncertain demand. Meal (1979) and Grubbstrom (1998) derive component plans with safety stocks in the MRP records. Miller (1979) proposes hedging of the master schedule to provide safety stocks within the system. However, all these assume fixed exogenous lead times.

Another common approach to production planning under fixed lead times and deterministic demand is the use of linear (LP) and integer programming (IP) models, of which a wide variety exist (Johnson and Montgomery 1974; Hackman and Leachman 1989; Voss and Woodruff 2003). These represent capacity as a fixed upper bound on the number of hours available at the resource in a period, and model input and output time lags between stages as well as planned lead times for different production units. However, these time lags are generally independent of workload, although these models can accommodate a wide range of specific technological and managerial constraints specific to particular production environments. The basic structure of these models is described in detail in Chap. 5.

## References

Ankenman BE, Bekki JM, Fowler J, Mackulak GT, Nelson BL, Yang F (2010) Simulation in production planning: an overview with emphasis in recent developments in cycle time estimation. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise: a state of the art handbook, vol 1. Springer, New York, pp 565–592

Arrow KJ, Karlin S, Scarf H (1958) Studies in the mathematical theory of inventory and production. Stanford University Press, Stanford, CA

Aurand S, Miller P (1997) The operating curve: a method to measure and benchmark manufacturing line productivity. In: IEEE/SEMI Advanced Semiconductor Manufacturing Conference, pp 391–397

Bechte W (1980) Steuerung der Durchlaufzeit durch Belastungsorientierte Auftragsfreigabe bei Werkstattfertigung. TU Hanover, Hanover

Bechte W (1988) Theory and practice of load-oriented manufacturing control. Int J Prod Res 26(3):375–395

Bergamaschi D, Cigolini R, Perona M, Portoli A (1997) Order review and release strategies in a job shop environment: a review and a classification. Int J Prod Res 35:399

Bertrand JWM, Van Ooijen HPG (2002) Workload based order release and productivity: a missing link. Prod Plan Control 13(7):665–678

Bertrand JWM, Van Ooijen HPG (2008) Optimal work order release for make-to-order job shops with customer order lead-time costs, tardiness costs and work-in-process costs. Int J Prod Econ 116(2):233–241

Bertrand JWM, Wortmann JC (1981) Production control and information systems for component-manufacturing shops. Elsevier, Amsterdam, Oxford, New York

Bertrand JWM, Wortmann JC, Wijngaard J (1990) Production control: a structural and design oriented approach. Elsevier, Amsterdam

Billington PJ, Mcclain JO, Thomas JL (1983) Mathematical programming approaches to capacity-constrained MRP systems: review, formulation and problem reduction. Manag Sci 29:1126–1141

Bobrowski PM (1989) Implementing a loading heuristic in a discrete release job shop. Int J Prod Res 27(11):1935–1948

Bowman EB (1956) Production scheduling by the transportation method of linear programming. Oper Res 4(1):100–103

Buffa ES, Taubert WH (1972) Production-inventory systems; Planning and control. R. D. Irwin, Homewood, IL

Charnes A, Cooper WW, Mellon B (1955) A model for optimizing production by reference to cost surrogates. Econometrica 23(3):307–323

Cohen O (1988) The Drum-Buffer-Rope (DBR) approach to logistics. In: Rolstadas A (ed) Computer-aided production management. Springer, New York

de Kok AG, Fransoo JC (2003) Planning supply chain operations: definition and comparison of planning concepts. In: de Kok AG, Graves SC (eds) Handbooks in operations research and management science. Supply chain management: design, coordination and operation, vol 11. Elsevier, Amsterdam, pp 597–675

Fredendall LD, Ojha D, Patterson W (2010) Concerning the theory of workload control. Eur J Oper Res 201:99–111

Goldratt E, Fox RE (1986) The race. North River Press, Croton-on-Hudson, NY

Grubbstrom RW (1980) A principle for determining the correct capital costs of work-in-progress and inventory. Int J Prod Res 18(2):259–271

Grubbstrom RW (1998) A net present value approach to safety stocks in planned production. Int J Prod Econ 56-57:213–229

Gupta M (2005) Constraints management—recent advances and practices. Int J Prod Res 41(4):647–659

Hackman ST, Leachman RC (1989) A general framework for modeling production. Manag Sci 35(4):478–495

Hanssmann F, Hess SW (1960) A linear programming approach to production and employment scheduling. Manag Technol 1(1):46–51

Harris FW (1915) Operations and cost. Factory management series. A. W. Shaw, Chicago

Hax AC, Candea D (1984) Production and inventory management. Prentice-Hall, Englewood Cliffs, NJ

Hendry L, Kingsman BG (1991a) Job release: part of a hierarchical system to manage manufacturing lead times in make-to-order companies. J Oper Res Soc 42(10):871–883

Hendry LC, Kingsman BG (1991b) A decision support system for job release in make to order companies. Int J Oper Prod Manag 11:6–16

Hendry LC, Land M, Gaalman GJC (2008) Investigating implementation issues for workload control (WLC): a comparative case study analysis. Int J Prod Econ 112(1):452–469

Hendry L, Huang Y, Stevenson M (2013) Workload control: successful implementation taking a contingency-based view of production planning and control. Int J Oper Prod Manag 33(1):69–103

Henrich P, Land M, Gaalman G (2004) Exploring applicability of the workload control concept. Int J Prod Econ 90(2):187–198

Heragu S, Srinivasan M (2011) Analysis of manufacturing systems via single-class, semi-open queuing networks. Int J Prod Res 49(2):295–319

Holt CC, Modigliani F, Simon HA (1955) A linear decision rule for production and employment scheduling. Manag Sci 2(1):1–30

Holt CC, Modigliani F, Muth JF (1956) Derivation of a linear rule for production and employment. Manag Sci 2(2):159–177

Holt CC, Modigliani F, Muth JF, Simon HA (1960) Planning production, inventories and work force. Prentice Hall, Englewood Cliffs, NJ

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Hutter T, Häussler S, Missbauer H (2018) Successful implementation of an order release mechanism based on workload control: a case study of a make-to-stock manufacturer. Int J Prod Res 56(4):1565–1580

Irastorza JC, Deane RH (1974) A loading and balancing methodology for job shop control. AIIE Trans 6(4):302–307

Jackson JR (1955) Scheduling a production line to minimize maximum tardiness. University of California, Los Angeles, Los Angeles

Jansen M (2012) Anticipation in supply chain operations planning. In: Beta research school for operations management and logistics. Eindhoven University of Technology, Eindhoven

Jia J, Heragu S (2009) Solving semi-open queuing networks. Oper Res 57(2):392–401

Johnson LA, Montgomery DC (1974) Operations research in production planning, scheduling and inventory control. Wiley, New York

Kanet JJ (1988) Load-limited order release in job shop scheduling systems. J Oper Manag 7:413–422

Kayton D, Teyner T, Schwartz C, Uzsoy R (1997) Focusing maintenance improvement efforts in a wafer fabrication facility operating under theory of constraints. Prod Invent Manag 38(Fourth Quarter):51–57

Kingsman BG (2000) Modelling input-output workload control for dynamic capacity planning in production planning systems. Int J Prod Econ 68:73–93

Kingsman BG, Mercer A (1997) Strike rate matrices for integrating marketing and production during the tendering process in make-to-order subcontractors. Int J Oper Res 4(1):251–257

Koopmans T (ed) (1951) Activity analysis of production and allocation. Wiley, New York

Krajewski L, Ritzman LP, Malhotra M (2013) Operations management. Pearson, Upper Saddle River, NJ

Kushner HJ (2001) Heavy traffic analysis of controlled queueing and communication networks. Springer, New York

Land M (2004) Workload control in job shops, grasping the tap. Labyrinth Publications, Richmond, CA

Land MJ, Gaalman GJC (1998) The performance of workload control concepts in job shops: improving the release method. Int J Prod Econ 56:347–364

Leontief WW (1951) The structure of the american economy. Oxford University Press, Oxford

Liker J (2004) The Toyota way: 14 management principles from the world's greatest manufacturer. McGraw-Hill, New York

Lu S, Ramaswamy D, Kumar PR (1994) Efficient scheduling policies to reduce mean and variance of cycle time in semiconductor plants. IEEE Trans Semicond Manuf 7:374–388

Manne AS (1957) A note on the Modigliani-Hohn production smoothing model. Manag Sci 3(4):371–379

Manne AS (1960) On the job-shop scheduling problem. Oper Res 8(2):219–223

McKay KN (2010a) Field-based research on production control. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise: a state of the art handbook, vol 1. Springer, New York, pp 205–232

McKay KN (2010b) The historical foundations of manufacturing planning and control practices. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise, vol 12. Springer, New York, pp 21–32

Meal H (1979) Safety stocks in MRP systems. Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA

Melnyk S, Ragatz GL (1989) Order review/release: research issues and perspectives. Int J Prod Res 27(7):1081–1096

Miller JG (1979) Hedging the master schedule. In: Ritzman LP (ed) Dissagregation problems in manufacturing and service organizations. Martinus Nijhoff, Boston, MA

Missbauer H (1998) Bestandsregelung als Basis für eine Neugestaltung von PPS-Systemen. Physica, Heidelberg

Missbauer H (2009) Models of the transient behaviour of production units to optimize the aggregate material flow. Int J Prod Econ 118(2):387–397

Modigliani F, Hohn FE (1955) Production planning over time and the nature of the expectation and planning horizon. Econometrica 23(1):46–66

Moench L, Fowler JW, Mason SJ (2013) Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis and systems. Springer, Berlin

Nyhuis P, Wiendahl HP (2009) Fundamentals of production logistics: theory, tools and applications. Springer, Berlin

Oosterman B, Land MJ, Gaalman GJC (2000) The influence of shop characteristics on workload control. Int J Prod Econ 68:107–119

Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York

Pande P, Neuman RP, Cavanagh RR (2000) The six sigma way: how GE, Motorola and other top companies are honing their performance. McGraw-Hill, New York

Perona M, Portioli A (1998) The impact of parameters setting in load oriented manufacturing control. Int J Prod Econ 55:133–142

Ragatz GL, Mabert VA (1988) An evaluation of order release mechanisms in a job shop environment. Decis Sci 19:167–189

Silva C, Stevenson M, Thürer M (2015) A case study of the successful implementation of workload control. J Manuf Technol Manag 26(2):280–296

Singhal J, Singhal K (2007) Holt, Modigliani, Muth and Simon's work and its role in the renaissance and and evolution of operations management. J Oper Manag 25:300–309

Soepenberg E, Land M, Gaalman GJC (2012) Workload control dynamics in practice. Int J Prod Res 50(2):443–460

Spearman ML, Zazanis M (1992) Push and pull production systems: issues and comparisons. Oper Res 40(3):521–532

Spearman ML, Hopp WJ, Woodruff DL (1989) A hierarchical control architecture for constant work-in-process (CONWIP) production systems. J Manuf Oper Manag 2(3):147–171

Spearman ML, Woodruff DL, Hopp W (1990) CONWIP: a pull alternative to Kanban. Int J Prod Res 28(5):879–894

Stadtler H, Kilger C, Meyr H (2015) Supply chain management and advanced planning. Concepts, models, software, and case studies. Springer Verlag, Berlin

Stevenson M, Hendry LC (2006) Aggregate load-oriented workload control: a review and a re-classification of a key approach. Int J Prod Econ 104(2):676–693

Stevenson M, Silva C (2008) Theoretical development of a workload control methodology: evidence from two case studies. Int J Prod Res 46(11):3107–3131

Stevenson M, Huang Y, Hendry LC, Soepenberg E (2011) The theory and practice of workload control: a research agenda and implementation strategy. Int J Prod Econ 131(2):689–700

Sugimori Y, Kusunoki K, Cho F, Uchikawa S (1977) Toyota production system and Kanban system: materialization of just-in-time and respect for human system. Int J Prod Res 15(6):553–564

Tatsiopoulos IP, Kingsman BP (1983) Lead time management. Eur J Oper Res 14:351–358

Thürer M, Silva C, Stevenson M (2010) Workload control release mechanisms: from practice back to theory building. Int J Prod Res 48(12):3593–3617

Thürer M, Stevenson M, Silva C (2011) Three decades of workload control research: a systematic review of the literature. Int J Prod Res 49(23):6905–6935

Thürer M, Silva C, Stevenson M, Land M (2012) Improving the applicability of workload control (WLC): the influence of sequence dependent setup times on workload controlled job shops. Int J Prod Res 50(22):6419–6430

Thürer M, Stevenson M, Silva C, Land M, Fredendall LD, Melnyk SA (2014) Lean control for make-to-order companies: integrating customer enquiry management and order release. Prod Oper Manag 23(3):463–476

Van Ooijen HPG (1998) Delivery performance improvement by controlled work-order release and work-center load balancing. Int J Prod Econ 56-57:661–675

Vollmann TE, Berry WL, Whybark DC, Jacobs FR (2005) Manufacturing planning and control for supply chain management. McGraw-Hill, New York

Voss S, Woodruff DL (2003) Introduction to computational optimization models for production planning in a supply chain. Springer, Berlin, New York

Voss S, Woodruff DL (2006) Introduction to computational optimization models for production planning in a supply chain. Springer, New York

Wagner HM, Whitin TM (1958) Dynamic version of the economic lot size model. Manag Sci 5(1):89–96

Whitin TM (1954) Erich Schneider's inventory control analysis. J Oper Res Soc Am 2(3):329–334

Wiendahl HP (1992) Anwendung der belastungsorientierten Fertigungsteuerung. Munich, Hanser-Verlag

Wiendahl HP (1995) Load oriented manufacturing control. Springer, Heidelberg

Womack JP, Jones DT, Ross D (1990) The machine that changed the world. Harper Perennial, New York

Yan H, Stevenson M, Hendry LC, Land MJ (2016) Load-oriented order release (LOOR) revisited: bringing it back to the state of the art. Prod Plan Control 27:1078–1091

Yang F, Ankenman B, Nelson BL (2006) Efficient generation of cycle time-throughput curves through simulation and metamodeling. Naval Res Logistics 54(1):78–93

Zäpfel G, Missbauer H (1993a) Production planning and control (PPC) systems including load-oriented order release—problems and research perspectives. Int J Prod Econ 30:107–122

Zäpfel G, Missbauer H (1993b) New concepts for production planning and control. Eur J Oper Res 67:297–320

Zäpfel G, Missbauer H, Kappel W (1992) PPS-Systeme mit belastungsorientierter Auftragsfreigabe—Operationscharakteristika und Möglichkeiten zur Weiterentwicklung. Zeitschrift für Betriebswirtschaft 62:897–919

# Chapter 5
# Planning Models with Stationary Fixed Lead Times

In this chapter, we present optimization models for order release using exogenous planned lead times that remain constant (stationary) over the planning horizon. We describe the material flow dynamics implied by these models, beginning by assuming lead times that are integer multiples of the underlying planning period. We construct a series of linear programming models for this problem and examine their dual, noting several implications that are inconsistent with insights from the queueing models discussed in Chap. 2. We then extend this approach to consider fractional lead times and a more general formulation where a production order may consume capacity in multiple, not necessarily consecutive, periods.

## 5.1 Preliminaries

The previous chapters have, we hope, set the stage upon which we propose to address the principal topic of this volume: the development of novel, and hopefully more effective, optimization models to support the goods flow problem faced by the planning level, whose purpose is to coordinate the releases of work across multiple production units in the production system or supply chain to meet demand in the best possible manner. Due to the need to match production with demand, the models must take into account the cycle times, the delay between work being released into the production unit and its emergence as completed product that can meet demand.

We shall refer to the smallest unit of work recognized by the goods flow problem as an *order*. Orders may be of external or internal origin; *external* orders represent a specific quantity of a specific product ordered by a specific customer, while *internal* orders are generated by the PPC system for purposes of production management within the production unit, and thus may represent material destined for several customers, a portion of a larger customer order, or simply material intended to

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_5

replenish inventory positions along the supply chain. For the purposes of the goods flow problem, both can be treated in the same manner, so we will use the term "order" for both.

Following the discussion in Chap. 1, a production unit is an organizational unit whose internal operations are not under the control of the planning level, which is tasked with managing the goods flow problem. A production unit consists of several workcenters with limited capacity, through which each order processed in the production unit follows a specified routing. For exposition we assume the routing to be deterministic, ignoring the possibility of random routing due to causes such as alternative resources or rework. While this is certainly not the most general model that could be presented, it is sufficient to convey the essence of the problems we consider. Hackman and Leachman (1989a, b) and Hackman (1990, 2008) provide a much more general treatment encompassing other modes of production such as resource-constrained project scheduling. Per Chap. 2, the cycle time of a unit of work is a random variable that follows some probability distribution, but can only be observed after the fact. We shall use the term *lead time* to denote an estimate of the cycle time used in planning models for the goods flow problem. The focus of this chapter is on planning models that use constant, exogenous lead times to represent the progress of orders through the production unit. For brevity of exposition, we shall refer to these lead times as *fixed lead times*. In this chapter, we consider the simpler case where the planned lead times associated with a production unit and its workcenters remain constant over time, i.e., do not vary across time periods. The more complex case of time-varying planned lead times is treated in the next chapter.

## 5.2 A Generic Production Unit

Figure 5.1 illustrates a generic production unit that produces a set $J$ of products $j = 1, \ldots, |J|$, for which it has a queue of orders waiting to be processed that have been released by the planning level, and a finished goods inventory location where finished items are stored. The production process uses a set $K$ of different workcenters $k = 1, \ldots, |K|$, with limited capacity, each of which, per Chap. 2, can be modeled as a queueing system. We denote the set of workcenters used by product $j$ by $K(j)$ and the time required to process a unit of product $j$ on workcenter $k$ as $a_{jk}$. The set of products requiring a workcenter $k$ will be denoted by $J(k)$. The planning horizon is divided into discrete time periods, which we shall assume without loss of generality to be of equal length $\Delta$, such that period $t$ ends at time $t\Delta$. When it causes no ambiguity, we shall assume the time periods to be of unit length so that period $t$ ends at time $t$. The basic sequence of events taking place in the production unit is as follows:

1. The planning level authorizes the release of an order consisting of a specific quantity $R_{jt}$ of product $j$ to the production unit at time $t$.
2. The order is released for production and enters the queue for the first workcenter in its routing. Control over its progress through the production unit is transferred to the internal management of the production unit. Upon completion of its

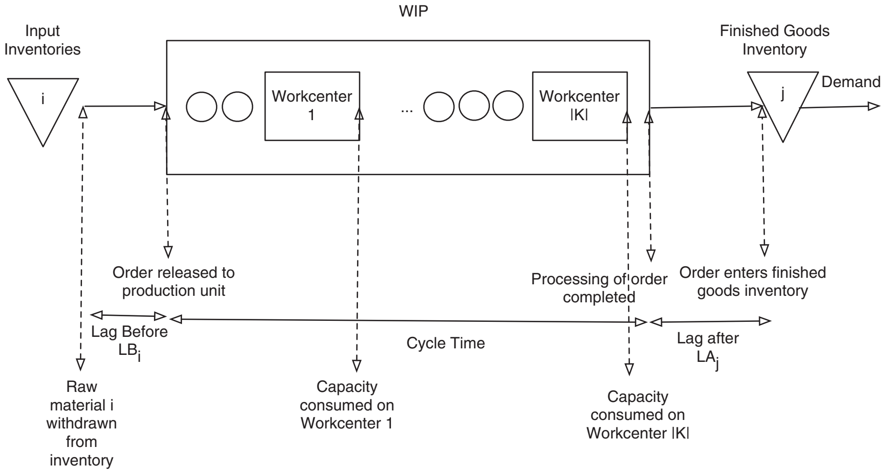

WIP
Input Inventories
i
Workcenter 1
...
Workcenter |K|
Finished Goods Inventory
j
Demand
Order released to production unit
Processing of order completed
Order enters finished goods inventory
Lag Before LB$_i$
Cycle Time
Lag after LA$_j$
Raw material i withdrawn from inventory
Capacity consumed on Workcenter 1
Capacity consumed on Workcenter |K|

**Fig. 5.1** Generic Production Unit with Time Lags

processing at each workcenter, the order moves directly to the next workcenter in its routing.

3. The order completes processing on the last workcenter in its routing and moves from the production facility to the production unit's finished inventory location.

Each order of product $j$ released in period $t$ will wait in the queue for workcenter $k$ for an average of $Q_{jkt}$ time units and will require an expected processing time of $P_{jkt} = R_{jt}a_{jk}$ time units, which we assume includes any necessary setup times. The expected cycle time of the order from its release to its entry into finished inventory is thus given by

$$T_{jkt} = \sum_{k \in K(j)} \left( Q_{jkt} + P_{jkt} \right) \tag{5.1}$$

The cycle time of an order at a workcenter $k$ is thus the sum of its processing time and its queue time. Per Chap. 2, the queue time is a random variable whose probability distribution depends on the utilization of the workcenter, which is determined by the work release decisions $R_{jt}$, while the service time is also a random variable. These random variables are represented in (5.1) by their expectations. The expected cycle time of an order is thus given by the sum of its expected processing and queue times at each workcenter $k$ in its routing. In practice, additional delays may be incurred, such as transportation time between workstations, preparation of components and raw materials, or transfer of the finished order to finished inventory, which are also likely to be random variables. A wide variety of domain-specific events may need to be considered, such as the need to allow a specified time for lumber to cure before its use in furniture manufacturing or the need to perform a thin-film deposition step within a specified time of a cleaning step in semiconductor manufacturing. The modeling of fixed delays between such events is discussed at length by Hackman and Leachman (1989b) and Hackman (2008). However, the events shown in Fig. 5.1 are sufficient to account for most cases of interest.

## 5.3 Lead Times in Models of the Goods Flow Problem

The wide range of planning models using fixed exogenous lead times, including both MRP (Orlicky 1975; Baker 1993; Vollmann et al. 2005) and mathematical programming models (Voss and Woodruff 2006; Missbauer and Uzsoy 2011), all assume that as long as all constraints in the model are satisfied, the production unit will be able to produce its output in a manner consistent with the lead time values specified. Thus the lead times serve the planning level as an anticipation function (Schneeweiss 2003) describing the impact of its release decisions on the output of the production units. We view a lead time $L_{jk}$ as a parameter whose value is an estimate of a suitably high percentile of the order cycle time distribution whose mean $T_{jkt}$ is given by (5.1). Hence under normal conditions any order released to the production unit will enter finished goods inventory within $L_j = \sum_k L_{jk}$ time units of its release with high probability. Under this view the lead time is treated as a delay between the release of an order into the production unit and its completion.

Billington et al. (1983) suggest using only the minimum time required to transfer material between operations without considering queue time or processing time; they argue that delays due to limited capacity will be computed by the planning model itself, which should produce materials ahead of time and hold it in finished inventory until needed to meet demand, ignoring the workload-dependent nature of the queue time $Q_{jkt}$. These transfer times between operations can be modeled as fixed delays following Hackman and Leachman (1989b) if their duration is significant relative to that of the planning period. Another class of planning models treats the fixed lead time not as a delay, but as a time interval within which the production unit must process the order once it is released. We shall first discuss models that treat lead times as delays and treat this latter view in Sect. 5.6.

### 5.3.1 *Planning Models with Fixed Exogenous Lead Times*

The vast majority of the mathematical programming models of interest to this volume approach the goods flow problem faced by the planning level following the early formulations of Modigliani and Hohn (1955), Manne (1957), Hanssmann and Hess (1960), and Holt et al. (1955). A finite time horizon is divided into discrete time periods, usually, but not necessarily, of the same length. Decision variables are associated with each period, and the objective is either to minimize total cost or to maximize total contribution (revenue minus variable costs) over the planning horizon. All quantities are treated as deterministic. Following Hackman and Leachman (1989b), such models require three basic sets of constraints:

1. *Inventory or material balance constraints* for all input and finished inventory points, which coordinate material flows through both space and time. These also enforce the satisfaction of demand, which is treated as a material flow out of the production system to an external demand source.

2. *Capacity constraints*, which model how the production activities capture and consume production resources.
3. *Domain-specific constraints* reflecting the special structure and requirements of the particular production environment being modeled. The structure of these constraints will differ widely based on the specific environment under study and hence will not be discussed in detail. We shall focus on the first two constraint sets, which are critical to the model's ability to accurately reflect the realized behavior of the production system for which the plans are developed.

Two points in time are of particular interest: the point at which the production order actually consumes capacity on the resources required to process it and the time it is completed and can be used to meet demand. Knowledge of the former is necessary to ensure that capacity constraints are not violated over time and of the latter to allow accurate prediction of the amount of material available to meet demand over time.

## *5.3.2 A Single Production Unit*

We begin by considering a production unit modeled as in Fig. 5.1. Since the timing and quantity of order releases constitute the link between planning and detailed scheduling within the production unit, release quantities are the primary decision variables of interest. We shall assume all demand must be met without backlogging; this will allow us to focus on representing the behavior of the production unit. Thus negative inventory levels are not permitted at any inventory location. Material flows within the production unit itself are of interest only to ensure that releases are capacity feasible for all workcenters $k \in K$, and hence the production unit can meet demand within the specified lead times.

### 5.3.2.1 Single Product, Instantaneous Production, Unlimited Inputs

The simplest model of production, encountered in classical inventory models such as the Economic Order Quantity model and the Wagner-Whitin dynamic lot-sizing model (Zipkin 2000; Hopp and Spearman 2008), is instantaneous production where the quantity ordered at a given point in time becomes available immediately upon production being initiated. In mathematical programming models, this implies that cycle time is negligible relative to the length of the planning period, so that the entire quantity $R_t$ of material released into the system during period $t$ is available to meet demand by the end of that period. The assumption of unlimited inputs implies either instantaneous acquisition or sufficient on-hand inventory of all inputs. Thus inputs will never constrain the ability of the production unit to meet demand, and there is no need to model input inventories. Since we have only a single product, the product subscript $j$ is suppressed.

To ensure consistent material flows over time, we model the finished goods inventory level across periods with the material balance equations

$$I_t = I_{t-1} + X_t - D_t, \quad t = 1,\ldots,T \tag{5.2}$$

where $I_t$ denotes the amount of finished goods on hand at the end of period $t$, $X_t$ the output of the production unit in period $t$, and $D_t$ the demand during that period. Under instantaneous production, we have $R_t = X_t$; all materials released into the production unit in a period are converted into output by the end of the period. Denoting the amount of finished goods inventory at the start of the first period (the end of period 0) by $I_0$, (5.2) can be rewritten as

$$I_t = I_0 + \sum_{\tau=1}^{t} X_\tau - \sum_{\tau=1}^{t} D_\tau, \quad t = 1,\ldots,T \tag{5.3}$$

by summing the constraints (5.2) for consecutive periods 1, …, $t$.

The most common capacity constraint encountered in the literature seeks to ensure that the total production $X_t$ for a given period $t$, and hence the planned releases $R_t$, cannot exceed the available capacity $C_{kt}$ of any workcenter $k$. Since we produce a single product, $C_{kt}$ can be expressed in units of the end item, allowing this constraint to be written as:

$$R_t \le C_{kt}, \quad t = 1,\ldots,T; \quad k = 1,\ldots,K \tag{5.4}$$

Taken together, (5.2) and (5.4) imply that as long as releases do not violate capacity constraints on any workcenter, materials released in period $t$ will be available to meet demand by the end of the same period. If demand $D_t$ in any period $t$ exceeds the capacity of some workcenter $k$, the only course open to the model is to produce the excess demand in an earlier period $s < t$, holding finished inventory in the periods $s$ to $t$. Combining (5.3) and (5.4) yields

$$\sum_{\tau=1}^{t} C_{k\tau} \ge \sum_{\tau=1}^{t} X_\tau = \sum_{\tau=1}^{t} R_\tau \ge \sum_{\tau=1}^{t} D_\tau - I_0, \quad t = 1,\ldots,T, \quad k \in K, \tag{5.5}$$

as a necessary condition for a feasible solution to exist. The only reason to release an order in advance of the period in which it is due is lack of capacity at some workcenter $k$ in that period. Denoting the unit cost of holding FGI for one period by $h_t$ and the unit incremental cost of production by $c_t$, the planning model can be written as:

$$\min \sum_{t=1}^{T} \left( h_t I_t + c_t R_t \right) \tag{5.6}$$

subject to

$$I_t = I_{t-1} + R_t - D_t, \quad t = 1,\ldots,T \tag{5.7}$$

$$R_t \le C_{kt}, \quad t = 1,\ldots,T, \quad k \in K \tag{5.8}$$

$$I_t, R_t \geq 0, \quad t = 1,\ldots,T \tag{5.9}$$

This model, although simplistic in its assumptions, has all the basic components of a production planning model: decision variables associated with each period (the $R_t$), state variables arising from the decision variables and the constraints (the $I_t$), an objective function minimizing the sum of production and inventory holding costs (5.6), material balance constraints (5.7) for the finished inventory location, and capacity constraints (5.8) for each resource $k$ in each period $t$.

The capacity constraint (5.8) ensures that the total planned resource usage during the planning period does not exceed the amount of the resource available during the period. This is necessary, but not sufficient, to ensure that the planned releases can actually be processed within the planning period, since the model does not control the timing of work arrivals at the workcenter within the period. If for some reason such as a machine failure on the shop floor, 75% of the amount released became available only in the second half of the planning period, the workcenter might well not be able to process all of it by the end of the period.

#### 5.3.2.2 Single Product, Non-instantaneous Production

The model (5.6)–(5.9) is not realistic when the magnitude of the workcenter cycle times $Q_{jkt} + P_{jkt}$ is significant relative to that of the planning period. The most common representation of this situation in the literature is a fixed lead time $L$ representing the estimated time required for work released in a given period to become available to meet demand, most commonly expressed as an integer number of planning periods.

Under these assumptions, material released into the production unit during period $t$ becomes available for use $L$ time periods later during period $t + L$, implying that $X_t = R_{t-L}$. The material balance constraints for the finished inventory are now

$$I_t = I_{t-1} + X_t - D_t = I_{t-1} + R_{t-L} - D_t, \quad t = 1,\ldots,T \tag{5.10}$$

This is exactly the model of lead times used in MRP in its backward scheduling phase, where the fixed lead time represents the amount of time elapsing between the time an order for a BOM item is placed and its receipt (Baker 1993; Voss and Woodruff 2003). Since we have only one product (end item), the product index $j$ remains suppressed.

Under instantaneous production, an order consumes capacity at each resource $k$ in the period in which it is released, rendering constraints (5.8) sufficient to ensure capacity feasible releases. However, when lead times exceed one period a question of timing arises—at what point in the lead time $L$ does the job consume capacity on a given resource $k$? This requires knowledge of the process routing, the sequence in which the different resources are utilized by the order. Without loss of generality, we shall assume that the order visits each resource exactly once in a known, deterministic sequence and that the resources are indexed in the order of their use. Thus resource

$k = 1$ is the first resource used in the routing, and resource $k = |K|$ the last one before the order enters finished inventory. Let $L_k$ denote the estimated delay between the release of the order to the production unit and its becoming available for processing on workcenter $k$. Thus $L_k$ represents an estimate of the total cycle time of the order at all workcenters in its routing prior to $k$, implying that

$$E\left[Q_{kt} + P_{kt}\right] = L_k - L_{k-1} \tag{5.11}$$

Clearly we must have

$$\max_{1 \le k \le |K|} \left\{L_k\right\} \le L \tag{5.12}$$

for consistency. Our capacity constraints (5.8) now take the form

$$R_{t-L_k} \le C_{kt}, \quad \text{for all} \quad k \in K; \quad t = 1,\ldots,T \tag{5.13}$$

Since no inventory is held within the production unit other than the WIP waiting for processing or in transit between stages, the output of individual workcenters is represented to capture their incremental costs of production and their limited capacity in each period. By the definition of the lead times $L_k$, an order processed on workcenter $k$ in period $t$ will have been released in period $t - L_k$. For simplicity of exposition, we shall assume that the total production cost of an order completed in period $t$, given by

$$c_t = \sum_{k=1}^{|K|} c_{k.t-\left(L-L_k\right)} \tag{5.14}$$

where $c_{kt}$ denotes the unit cost of production on workcenter $k$ in period $t$, is assessed in period $t$; this could easily be relaxed at the expense of additional notation. The single-product multiple workcenter model with integer lead times $L_k$ associated with each resource $k$, and an overall lead time $L$ associated with the entire production unit, is as follows:

$$\min \sum_{t=1}^{T} \left(h_t I_t + c_t R_{t-L}\right) \tag{5.15}$$

subject to

$$I_t = I_{t-1} + R_{t-L} - D_t, \quad t = 1,\ldots,T \tag{5.16}$$

$$R_{t-L_k} \le C_k, \quad t = 1,\ldots,T, \quad k \in K \tag{5.17}$$

$$I_t,\ R_t \ge 0, \quad t = 1,\ldots,T \tag{5.18}$$

Decision variables with non-positive subscripts correspond to decisions made prior to the start of the planning horizon that are known with certainty, and as such are parameters of the model. This is essentially the step-separated formulation of

Leachman and Carmon (1992), without the alternative production routings considered in that paper. The amount of production that can take place on resource $k$ in a given period $t$ is limited by both the capacity $C_{kt}$ and the amount of work available for processing, given by past releases $R_{t-L_k}$ per (5.17). Hence the amount of WIP available to process on workcenter $k$ in period $t$ is simply $R_{t-L_k}$. The total amount of WIP in the production unit—the amount of material that has been released but not yet completed—is given by

$$W_t = \sum_{\tau=t-L+1}^{t} R_\tau = \sum_{\tau=t+1}^{t+L} X_\tau \tag{5.19}$$

This quantity does not appear in LP models of production planning, such as those discussed in Johnson and Montgomery (1974), Hackman and Leachman (1989a, b), and Voss and Woodruff (2006) that treat fixed lead times as a delay between order release and completion. The reason for this is apparent from (5.19): when a fixed lead time represents a delay the amount of WIP is determined by the lead time $L$ and the releases $R_t$; any WIP holding cost can be incorporated into the incremental production cost $c_t$.

The movement of material through a system with four machines in series under this model is traced in Fig. 5.2. The vertical axis shows the lead times for each machine, and each timeline the material processed by each machine in each period, identified by the period of its release to the first machine, machine 1. The material released in each period is indicated by the numeral above it; thus, material released at the start of period 1 is indicated by a "1" above the line indicating the material. Material released in a given period is shown with a bar of a given color until it exits the system; thus the material released in period 1 is shown as a red bar as it proceeds through the machines. Material released at the start of the planning horizon, at the start of period 1, indicated by the red bars, becomes available to machine 2 at the start of period 2, is in WIP at machine 2 at the start of period 3, is available to

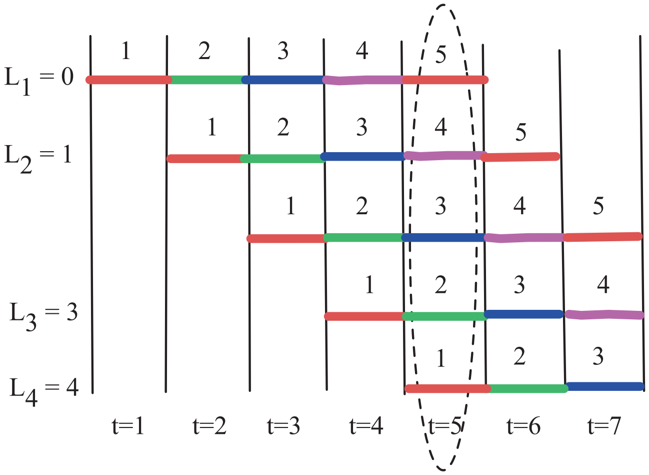


Fig. 5.2 Timing of material flow under fixed lead times

machine 3 at the start of period 4, and is available to machine 4 at the start of period 4. At the end of period 5 or, equivalently, the start of period 6, the WIP at machine 4 consists of the material entering the system (i.e., released to machine 1) in period 2; at machine 3, the material released in periods 3 and 4, and at machine 2, that released in period 5. The figure also indicates that not all the WIP at a machine at the start of a period is necessarily available to be processed at the machine in the period. For example, at the start of period 6, material entering the system in periods 3 and 4 is in WIP at machine 4, but even if the machine has sufficient capacity, only the material entering in period 3 will be processed. In other words, WIP cannot accumulate, but flows through the system in discrete units equal to the quantity released in each period.

The model assumes that WIP will not accumulate in the system over time; only the material released in period $t - L_k$ is available to resource $k$ for processing in period $t$. Equivalently, all $R_t$ units of product released in period $t$ are assumed to move through the production process as a single entity, occupying capacity on each workcenter within a single period. Since (5.17) ensures that releases do not exceed capacity, the system can always process this quantity in a single period. The remaining WIP still to be processed by the workcenter, given by

$$\tilde{W}_{kt} = \sum_{\tau=t-L_k+1}^{t-(L_k-L_{k-1})+1} R_\tau \tag{5.20}$$

has no effect on the cycle time of the workcenter, which can never exceed $L_k$–$L_{k-1}$ as long as the capacity $C_{kt}$ of the resource in period $t$ is not exceeded. The lead time $L_k$ simply delays the arrival of work to the workcenter after its release into the production unit; it does not describe the behavior of the workcenter itself.

Examination of constraints (5.16) and (5.17) reveals another consequence of the fixed lead times: the output of the production unit in periods 1 through $L$ cannot be influenced by release decisions in periods 1, …, $L$–$1$ but is determined by release decisions in periods –$L$+$1$ through 0 which, since they lie in the past, are assumed to be known with certainty. Thus positive fixed lead times bring the need to initialize the model with information about decisions in the early periods of the planning horizon. These quantities are analogous to the scheduled receipts used in MRP calculations (Baker 1993; Jacobs et al. 2011). Similarly, the model will not plan releases in periods $T - L + 1$ through $T$, since this material can only meet demand in periods $T + 1$ through $T + L - 1$ that lie outside the planning horizon. Thus the use of fixed lead times requires specifying boundary conditions for the planning models at the beginning and end of the planning horizon.

The timing of releases and output under fixed lead times is illustrated in Fig. 5.3, which assumes a fixed lead time of $L = 2$ periods. Releases $R_t$ in each period $t$ are assumed to be uniformly distributed across the period. Hence the output $X_3$ in period 3 is determined by the amount of releases $R_1$ in period 1. However, the output $X_1$ of the production unit in period 1 lies within the fixed lead time, and hence depends on decisions made in the past, in period $t = -1$. To avoid introducing additional notation for these historical release decisions associated with periods $t = -L + 1$ through $t = 0$, we assume henceforth that any decision variable with a non-positive subscript

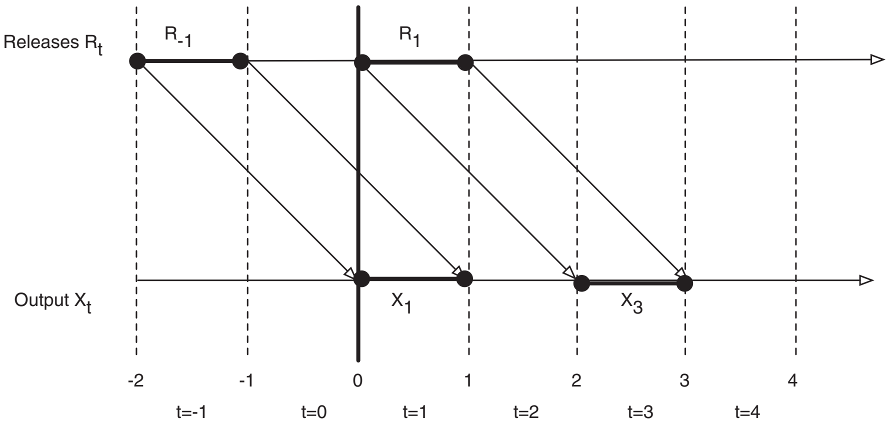


**Figure 5.3** Timing of material flows under integer fixed lead times

is a parameter corresponding to a historical decision. Under this model of fixed lead times, the time series $X_t$, $t = 1,\ldots,T$ representing the output of the workcenter is simply the time series $R_t$, $t = 1,\ldots, T$ of the releases shifted $L$ periods to the right.

Hence under fixed lead times, the output variables $X_t$ and release variables $R_t$ are completely interchangeable. We have written our formulation in terms of the release variables $R_t$, but since $X_t = R_{t-L}$ it is straightforward to write it in terms of the output variables $X_t$.

Finally, the model (5.15)–(5.18) can be rewritten using (5.3) to eliminate the inventory variables. Defining $I_0$ to be the amount of finished goods inventory on hand at the start of the first period in the planning horizon, we see that

$$\begin{aligned} I_t &= I_0 + \sum_{\tau=1}^{t} R_{\tau-L} - \sum_{\tau=1}^{t} D_\tau \\ &= I_0 + \sum_{\tau=1}^{t} X_\tau - \sum_{\tau=1}^{t} D_\tau \end{aligned} \tag{5.21}$$

Substituting (5.21) into (5.15) yields

$$\begin{aligned} \sum_{t=1}^{T}\left(h_t I_t + c_t X_t\right) &= \sum_{t=1}^{T}\left[h_t\left(I_0 + \sum_{\tau=1}^{t} X_\tau - \sum_{\tau=1}^{t} D_\tau\right) + c_t X_t\right] \\ &= \sum_{t=1}^{T}\left[h_t \sum_{\tau=1}^{t} X_\tau + c_t X_t\right] + \sum_{t=1}^{T} h_t\left[I_0 - \sum_{\tau=1}^{t} D_\tau\right] \\ &= \sum_{t=1}^{T}\left[\sum_{\tau=1}^{t} h_\tau R_{\tau-L} + c_t R_{t-L}\right] + \sum_{t=1}^{T} h_t\left[I_0 - \sum_{\tau=1}^{t} D_\tau\right] \\ &= \sum_{t=1}^{T}\left[\sum_{\tau=t}^{T} h_\tau R_{t-L} + c_t R_{t-L}\right] + \sum_{t=1}^{T} h_t\left[I_0 - \sum_{\tau=1}^{t} D_\tau\right] \end{aligned}$$

Discarding constants independent of the decision variables, we can rewrite (5.15)–(5.18) as

$$\min \sum_{t=1}^{T} \left( \sum_{\tau=t}^{T} h_\tau + c_t \right) R_{t-L} \tag{5.22}$$

$$\sum_{\tau=1}^{t-L+1} R_\tau \geq \sum_{\tau=1}^{t} D_\tau - I_0, t = 1, \ldots, T \tag{5.23}$$

$$R_{t-L_k} \leq C_k, k = 1, \ldots, K; t = 1, \ldots, T \tag{5.24}$$

$$R_t \geq 0, t = 1, \ldots, T \tag{5.25}$$

Model (5.22)–(5.25) shows that the $I_t$ variables are not essential; they are a consequence of the primary decisions, given by the releases $R_t$, and the constraints describing the behavior of the system. While the model (5.6)–(5.9) is probably more familiar to the reader, as it is widely used in textbooks, the model (5.22)–(5.25) provides some advantages when analyzing the structure of optimal solutions, particularly the dual solutions that we shall examine later in this chapter.

This basic formulation can be extended in a number of directions without materially affecting its structure. Models involving lot-sizing considerations due to the presence of setup costs or setup times, such as that of Billington et al. (1983) or those studied by Pochet and Wolsey (2006), involve integer variables—a significant difference from a computational perspective—but their treatment of capacity and lead times is essentially the same. Far more elaborate objective functions are possible, but our emphasis is on the representation of production capacity and material flow. The assumption of no backlogging can be relaxed in the standard manner (Johnson and Montgomery 1974). Since a backlog corresponds to a negative inventory level, we can represent the net inventory level $N_t$ as the difference of two nonnegative variables $N_t = I_t - B_t$, where $I_t$ denotes on-hand, positive inventory, at the end of period $t$, and $B_t$ the backlog. Since the column vectors for $I_t$ and $B_t$ in the constraint matrix of the linear programming model will be linearly dependent, both variables cannot take positive values in an optimal solution.

### 5.3.2.3 Multiple Items

The single-item multiple workcenter model (5.22)–(5.25) extends to the multi-item case with items $j \in J_F$ in a very natural manner. Capacity constraints at each workcenter $k$ must now consider the total capacity consumption by all items $j \in J(k)$ using that workcenter in each period, and separate finished goods inventory balance equations must be written for each product $j$. All lead time parameters are now product-dependent, with $L_j$ denoting the lead time of product $j$ from release until completion and $L_{jk}$ its lead time from release until its availability for processing at workcenter $k$. With these changes, the multi-item model can be written as:

$$\min \sum_{t=1}^{T} \left[ \sum_{j \in J_{\mathrm{F}}} \left( h_{jt} I_{jt} + \sum_{k \in K(j)} c_{jkt} R_{j,t-L_k} \right) \right] \tag{5.26}$$

subject to

$$I_{jt} = I_{j,t-1} + R_{t-L_j} - D_{jt}, \quad \forall j \in J_{\mathrm{F}}, \quad t = 1,\ldots,T \tag{5.27}$$

$$\sum_{j \in J(k)} a_{jk} R_{t-L_{jk}} \le C_{kt}, \quad \forall k \in K, \quad t = 1,\ldots,T \tag{5.28}$$

$$I_{jt}, \quad R_{jt} \ge 0, \quad \forall j \in J_{\mathrm{F}}, \quad t = 1,\ldots,T \tag{5.29}$$

The only representation of resource contention between the products $j$ at the workcenters $k$ is the left hand side of (5.28), which is linear in the release quantities of each product. This is in marked contrast to Fig. 2.2, where the output of the resources is a concave non-decreasing function of the workload, determined by the production quantities. The presence of multiple products with different processing times on the workcenter will result in increased coefficients of variation of the processing times $P_{jkt}$ and a downward shift in the output function. The lead times $L_{jk}$ are also unaffected by production quantities, in contrast to the highly nonlinear behavior of the cycle time with workload seen in Fig. 2.1. It begins to be apparent that the workcenter behavior described by this model differs quite fundamentally from that of the queueing models discussed in Chap. 2. The inventory variables can also be eliminated using (5.21), resulting in the formulation

$$\min \sum_{j \in J} \sum_{t=L_j+1}^{T} \left( \sum_{\tau=t}^{T} h_j \right) R_{j,t-L_j} = \sum_{j \in J} \sum_{t=L_j+1}^{T} \big( (T-t+1) h_j \big) R_{j,t-L_j} \tag{5.30}$$

subject to

$$\sum_{\tau=1}^{t} R_{j,\tau-L_j} \ge \sum_{\tau=1}^{t} D_{j\tau} - I_{j0}, \quad \forall j \in J, \quad t = 1,\ldots,T \tag{5.31}$$

$$\sum_{j \in J(k)} a_{jk} R_{t-L_{jk}} \le C_{kt}, \quad \forall k \in K, \quad t = 1,\ldots,T \tag{5.32}$$

$$R_{jt} \ge 0, \quad \forall j \in J, \quad t = L_j,\ldots,T-L_j \tag{5.33}$$

We now use this formulation to discuss the dual model and its interpretation.

## 5.4 Dual Formulation

Unless it is infeasible or its optimal value is unbounded, any linear program is associated with another linear program, its dual, whose optimal value is equal to that of the original (the primal) at optimality (Bazaraa et al. 2004). Each decision variable in the dual is associated with a constraint in the primal and each dual constraint with a primal decision variable. Thus, the generic linear program

$$\min \sum_{j=1}^{n} c_j x_j \tag{5.34}$$

subject to

$$\sum_{j=1}^{n} a_{ij} x_j \geq b_i, \quad i = 1, \ldots, m \tag{5.35}$$

$$x_j \geq 0, \quad j = 1, \ldots, n \tag{5.36}$$

will be associated with its dual

$$\max \sum_{i=1}^{m} b_i y_i \tag{5.37}$$

subject to

$$\sum_{i=1}^{m} a_{ji} y_i \leq c_j, \quad j = 1, \ldots, n \tag{5.38}$$

$$y_i \geq 0, \quad i = 1, \ldots, m \tag{5.39}$$

The dual variables $y_i$ associated with each primal constraint $i$ correspond to the Lagrange multipliers associated with that constraint, representing the partial derivative of the optimal objective function value with respect to the right-hand side $b_i$ of constraint $i$ at optimality. An important property arising from the Kuhn–Tucker optimality conditions for linear programs is the complementary slackness condition

$$y_i \left( \sum_{j=1}^{n} a_{ij} x_j - b_i \right) = 0, \quad i = 1, \ldots, m \tag{5.40}$$

The dual variables have an economic interpretation that is often helpful in interpreting the results of a model. An important advantage of the models developed in Chap. 7 is their ability to provide richer dual information than that obtained from the models discussed in this chapter.

Since our primary concern lies with production planning models, we discuss duality in an intuitive, heuristic fashion; rigorous mathematical treatments are given by Bazaraa et al. (2004) and Bertsimas and Tsitsiklis (1997). Correct interpretation of dual variables can be quite subtle, especially in the presence of a degenerate optimal solution where some constraints are redundant; extensive discussions of these issues are given by Jansen et al. (1997), Koltai and Terlaky (2000), and Rubin and Wagner (1990). To avoid the extensive mathematical digressions required to address the issues in estimating dual prices in the face of degenerate optimal solutions, our discussion will assume that all optimal solutions are non-degenerate, closely following the development in Kefeli (2011) but omitting some details to focus on insights.

We will develop the dual formulation for the model (5.30)–(5.33). For further simplicity in exposition, we shall assume all costs are time-stationary such that, for example, $c_{jt} = c_j$ for all periods $t$. In this case, the no-backlogging assumption implies that as long as a feasible solution exists, in any optimal solution total production of any product will exactly equal its total demand net of the initial inventories $I_{j0}$, and the production costs will have no influence on the optimal solution. This results in the simplified primal linear program

$$\min \sum_{j\in J} \sum_{t=L_j+1}^{T} \left( \sum_{\tau=t}^{T} h_j \right) R_{j,t-L_j} = \sum_{j\in J} \sum_{t=L_j+1}^{T} \left( \left( T-t+1 \right) h_j \right) R_{j,t-L_j} \tag{5.41}$$

subject to

$$\sum_{\tau=1}^{t} R_{j,\tau-L_j} \geq \sum_{\tau=1}^{t} D_{j\tau} - I_{j0}, \quad \forall j \in J, \quad t = 1,\ldots,T \quad \left( \gamma_{jt} \right) \tag{5.42}$$

$$\sum_{j\in J(k)} a_{jk} R_{t-L_{jk}} \leq C_{kt}, \quad \forall k \in K, \quad t = 1,\ldots,T \quad \left( \sigma_{kt} \right) \tag{5.43}$$

$$R_{jt} \geq 0, \quad \forall j \in J, \quad t = L_j,\ldots,T-L_j \tag{5.44}$$

The Greek letters in parentheses denote the dual variables associated with each constraint set. The dual of this linear program is given by

$$\max \sum_{t=1}^{T} \sum_{j\in J} \left[ \left( \sum_{\tau=1}^{t} D_{j\tau} - I_{j0} \right) \gamma_{jt} - \sum_{k\in K} C_{kt} \sigma_{kt} \right] \tag{5.45}$$

subject to

$$\begin{gathered} \sum_{\tau=t}^{T} \gamma_{j\tau} - \sum_{k\in K(j)} a_{jk} \sigma_{k,t-\left(L_j - L_{jk}\right)} \leq \left( T-t+1 \right) h_j, \quad \left( R_{j,t-L_j} \right) \\ \forall j \in J, \quad t = \left( L_j - L_{jk} \right) + 1,\ldots,T \end{gathered} \tag{5.46}$$

$$\gamma_{jt} \geq 0, \quad \forall j \in J, \quad t = 1,\ldots,T \tag{5.47}$$

$$\sigma_{kt} \geq 0, \quad \forall k \in K, \quad t = 1,\ldots,T \tag{5.48}$$

The primal variables corresponding to the dual constraints are shown next to each dual constraint set. While the primal problem chooses releases $R_{jt}$ in each period $t$ to minimize the cost of meeting demand under capacity constraints, the dual problem chooses prices $\gamma_{it}$ and $\sigma_{kt}$ to maximize revenue. $\gamma_{jt}$ can be interpreted (subject to the mathematical caveats discussed by Rubin and Wagner (1990) and others) as the minimum amount the firm should charge an additional unit of demand for item $j$ in period $t$. $\sigma_{kt}$, on the other hand, represents the maximum amount the firm should be willing to pay to acquire an additional unit of resource $k$ in period $t$. The cost coefficients $(T - t + 1)h_j$ of the primal problem represent the contribution to the total cost of a unit of item $j$ produced in period $t$, given by its incremental contribution to holding cost until the end of the planning horizon.

The first term in (5.45) represents the revenue from an additional unit of demand for item $j$ in period $t$, which will increase the cumulative net demand $\sum_{\tau=1}^{t} D_{j\tau} - I_{j0}$ in each subsequent period until the end of the horizon. The second term in (5.45) represents the marginal cost of all resources required to process this additional unit of demand; recall that all demands must be met without backlogging. Hence the right-hand side of (5.46) represents the net marginal revenue (marginal revenue minus marginal resource costs) associated with an additional unit of demand for item $j$ in period $t$. (5.46) ensures that the total marginal cost of the additional item cannot exceed its marginal net revenue. The complementary slackness property (5.40) implies that when there is positive slack in constraint (5.46) for some item $j$ and period $t$ at optimality, we will have $R_{jt} = 0$ in an optimal solution. Conversely, $R_{jt} > 0$ at optimality implies (5.46) is satisfied at equality.

Our primary interest in this discussion is the dual variables $\sigma_{kt}$ associated with the primal capacity constraints (5.43). These dual variables represent the impact on the objective function of an additional unit of capacity at resource $k$ in period $t$, which is of interest for several reasons. A high value of this dual variable indicates that limited capacity at this machine is significantly affecting the ability of the production unit to meet demand in a cost-effective manner, suggesting particular attention by management to improving its performance. It will also turn out, as we shall see in Chap. 7, that the clearing function formulations introduced in that chapter yield much more informative dual information than that obtained from this model, as we shall illustrate below.

Recall that a unit of product $j$ that completes processing in period $t$ will consume capacity on its $k$'th workcenter in period $t - L_j + L_{jk}$. Thus the output $X_{jt} = R_{j,t-L_j}$ of each item $j$ in any period $t$ is potentially constrained by at most $|K(j)|$ of the capacity constraints (5.43), each corresponding to a workcenter $k$ in period $t - L_j + L_{jk}$. To ensure a non-degenerate optimal solution, we shall assume that for each item $j$ at most one of these associated capacity constraints is satisfied at equality; this condition can be enforced if necessary by perturbing the right-hand side of the constraints by an arbitrarily small quantity. The specific workcenter $k$ whose capacity constraint is binding in period $t - L_j + L_{jk}$ will be denoted by $k^*(j,t)$, indicating that this workcenter limits the output of item $j$ in period $t$. We will refer to resource $k^*(j,t)$ as the limiting workcenter for item $j$ in period $t$. Our assumption of non-degeneracy implies at most one limiting workcenter for each product $j$ in each period $t$ of the planning horizon. The limiting workcenter of an item $j$ may be used concurrently by other items and may change from period to period, i.e., it is perfectly possible to have $k^*(j,t) \neq k^*(j,s)$ for $t \neq s$. Different items $j$ may have different limiting resources in a given period.

As long as the production costs $c_t$ are non-decreasing in the time period $t$, it is straightforward to show that an optimal solution to the primal will satisfy

$$\left(\sum_{j\in J(k)} a_{j,k^*(j,t)} R_{j,t-L_j+L_{j,k^*(j,t)}} - C_{k^*(j,t),t}\right) I_{j,t-1} = 0 \tag{5.49}$$

implying that the model will hold finished inventory against future demand in period $t - 1$ if and only if capacity at a resource $k^*(j,t)$ is fully utilized in period $t - L_j + L_{j,k^*(j,t)}$. Hence the model will hold finished inventory of product $j$ in some period $t$ only if the total demand for all items $\sum_{j\in J} D_{js}$ in some future period $s > t$ overloads the available capacity on its limiting resource $k^*(j,s)$ for period $s$, i.e.,

$$\sum_{j\in J} D_{js} > C_{k^*(j,s),s} \tag{5.50}$$

requiring the model to meet the demand in period $s$ by building up finished inventory in periods prior to $s$. Periods $s$ into which no finished inventory is carried in the optimal solution indicate that the optimal decisions for periods $s < t$ are independent of those for periods $s \geq t$. Hence an optimal solution to (5.41)–(5.44) will consist of one or more busy intervals, each consisting of $q \geq 0$ consecutive periods $S = \{s-q, s-q+1, \ldots, s\}$ with $I_{jq} > 0$ for some items $j \in J$ such that

$$\sum_{j\in J(k^*(j,s))} a_{jk} R_{j,s-L_j+L_{j,k^*(j,s)}} = C_{k^*(j,s),s} \tag{5.51}$$

Since the limiting workcenter $k^*(j,s)$ has a binding capacity constraint in period $s$ by definition, our assumption of a non-degenerate solution implies that all products $i \neq j$ requiring this resource in this period will have positive inventory in this busy interval, implying that $\gamma_{jt} = 0$ by the complementary slackness condition for constraints (5.42). Based on these observations, we will have dual prices $\sigma_{kt} > 0$ in periods $(s-q)-\left(L_j - L_{jk^*}\right)$ for all products $j$ that use workcenter $k^*(j,s)$ in period $s$. We shall restrict our attention to periods in this interval where production activity is taking place, i.e., $X_{jt} = R_{j,t-L_j} > 0$. By complementary slackness, the dual constraints (5.46) will be tight in periods $s - q$ through $s$. Solving recursively from period $s + 1$ backwards in time to period $s$, Kefeli (2011) shows that

$$a_{j,k^*(j,t-1)}\sigma_{k^*(j,t-1),t-L_j+L_{j,}k^*(j,t-1)} - a_{j,k^*(j,t)}\sigma_{k^*(j,t),t-L_j+L_{j,}k^*(j,t)} = h_j \tag{5.52}$$

Our assumption of non-degeneracy implies that the output of each product $j$ is limited by at most one resource $k$ in each period, but there may be multiple workcenters with positive dual prices corresponding to different subsets of products. The limiting workcenter for a product $j$ may also change from one period to the next, i.e., $k^*(j,t-1) \neq k^*(j,t)$. When the same workcenter is limiting for item $j$ in two consecutive periods $t - 1$ and $t$, (5.52) simplifies to

$$\left(\sigma_{k^*(j,t),(t-1)-L_j+L_{j,}k^*(j,t)} - \sigma_{k^*(j,t),t-L_j+L_{j,}k^*(j,t)}\right) = \frac{h_j}{a_{j,k^*(j,t)}} \tag{5.53}$$

Examining this expression shows that the absolute value of the dual price of capacity increases linearly with time over the busy interval, starting with a value of zero and increasing in absolute value by $h_j / a_{j,k^*(j,t)}$ in each period. This is intuitive; an additional unit of capacity at workcenter *k*(j,t)* in period $t - L_{j,k^*(j,t)}$ will allow $1 / a_{j,k^*(j,t)}$ units held in inventory in period *s* to be produced in this period, reducing holding costs by $\left(t - s + 1\right) / a_{j,k^*(j,t)}$.

The following numerical example illustrates this structure of optimal solutions.

**Example 5.1** To illustrate the structure of the optimal solution and the dual variables, consider a production unit with two products and four workcenters with the data given in Table 5.1. The unit finished goods holding costs are given by $h_1 = h_2 = 5$, and the overall lead times by $L_1 = L_2 = 5$. Initial inventory at the end of period 0 is $I_{10} = 20$ units for Product 1 and $I_{20} = 25$ units for Product 2. Both products require processing on all four resources in increasing order of machine number. The demand for each product in each period is given in Table 5.2.

**Table 5.1** Parameter values for Example 5.1

| | Item | Machine 1 | Machine 2 | Machine 3 | Machine 4 |
|---|---|---|---|---|---|
| Production cost | 1 | 1 | 1 | 1 | 2 |
| | 2 | 2 | 2 | 2 | 2 |
| Processing time | 1 | 3 | 3 | 2 | 4 |
| | 2 | 3 | 4 | 4 | 4 |
| Lead time $L_{jk}$ | 1 | 0 | 1 | 2 | 4 |
| | 2 | 0 | 1 | 2 | 4 |
| Capacity $C_{kt}$ | | 100 | 100 | 18 | 20 |

**Table 5.2** Demand data for Example 5.1

| Period | Item 1 demand | Item 2 demand |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 0 | 0 |
| 3 | 5 | 0 |
| 4 | 4 | 0 |
| 5 | 4 | 2 |
| 6 | 5 | 4 |
| 7 | 5 | 5 |
| 8 | 5 | 5 |
| 9 | 6 | 3 |
| 10 | 7 | 4 |
| 11 | 6 | 3 |
| 12 | 6 | 3 |
| 13 | 0 | 6 |
| 14 | 0 | 5 |
| 15 | 0 | 0 |

Solving the primal model (5.41)–(5.44) yields the optimal solution given in Table 5.3, with optimal objective function value 2636.25. Machine 3 is the limiting resource for Product 1 in periods 4 through 8, and for Product 2 in periods 9 and 10. The dual prices $\sigma_{kt}$ associated with this machine in the optimal solution are plotted in Fig. 5.4; only Machine 3 has nonzero dual prices. The relation (5.53) can be clearly observed, with the dual price increasing linearly until the capacity loading falls below resource capacity. Note that although Machine 4 has utilization of 0.9 in periods 12 and 13, and would thus be expected to have high cycle time and WIP, its dual price remains at zero since the capacity constraint is not binding.

**Table 5.3** Optimal solution for Example 5.1

| | Releases, $R_j$ | | Resource loading | | | | Ending inventory, $I_{jt}$ | |
|---|---|---|---|---|---|---|---|---|
| Period | Item 1 | Item 2 | Machine 1 | Machine 2 | Machine 3 | Machine 4 | Item 1 | Item 2 |
| −3 | | | | | | | | |
| −2 | | | | | | | | |
| −1 | | | | | | | | |
| 0 | | | | | | | 20 | 25 |
| 1 | 0 | 3.75 | 11.25 | 0 | 0 | 0 | 20 | 25 |
| 2 | 3 | 2.25 | 15.75 | 15 | 0 | 0 | 20 | 25 |
| 3 | 6 | 0 | 18 | 18 | 15 | 0 | 15 | 25 |
| 4 | 6 | 0 | 18 | 18 | 18 | 0 | 11 | 25 |
| 5 | 6 | 0 | 18 | 18 | 18 | 15 | 7 | 23 |
| 6 | 6 | 0 | 18 | 18 | 18 | 15 | 2 | 22.75 |
| 7 | 6 | 0 | 18 | 18 | 18 | 12 | 0 | 20 |
| 8 | 0 | 4.5 | 13.5 | 18 | 18 | 12 | 1 | 15 |
| 9 | 0 | 4.5 | 13.5 | 18 | 18 | 12 | 1 | 12 |
| 10 | 0 | 0 | 0 | 18 | 18 | 12 | 0 | 8 |
| 11 | 0 | 0 | 0 | 0 | 18 | 12 | 0 | 5 |
| 12 | 0 | 0 | 0 | 0 | 0 | 18 | 0 | 2 |
| 13 | 0 | 0 | 0 | 0 | 0 | 18 | 0 | 0.5 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

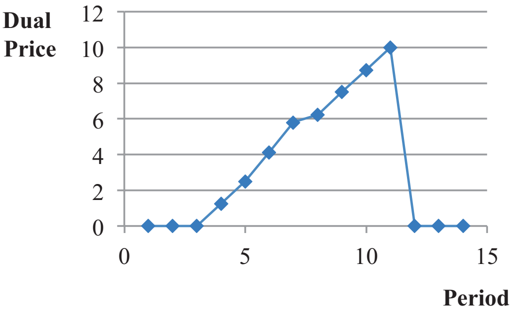


**Fig. 5.4** Dual variables associated with Machine 3 in optimal solution to Example 5.1

### 5.4.1 Insights from the Dual

Our analysis of the dual solution indicates a number of drawbacks of the formulation (5.41)–(5.44), particularly its representation of workcenter behavior. By complementary slackness, the dual variables $\sigma_{kt}$ associated with the capacity constraints (5.43) will only take nonzero values if the associated primal constraint is binding at optimality, implying that the workcenter's capacity is fully utilized. Since $\sigma_{kt}$ represents the maximum amount the firm should be willing to pay for an additional unit of output from workcenter $k$ in period $t$, this implies that no improvement in the optimal objective function value can be obtained from additional output at a workcenter unless its capacity is fully utilized. However, as discussed in Chap. 2, queueing models suggest qualitative changes in the behavior of a capacitated workcenter at utilization levels well below 1; more precisely, they show a nonlinear increasing relation proportional to $1/(1-u)$ between cycle time and utilization (Hopp and Spearman 2008), implying that additional capacity at the workcenter might improve system performance even though currently capacity is not fully utilized. Likewise, improving the performance of a workcenter such that it can generate more output for the same average WIP level, shifting the curves in Fig. 2.2 to the left, should allow reduced cycle time and hence reduced costs, which the current model is unable to capture. Note, however, that this does not necessarily imply that adding capacity would be economically beneficial, especially if capacity can only be added in discrete increments.

A second drawback of this model can be observed directly in (5.53): the dual price of a resource in a period is independent of events at other resources as long as the limiting resources do not change. This again contradicts insights from queueing models (Hopp and Spearman 2008), which show that the behavior of downstream resources is affected by the utilization of upstream ones. Consider two resources operating in series where work flows from workcenter 1 to workcenter 2. Per Hopp and Spearman (2008) Chap. 8, the squared coefficient of variation (SCV) of the interarrival times at workcenter 2 is given by the SCV of the departure process from workcenter 1, which, in turn, can be approximated by

$$c_{\mathrm{d}}^2 = u^2 c_{\mathrm{e}}^2 + \left(1-u^2\right) c_{\mathrm{a}}^2 \tag{5.54}$$

where $u$ denotes the average utilization of workcenter 1, $c_{\mathrm{e}}^2$ the SCV of the effective processing time distribution at workcenter 1, and $c_{\mathrm{a}}^2$ the SCV of the external arrival process to workcenter 1. This relation suggests that the dual price of capacity at workcenter 2 ought to be influenced by decisions at workcenter 1; under most conditions, unless $c_{\mathrm{e}}$ is small relative to $c_{\mathrm{a}}$, adding capacity to workcenter 1 will reduce $u$, reducing the average cycle time at workcenter 2 which ought to improve overall performance, or at least leave it no worse.

This analysis of the dual prices of capacity suggests that the use of fixed lead times can model the behavior of production resources subject to queueing behavior to at best a limited degree. The largest discrepancies are to be expected when resource utilization levels vary significantly over time, causing the fixed lead times to over- and/or underestimate actual cycle times; and when multiple resources have high utilization levels close to, but not quite equal to, 1, such that small changes in utilization lead to large changes in cycle times.

## 5.5 Fractional Lead Times

Our discussion of fixed lead times up to this point has assumed lead times expressed as integer multiples of the planning period length $\Delta$, recalling that period $t$ ends at time $t\Delta$. Assuming that all release and demand rates are uniform over each planning period, Hackman and Leachman (1989b) have shown that non-integer fixed lead times can be incorporated easily. We first illustrate the basic idea with a single-product single-workcenter model and then discuss generalizations.

Any fractional fixed lead time $L$ can be decomposed into integer and fractional parts as $L = \lfloor L \rfloor + \phi$, where $\lfloor L \rfloor$ denotes the largest integer less than or equal to $L$ and $\phi = L - \lfloor L \rfloor$ the fractional part of the lead time. We assume $L$ remains constant in all planning periods; the case where lead times can vary over time is addressed in the next chapter. Under uniform release and demand rates over the planning period, if $R_t$ units of a product are released during this period, the material will enter the production unit at a rate of $R_t/\Delta$ units per unit time. The material flow through the workcenter can then be represented as in Fig. 5.5. The upper timeline represents the progression of releases into the production unit over time and the lower timeline the entry of this material into finished inventory. The amount of material becoming available to meet demand in period $t$ is given by

$$Y_t = \phi R_{t-\lfloor L \rfloor -1} + (1-\phi) R_{t-\lfloor L \rfloor} \tag{5.55}$$

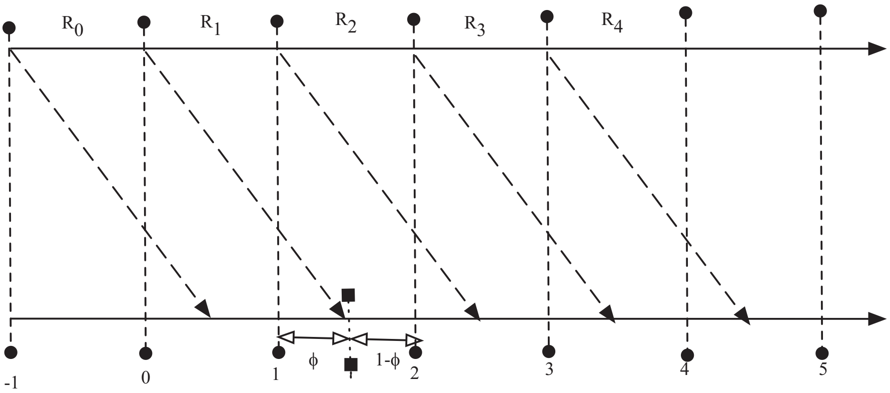


**Fig. 5.5** Fractional lead times

**Table 5.4** Data for fractional lead time example

| Period | Releases | Outputs | Demand | Ending inventory |
|---|---|---|---|---|
| −1 | 50 | | – | – |
| 0 | 80 | | – | 0 |
| 1 | 120 | 65 | 50 | 15 |
| 2 | 150 | 100 | 110 | 5 |
| 3 | 150 | 135 | 140 | 0 |
| 4 | 160 | 150 | 150 | 0 |
| 5 | 150 | 155 | 155 | 0 |
| 6 | 80 | 155 | 150 | 5 |
| 7 | 25 | 115 | 100 | 20 |
| 8 | 0 | 52.5 | 60 | 12.5 |
| 9 | 0 | 12.5 | 25 | 0 |

and the material balance constraint analogous to (5.16) takes the form

$$I_t = I_{t-1} + \phi R_{t-\lfloor L \rfloor -1} + (1-\phi) R_{t-\lfloor L \rfloor} - D_t, \quad t = 1,\ldots,T \tag{5.56}$$

where $I_t$ denotes the amount of finished goods inventory at the end of period $t$.

However, let us take a closer look at the implications of this formulation. Recall that we assume constant release and production rates throughout each planning period. Now consider the data given in Table 5.4, under a fixed lead time of $L = 1.5$ periods and $R_t = 0$ for $t < -1$.

The output (production) in each period $t$ is computed assuming that releases $R_t$ and demands $D_t$ are uniformly distributed across their associated planning periods as in (5.56). The ending inventory is computed using the inventory balance equation (5.56) at the end of each period. The reader should verify these calculations to confirm that inventory levels are nonnegative at the end of all planning periods.

However, all is not as it seems. Although the release rate over each planning period is constant, the output rate, which defines the rate of inflow into the inventory, is not. Due to the fractional nature of the lead time, material released at the start of period $t$ emerges as output in the middle of period $t + 1$, as illustrated in Fig. 5.6, where each period is divided into two subintervals of length $\phi$ and $1 - \phi$, in this case both equal to 0.5 periods. In periods 1, 2, and 3, the output rate of the production resource during the first subinterval of the period is different from that in the second subinterval.

Table 5.5 recalculates Table 5.4 at each half-period. As the reader can (and should!) verify, changes in output rates within the planning periods result in negative inventory levels at some of these intermediate points.

As pointed out by Hackman and Leachman (1989b), there are two possible solutions to this problem. The most obvious, especially in the very structured example we have used here, is to reduce the size of the planning periods such that rate changes within planning periods are no longer possible, and enforce material balance and capacity constraints at the boundaries of each of these subintervals.

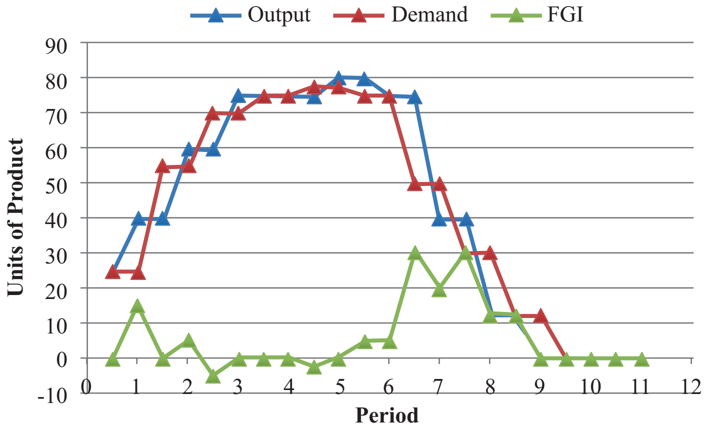


**Fig. 5.6** Output of production unit with fractional lead times per Table 5.5

**Table 5.5** Effect of fractional lead times at interior points of planning periods

| Period | Output | Demand | FGI |
|---|---|---|---|
| 0.5 | 25 | 25 | 0 |
| 1 | 40 | 25 | 15 |
| 1.5 | 40 | 55 | 0 |
| 2 | 60 | 55 | 5 |
| 2.5 | 60 | 70 | -5 |
| 3 | 75 | 70 | 0 |
| 3.5 | 75 | 75 | 0 |
| 4 | 75 | 75 | 0 |
| 4.5 | 75 | 77.5 | –2.5 |
| 5 | 80 | 77.5 | 0 |
| 5.5 | 80 | 75 | 5 |
| 6 | 75 | 75 | 5 |
| 6.5 | 75 | 50 | 30 |
| 7 | 40 | 50 | 20 |
| 7.5 | 40 | 30 | 30 |
| 8 | 12.5 | 30 | 12.5 |
| 8.5 | 12.5 | 12.5 | 12.5 |
| 9 | 0 | 12.5 | 0 |
| 9.5 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 |
| 10.5 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 |

To ensure consistency, the length of the subintervals must be equal to the least common divisor of the fractional lead times. This will significantly increase the size of the formulation, since both the number of constraints and the number of decision variables depend on the number of periods. It is also impractical in the presence of the time-varying lead times discussed in Chap. 6, where the lead time associated with each period may have a different fractional part. In this case a period length equal to the greatest common divisor of all lead times must be used, which may result in a much larger model than necessary.

Hackman and Leachman (1989b) propose a much simpler solution to this difficulty by noting that it is only necessary to write additional constraints at points in time where output or release rates may change. This includes the boundaries of the original planning periods and intermediate points where a fractional lead time causes a change in output (and hence the rate of inflow into finished inventory) or the amount of material requiring capacity at a particular resource. Under the time-stationary lead times assumed in this chapter, each planning period will have at most one intermediate point for which additional constraints for a given product need to be written.

Although we have focused on the overall lead time $L_j$ of the production unit for a particular item $j$, the same issues arise with respect to the capacity constraints for each workcenter $k$ and their associated lead times $L_k$. In this case the changes in release rate within a planning period may result in capacity constraints being violated at interior points of the period (Hackman and Leachman 1989b). To see this, consider the situation illustrated in Fig. 5.7 where we have two items whose respective lead times are $L_{1k} = 1.5$ and $L_{2k} = 1.75$ periods. The upper time line shows the releases of each item and the lower the arrival of each item at the resource under consideration. Recalling our convention that period $t$ ends at time $t$, the rate of material arriving at the resource $k$ during period $t$ can change at three potential points in time: $t - 1$, $t + \phi_{1k}$, and $t + \phi_{2k}$, where $\phi_{jk} = L_{jk} - \lfloor L_{jk} \rfloor$, requiring the capacity constraints

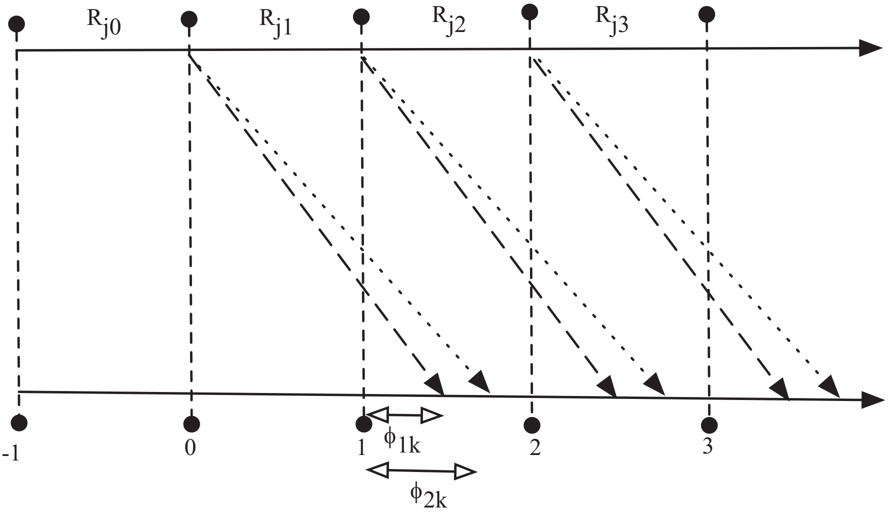


**Fig. 5.7** Impact of multiple products with fractional lead times on capacity constraints

$$\begin{aligned} \phi_{1k}\left(a_{1k}R_{t-L_{1k}-1}+a_{2k}R_{t-L_{2k}-1}\right) &\le \phi_{1k}C_{kt} \\ \left(\phi_{2k}-\phi_{1k}\right)\left(a_{1k}R_{t-L_{1k}}+a_{2k}R_{t-L_{2k}-1}\right) &\le \left(\phi_{2k}-\phi_{1k}\right)C_{kt} \\ \left(1-\phi_{2k}\right)\left(a_{1k}R_{t-L_{1k}}+a_{2k}R_{t-L_{2k}}\right) &\le \left(1-\phi_{2k}\right)C_{kt} \end{aligned} \tag{5.57}$$

This approach results in a large number of additional capacity constraints, especially in environments such as semiconductor wafer fabrication where a given resource may be used by tens of different unit operations. Leachman (2001) points out that the presence of many items $j$ with slightly different fractional components $\phi_{jk}$ is likely to yield a roughly uniform distribution of workload over the planning interval, allowing approximate capacity constraints of the form

$$\sum_{j\in J(k)} a_{jk}\left[\phi_{jk}R_{t-L_{jk}-1}+\left(1-\phi_{jk}\right)R_{t-L_{jk}}\right]\le C_{kt},\quad t=1,\ldots T;\quad k=1,\ldots,K \tag{5.58}$$

to be used without inducing excessive error. Note that (5.58) simply adds up the total amount of each product loading the resource within the planning period, without considering the specific timing of the loading within the period. The basic operation of these constraints is the same as that for material flow discussed above and can be illustrated in the following example.

**Example 5.2** Consider a single resource and three products with fixed fractional lead times $L_1 = 1.3$, $L_2 = 1.5$, and $L_3 = 1.75$ that remain constant over a planning horizon consisting of $T = 12$ periods. Thus we have $\phi_1 = 1.3 - \lfloor 1.3 \rfloor = 0.3$, $\phi_2 = 0.5$, and $\phi_3 = 0.75$ by the same logic. Following Fig. 5.7, the intervals within which the capacity loading from each product will remain constant, assuming constant release rates over each planning period, are calculated in Table 5.6.

Capacity loading of the resource remains constant over each interval with the given start and end points. Due to the fractional lead times, the rate of capacity

**Table 5.6** Uniform loading intervals for Example 5.2

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prod. 1 | Start | 0.3 | 1 | 1.3 | 2 | 2.3 | 3 | 3.3 | 4 | 4.3 | 5 | 5.3 | 6 | 6.3 |
| | End | 1 | 1.3 | 2 | 2.3 | 3 | 3.3 | 4 | 4.3 | 5 | 5.3 | 6 | 6.3 | 7 |
| Prod. 2 | Start | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 | 6.5 |
| | End | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 | 6.5 | 7 |
| Prod. 3 | Start | 0.75 | 1 | 1.75 | 2 | 2.75 | 3 | 3.75 | 4 | 4.75 | 5 | 5.75 | 6 | 6.75 |
| | End | 1 | 1.75 | 2 | 2.75 | 3 | 3.75 | 4 | 4.75 | 5 | 5.75 | 6 | 6.75 | 7 |
| Prod. 1 | Start | 7 | 7.3 | 8 | 8.3 | 9 | 9.3 | 10 | 10.3 | 11 | 11.3 | 12 | 12.3 | |
| | End | 7.3 | 8 | 8.3 | 9 | 9.3 | 10 | 10.3 | 11 | 11.3 | 12 | 12.3 | 13.3 | |
| Prod. 2 | Start | 7 | 7.5 | 8 | 8.5 | 9 | 9.5 | 10 | 10.5 | 11 | 11.5 | 12 | 12.5 | |
| | End | 7.5 | 8 | 8.5 | 9 | 9.5 | 10 | 10.5 | 11 | 11.5 | 12 | 12.5 | 13.5 | |
| Prod. 3 | Start | 7 | 7.75 | 8 | 8.75 | 9 | 9.75 | 10 | 10.8 | 11 | 11.8 | 12 | 12.8 | |
| | End | 7.75 | 8 | 8.75 | 9 | 9.75 | 10 | 10.8 | 11 | 11.8 | 12 | 12.8 | 13.8 | |

loading can change at the start of any of these intervals for any product; hence it is necessary to write capacity constraints for each subinterval arising from the intersections of consecutive loading intervals for the individual products. Thus constraints similar to (5.57) need to be written for the intervals (0.3, 0.5), (0.5, 0.75), (0.75, 1), (1, 1.3), (1.3, 1.5), and so on, resulting in a total of 53 intervals that need to be considered explicitly. The large number of constraints required by this approach is immediately evident.

Assuming that each unit of each product requires a single unit of capacity for one planning period to be completed, we now calculate the capacity loading, in terms of units of capacity, for each interval assuming the releases in Table 5.7.

Figure 5.8 plots the total loading of the resource by all three products for the release schedule shown in Table 5.7. The interval load plot shows the load, in terms of the number of parallel machines that would be required to process all the work available in the interval, for each of the 53 subintervals over which load remains constant, while the period load plots the total load within each planning period using (5.58). Discrepancies between the two plots arise where one would expect, in regions where the releases, and hence the loading of the resources, is changing, which in the example are at the start and end of the planning horizon and between periods 5 and 6, where the releases of Product 2 are temporarily interrupted. As the number of products and the number of different $\phi_j$ values increase, and especially if the $\phi_j$ values are distributed somewhat uniformly between 0 and 1, the error induced

**Table 5.7** Release schedule for Example 5.2

| Period | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prod. 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Prod. 2 | 2 | 2 | 2 | 2 | 1 | 0 | 0 | 0 | 2 | 2 | 2 | 2 | 2 |
| Prod. 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |

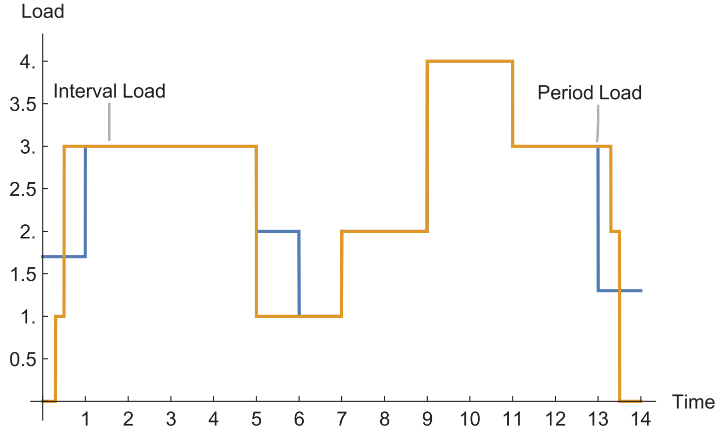


**Fig. 5.8** Capacity loading with fractional lead times

by using (5.58) is likely to be considerably smaller than that arising from other sources, such as errors in demand forecasts.

Kacar et al. (2016) compare the performance of planning models using integer and fractional lead times on a data set representing a large semiconductor wafer fabrication facility and find that incorporating fractional lead times for both finished inventory balance and capacity constraints yields markedly superior performance than including it for either one alone. The fractional lead time model including the aggregate capacity constraints (5.58) and the finished inventory balance constraints (5.56) yielded significantly better performance than a model using integer lead times, and comparable performance to the much larger model using clearing functions described in Chap. 7.

Given the magnitude of the performance improvement and the capabilities of today's commercial LP solvers, we see no reason not to use fractional lead times if they appear to be called for. The most likely case where fractional lead times will be beneficial is when the cycle times of the production system and individual resources span multiple planning periods, and the fractional parts of the lead times are substantial relative to the length of the planning period.

## 5.6 Input-Output Models: An Alternative View of Fixed Lead Times

Our discussion so far has assumed that the entire quantity $R_{jt}$ of product $j$ released into the production unit in period $t$ (the production orders released in period $t$) moves through the production unit as a single entity, such that all items released in that period consume capacity and enter finished inventory together. Given the assumption of releases taking place at a constant rate over the planning periods, each unit of product $j$ will be processed at resource $k$ $L_{jk}$ time units after its release. Hence these lead times represent the time elapsing between the release of the material to the first resource on its routing and its consuming capacity on resource $k$. Under integer lead times, this implies that all materials released in period $t$ will consume capacity at resource $k$ in period $t + L_{jk,}$, i.e., at the end of the specified lead time; the case of fractional lead times is a simple extension of this idea as discussed in the previous section.

An alternative view of fixed lead times allows a production order to consume capacity on a workcenter anywhere within the time it is expected to spend at the workcenter estimated by (5.11). This requires defining new lead time parameters $L'_{jk}$ representing the arrival times of the orders at the workcenters, i.e., the earliest possible time after its release that processing of the material at the $k$'th resource on its routing can start. Note that the lead times $L_{jk}$ we have used in the previous sections represent a different quantity, the time elapsing between release and capacity consumption. Thus a production order of product $j$ released in period $t$ can consume capacity on the $k$'th workcenter in its routing anywhere in the time interval $[t + L'_{jk},$

$t + L'_{j,k+1}{-}1]$, instead of in period $t + L_{jk}$. Management of the production unit may elect to process portions of a production order in several, not necessarily consecutive, periods, while still ensuring completion of the order within its planned lead time $L_j$. This timing flexibility reflects the possibility of production smoothing within the lead times through scheduling decisions, whereas in the models in the previous sections production quantities are entirely determined by the releases. Models of this kind have been proposed in several different contexts. Pürgstaller and Missbauer (2012) note that the Input-Output Control approach of Wight (1970) implies a model of this form, although the model is not explicitly stated. We have also shown in Chap. 4 that a similar model is implicit in the LUMS order release mechanism for make-to-order production (Hendry et al. 2013). The structure of these models is also related to a much older formulation by Bowman (1956). Spitter et al. (2005) and de Kok and Fransoo (2003) consider a production unit with a single bottleneck workcenter that may consist of a number of parallel machines. In these latter papers, the primary purpose of the model is for supply chain coordination rather than detailed release planning, so they do not directly accommodate modeling of production flows across multiple resources within a production unit. The formulation given below extends these models to incorporate such production flows.

Since the release quantities $R_{jt}$ no longer define the capacity loading of resources in a unique manner, we define additional decision variables $Z^k_{jts}$ specifying the amount of product $j$ released in period $t$ that consumes capacity on workcenter $k$ in period $s$. To ensure that the workcenters in the routing are visited in the correct sequence, we must define these variables to ensure that processing on workcenter $k$ can only take place in the correct time interval such that $t + L'_{jk} \le s \le t + L'_{j,k+1} - 1$ and $1 \le t \le T - L_j$.

Since all materials entering the system must be processed on every workcenter (neglecting details such as scrap or yield losses), we have

$$R_{jt} = \sum_{k \in K} \sum_{s=t+L'_{jk}}^{t+L'_{j,k+1}-1} Z^k_{jts}, \forall j \in J, s = 1,\ldots,T \tag{5.59}$$

Since the processing of a given production order may now be distributed over several periods, all materials associated with the production order released in period $t$ need not necessarily enter finished inventory together. If the production order can enter finished inventory only after the planned lead time has elapsed, irrespective of the actual time(s) the material is processed, the finished inventory balance equations will take the form

$$I_{jt} = I_{j,t-1} + R_{j,t-L_j} - D_{jt}, \quad t = 1,\ldots,T, \quad j \in J \tag{5.60}$$

If, however, material can enter finished inventory as it completes its processing, without having to wait for the remainder of the order, the finished inventory balance equation will be

$$I_{jt} = I_{j,t-1} + \sum_{s=t-L'_{j,K(j)}}^{t} Z^{K(j)}_{jst} - D_{jt}, \quad \forall j \in J, \quad t = 1,\ldots,T \tag{5.61}$$

where *K(j)* denotes the last resource in the process routing of item *j*. Since (5.60) is more consistent with the intent of a planned lead time to ensure availability of the material after the planned lead time with high probability while leaving internal resource allocation decisions to the local management, we shall adopt this assumption from now on. The $R_{jt}$ variables can, of course, be eliminated using (5.59) to reduce the number of variables when solving the model.

The capacity constraints for each workcenter *k* will now take the form

$$\sum_{j\in J}\sum_{s=t-L'_{jk}}^{t} a_{jk}Z^{k}_{jst} \le C_{kt}, \quad t=1,\dots,T-L_j, \quad \forall k\in K \tag{5.62}$$

where the summation on the left hand side represents the total amount of work allocated to workcenter *k* in period *t*. Hence while it is possible to incorporate time-dependent production costs at the different workcenters, if costs are time-stationary there is no need to do so due to the no backlogging assumption. The complete formulation can now be written as

$$\min \sum_{j\in J}\sum_{t=1}^{T-L_j} h_j I_{jt} \tag{5.63}$$

subject to (5.59), (5.60) or (5.61), (5.62) depending on assumptions, and

$$R_{jt} \ge 0, \forall j\in J, t=1,\dots,T \tag{5.64}$$

$$Z^{k}_{jst} \ge 0, \forall j\in J, \forall k\in K, t=1,\dots,T-L_j, s=t-L_{jk},\dots,t \tag{5.65}$$

Since imposing the additional constraint that

$$Z^{k}_{jst} = \begin{cases} R_{jt}, \text{for } s=t-L_{jk} \\ 0, \text{otherwise} \end{cases} \tag{5.66}$$

with $L_{jk}$ denoting the pre-specified lead time in the interval $[L'_{jk}; L'_{j,k+1}-1]$ recovers formulation (5.26)–(5.29), (5.63)–(5.65) is a relaxation of the former in the sense that any feasible solution to (5.26)–(5.29) is feasible for (5.63)–(5.65), but not vice versa. As with formulation (5.26)–(5.29), (5.63)–(5.65) can be rewritten to eliminate the $I_{jt}$ variables giving a model analogous to (5.41)–(5.44).

The formulation until this point has ignored WIP costs. Their inclusion requires some additional thought. If material released at *t* and processed at workcenter *k* cannot move to the next workcenter in its routing until time $t + L'_{j,k+1}$, two different types of WIP may exist at a workcenter: material that has been processed and is waiting to move to the next stage and material that has not yet been processed. If the value of the WIP depends on the timing of production that results from the mode, i.e., earlier production within the lead time means higher WIP holding costs, this can be accounted for by decomposing the WIP at the workcenters into WIP before and WIP after processing and assigning different WIP holding costs to each component.

Denoting $W^{\mathrm{b}}_{jkt}$ and $W^{\mathrm{a}}_{jkt}$ the WIP of product *j* at workcenter *k* at the end of period *t* before and after processing, respectively, the WIP balance equations are

$$W^{\mathrm{b}}_{jkt} = W^{\mathrm{b}}_{j,k,t-1} + R_{j,t-L'_{jk}} - \sum_{s=t-L'_{jk}}^{t} Z^{k}_{jst}, \quad \forall j \in J; \quad t = 1,\ldots,T; \quad k \in K \tag{5.67}$$

$$W^{\mathrm{a}}_{jkt} = W^{\mathrm{a}}_{j,k,t-1} + \sum_{s=t-L'_{jk}}^{t} Z^{k}_{jst} - R_{j,t-L'_{j,k+1}}, \quad \forall j \in J; \quad t = 1,\ldots,T; \quad k \in K \tag{5.68}$$

where in (5.68) $L'_{j,|K|+1} = L_j$. The complete model formulation is given by Pürgstaller and Missbauer (2012). We do not consider this extension in the following example, but note it as an illustration of an issue that the more flexible treatment of lead times may raise.

**Example 5.3** We implement the model (5.63)–(5.65) with the finished inventory balance constraints (5.60) on the problem instance solved in Example 5.1, where we set the values of the $L'_{jk}$ to the $L_{jk}$ values in that example. An optimal solution with objective function value 2505 is obtained as shown in Table 5.8.

The principal difference, as one would expect, lies in the distribution of the capacity loading on Machine 3. Since this machine has a local lead time of $L'_{j4} - L'_{j3} = 4 - 2 = 2$ periods for both products, it is able to allocate capacity across two different periods to releases made in a single period, unlike the previous model where all releases from a given period *t* will load a resource in the single period $t + L_{jk}$ (assuming integer lead times). This difference is seen in Table 5.9 that shows

**Table 5.8** Optimal solution for Example 5.3

| | Releases | | Capacity loading | | | | Ending inventory | |
|---|---|---|---|---|---|---|---|---|
| Period | Item 1 | Item 2 | Machine 1 | Machine 2 | Machine 3 | Machine 4 | 11 | 12 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 25 |
| 1 | 3 | 1.5 | 13.5 | 0 | 0 | 0 | 20 | 25 |
| 2 | 6 | 0 | 18 | 15 | 0 | 0 | 15 | 25 |
| 3 | 6 | 0 | 18 | 18 | 15 | 0 | 11 | 25 |
| 4 | 6 | 0 | 18 | 18 | 18 | 0 | 7 | 23 |
| 5 | 6 | 1.5 | 22.5 | 18 | 18 | 12 | 2 | 19 |
| 6 | 6 | 2 | 24 | 24 | 18 | 12 | 0 | 15.5 |
| 7 | 0 | 5 | 15 | 26 | 18 | 12 | 1 | 10.5 |
| 8 | 0 | 5 | 15 | 20 | 18 | 12 | 1 | 7.5 |
| 9 | 0 | 0 | 0 | 20 | 18 | 18 | 0 | 3.5 |
| 10 | 0 | 0 | 0 | 0 | 18 | 20 | 0 | 2 |
| 11 | 0 | 0 | 0 | 0 | 18 | 20 | 0 | 1 |
| 12 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 5.9** $Z^3_{jst}$ values for Machine 3 in Example 5.3

| | Release | Loading period | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Period | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Item 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 2 | | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 3 | | | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 4 | | | | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 5 | | | | | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 6 | | | | | | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 7 | | | | | | | 0 | 0 | 4 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 8 | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 9 | | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 10 | | | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 |
| | 11 | | | | | | | | | | | 0 | 0 | 0 | 0 | 0 |
| | 12 | | | | | | | | | | | | 0 | 0 | 0 | 0 |
| | 13 | | | | | | | | | | | | | 0 | 0 | 0 |
| | 14 | | | | | | | | | | | | | | 0 | 0 |
| | 15 | | | | | | | | | | | | | | | 0 |
| Item 2 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 2 | | 0 | 0 | 1.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 3 | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 4 | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 5 | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 6 | | | | | | 0 | 0 | 0 | 1.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 7 | | | | | | | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| | 8 | | | | | | | | 0 | 0 | 1 | 4 | 0 | 0 | 0 | 0 |
| | 9 | | | | | | | | | 0 | 0 | 0.5 | 4.5 | 0 | 0 | 0 |
| | 10 | | | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 |
| | 11 | | | | | | | | | | | 0 | 0 | 0 | 0 | 0 |
| | 12 | | | | | | | | | | | | 0 | 0 | 0 | 0 |
| | 13 | | | | | | | | | | | | | 0 | 0 | 0 |
| | 14 | | | | | | | | | | | | | | 0 | 0 |
| | 15 | | | | | | | | | | | | | | | 0 |

the values of the $Z^3_{jst}$ variables for Machine 3. Releases of Product 1 made in period 7 are processed in periods 9 and 10; releases of Product 2 in period 8 are processed in periods 10 and 11, and those from period 9 in periods 11 and 12. Thus in period 11, releases of Product 2 from two different, but consecutive, periods are being processed.

The dual prices associated with this optimal solution are plotted in Fig. 5.9. Note that now both Machines 3 and 4 have binding capacity constraints and hence positive absolute dual prices, in the later periods of the planning horizon. While the

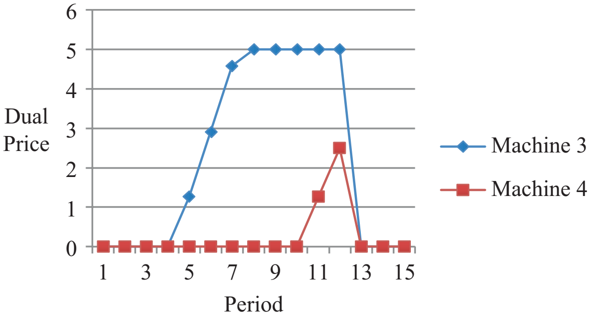


**Fig. 5.9** Dual variables for Example 5.3

formulation in Example 5.1 results in positive dual prices for Machine 3 over the same time interval as in this case, Machine 4 never achieves a positive dual price under the previous formulation.

## 5.7 A Caveat: Lot-Sizing Issues

The models described so far yield the release quantities per product and period $R_{jt}$ as the essential result that is used by the planning level. Executing these decisions in a straightforward manner means releasing production orders of size $R_{jt}$ in the respective periods. However, if the sizes of the production orders are fixed, as is the case when production orders are released to the shop floor by an MRP system that specifies standard lot sizes, the $R_{jt}$ quantities should be viewed as release "budgets" that are filled or consumed by the orders. This is also the case if the model is formulated for aggregate products or product families $j$ with similar routing and resource requirements for the products within one family. Even if the demand $D_{jt}$ is derived from the order sizes, the release quantities need not be a sum of the order sizes due to the capacity constraints and the real-valued $R_{jt}$. In this case the orders to release within the quantities $R_{jt}$ must be determined by a separate planning step. One obvious possibility is to release the orders of product $j$ in period $t$ in the sequence of increasing due date until the cumulative actual release quantity reaches its planned value, that is

$$\sum_{\tau=1}^{t} Actual\ release\ quantity_{j\tau} \leq \sum_{\tau=1}^{t} R_{j\tau},\ \forall j \in J,\ t = 1,\ldots,T \tag{5.69}$$

perhaps with the possibility to exceed the cumulative planned releases by the last order as applied in Load-Oriented Order Release discussed in Sect. 4.2.2.

Alternatively, the release model can be formulated at the level of production orders $p = 1, \ldots, P_j$ with order size, due date, and capacity requirements (setup and processing time) given for each order. Without loss of generality, we assume that the orders are indexed in the order of increasing due dates. The model then determines the period in which each order will be released. The decision variables are

$$\delta_{jpt} = \begin{cases} 1, \text{ if order } p \text{ of product } j \text{ is released in period } t \\ 0, \text{ otherwise} \end{cases}.$$

The release periods are subject to the constraints

$$\sum_{t=1}^{T} \delta_{jpt} = 1, \quad \forall j; \quad \forall p = 1, \ldots, P_j \tag{5.70}$$

$$\sum_{\tau=1}^{t} \delta_{jp\tau} \geq \sum_{\tau=1}^{t} \delta_{j,p+1,\tau}, \quad \forall j,t; \quad \forall p = 1, \ldots, P_j - 1 \tag{5.71}$$

where (5.70) ensures that each order is released exactly once and (5.71) maintains the correct release sequence of the orders. The release quantities can be obtained by

$$R_{jt} = \sum_{p=1}^{P_j} Q_{jp} \delta_{jpt}, \quad \forall j,t \tag{5.72}$$

where $Q_{jp}$ denotes the order size of order $p$ of product $j$.

This modeling technique can be applied to both the conventional fixed lead time model (5.26)–(5.29) and for the alternative model with variable timing of production described in this section, which is described in Missbauer (2014). At the present time, there is no experience with the solvability of the resulting MILP model for real-life problems. Heuristics, e.g., decomposing by product and coordinating the resulting subproblems by Lagrangian techniques or by column generation, are an obvious possibility.

## 5.8 Summary and Conclusions

In this chapter, we have examined the structure of production planning models based on fixed, exogenous lead times that remain constant over the planning horizon. This constitutes the most prevalent mechanism for representing cycle times in both the research literature and industrial practice. We have shown that different models are possible depending on what assumptions are made on the timing of different events, such as when capacity is consumed on specific resources relative to the release time. We have also shown that models with fixed positioning of production within the lead time treat WIP in a rather restrictive manner, assuming WIP cannot accumulate and only a portion of the total WIP in the system is available to be processed by a resource in a given period.

We have also illustrated several limitations of these models relative to the behavior of production resources discussed in Chap. 2. Queueing models show that average cycle time is nonlinear in the average resource utilization, which is directly determined by the work release decisions made by planning models. However, fixed exogenous lead times ignore this relationship, assuming that as long as all capacity constraints are satisfied changes in cycle time due to workload will be negligible. Queueing models also suggest that cycle times begin to degrade well before utilization reaches 1, suggesting there may be benefit to additional capacity at resources whose utilization is below 1. However, our analysis of the dual prices of capacity shows that until a resource is fully utilized, dual prices will be zero, suggesting no benefit from additional resources.

The limited research examining the benefits of more sophisticated models with workload-dependent lead times (Kacar et al. 2012, 2013, 2016) suggests that as long as the average resource utilization remains relatively constant, fixed lead time models with appropriately chosen values of the lead times yield performance very similar to that of much more complex models with workload-dependent lead times. The use of fractional lead times yields a significant improvement over integer lead times, at little additional cost in model complexity. However, when resource utilization and product mix vary significantly over time, the performance of fixed lead time models begins to deteriorate. For this reason, as well as to address the theoretical drawbacks of fixed lead time models discussed above, it is of interest to explore planning models capable of recognizing the nonlinear relation between utilization and cycle time. Put another way, fixed lead time models optimize over releases only; queueing results suggest that jointly optimizing releases and lead times may yield better results. We now explore these more advanced models in the next chapters.

## References

Baker KR (1993) Requirements planning. In: Graves SC, Kan AHGR, Zipkin PH (eds) Handbooks in operations research and management science. Logistics of production and inventory, vol 3. Elsevier Science, Amsterdam, pp 571–627

Bazaraa MS, Jarvis J, Sherali HD (2004) Linear programming and network flows. Wiley, New York

Bertsimas D, Tsitsiklis JN (1997) Introduction to linear optimization. Scientific, Athena

Billington PJ, Mcclain JO, Thomas JL (1983) Mathematical programming approaches to capacity-constrained MRP Systems: review, formulation and problem reduction. Manag Sci 29:1126–1141

Bowman EB (1956) Production scheduling by the transportation method of linear programming. Oper Res 4(1):100–103

de Kok AG, Fransoo JC (2003) Planning supply chain operations: definition and comparison of planning concepts. In: de Kok AG, Graves SC (eds) Handbooks in operations research and management science. Supply chain management: design, coordination and operation, vol 11. Elsevier, Amsterdam, pp 597–675

Hackman S (1990) An axiomatic framework of dynamic production. J Prod Anal 1:309–324

Hackman S (2008) Production economics. Springer, Berlin

Hackman S, Leachman RC (1989a) An aggregate model of project oriented production. IEEE Trans Syst Man Cybern 19(2):220–231

Hackman ST, Leachman RC (1989b) A general framework for modeling production. Manag Sci 35(4):478–495

Hanssmann F, Hess SW (1960) A linear programming approach to production and employment scheduling. Manag Technol 1(1):46–51

Hendry L, Huang Y, Stevenson M (2013) Workload control: successful implementation taking a contingency-based view of production planning and control. Int J Oper Prod Manag 33(1):69–103

Holt CC, Modigliani F, Simon HA (1955) A linear decision rule for production and employment scheduling. Manag Sci 2(1):1–30

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Jacobs FR, Berry WL, Whybark DC, Vollmann TE (2011) Manufacturing planning and control for supply chain management. McGraw-Hill Irwin, New York

Jansen B, de Jong JJ, Roos C, Terlaky T (1997) Sensitivity analysis in linear programming: just be careful! Eur J Oper Res 101(1997):15–28

Johnson LA, Montgomery DC (1974) Operations research in production planning, scheduling and inventory control. Wiley, New York

Kacar NB, Irdem DF, Uzsoy R (2012) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. IEEE Trans Semicond Manuf 25(1):104–117

Kacar NB, Moench L, Uzsoy R (2013) Planning wafer starts using nonlinear clearing functions: a large-scale experiment. IEEE Trans Semicond Manuf 26(4):602–612

Kacar NB, Moench L, Uzsoy R (2016) Modelling cycle times in production planning models for wafer fabrication. IEEE Trans Semicond Manuf 29(2):153–167

Kefeli A (2011) Production planning models with clearing functions: dual behavior and applications. Unpublished Ph.D. Dissertation. Edward P. Fitts Department of Industrial and Systems Engineering. North Carolina State University, Raleigh, NC

Koltai T, Terlaky T (2000) The difference between the managerial and mathematical interpretation of sensitivity results in linear programming. Int J Prod Econ 65:257–274

Leachman RC (2001) Semiconductor production planning. In: Pardalos PM, Resende MGC (eds) Handbook of applied optimization. Oxford University Press, New York, pp 746–762

Leachman RC, Carmon TF (1992) On capacity modeling for production planning with alternative machine types. IIE Trans 24(4):62–72

Manne AS (1957) A note on the Modigliani-Hohn production smoothing model. Manag Sci 3(4):371–379

Missbauer H (2014) From cost-oriented input-output control to stochastic programming? Some reflections on the future development of order release planning models. In: Gössinger R, Zäpfel G (eds) Management Integrativer Leistungserstellung. Festschrift für Hans Corsten. Duncker & Humblot GmbH, Berlin, pp 525–544

Missbauer H, Uzsoy R (2011) Optimization models of production planning problems. In: Planning production and inventories in the extended enterprise: a state of the art handbook. Springer, Boston, pp 437–508

Modigliani F, Hohn FE (1955) Production planning over time and the nature of the expectation and planning horizon. Econometrica 23(1):46–66

Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York

Pochet Y, Wolsey LA (2006) Production planning by mixed integer programming. Springer Science and Business Media, New York

Pürgstaller P, Missbauer H (2012) Rule-based vs. optimization-based order release in workload control: a simulation study of an MTO manufacturer. Int J Prod Econ 140:670–680

Rubin DS, Wagner HM (1990) Shadow prices: tips and traps for managers and instructors. Interfaces 20(4):150–157

Schneeweiss C (2003) Distributed decision making. Springer-Verlag, Berlin

Spitter JM, Hurkens CAJ, de Kok AG, Lenstra JK, Negenman EG (2005) Linear programming models with planned lead times for supply chain operations planning. Eur J Oper Res 163(3):706–720

Vollmann TE, Berry WL, Whybark DC, Jacobs FR (2005) Manufacturing planning and control for supply chain management. McGraw-Hill, New York

Voss S, Woodruff DL (2003) Introduction to computational optimization models for production planning in a supply chain. Springer, Berlin, New York

Voss S, Woodruff DL (2006) Introduction to computational optimization models for production planning in a supply chain. Springer, New York

Wight O (1970) Input-output control: a real handle on lead times. Prod Invent Manag J 11(3):9–31

Zipkin PH (2000) Foundations of inventory management. Burr Ridge, IL, Irwin

# Chapter 6
# Time-Varying Lead Times and Iterative Multi-Model Approaches

The planning models in the previous chapter assume the planned lead times to be workload-independent, exogenous parameters that remain constant over the entire planning horizon. We now consider models with exogenous lead times that vary over time, seeking to accommodate time-varying levels of resource utilization. Since, as discussed in Chap. 2, cycle times depend on capacity utilization, which is determined by release decisions, obtaining time-varying estimates of lead time parameters requires observation or prediction of resource utilization across the time periods in the planning horizon. This tight linkage of utilization and cycle time suggests that releases and lead times should be jointly determined, i.e., the lead times should be endogenous to the model.

We begin this chapter with formulations based on exogenous, time-varying lead times, discuss the issues that arise in estimating these parameters, and then describe order release models that treat time-varying lead times as decision variables linked to the order releases. Noting that many of these formulations result in non-convex optimization models, we then discuss a class of iterative multi-model approaches that have been proposed in the literature.

## 6.1 Preliminaries

It is important to distinguish the problem addressed in this chapter, that of estimating planned lead times to represent cycle times that vary over time, from that of updating existing lead time estimates as new information becomes available from the market and the shop floor. The lead time parameters of MRP systems are reviewed relatively infrequently in practice (Jonsson and Matsson 2006), but must be updated periodically as the production system and its products evolve over time. In this chapter we consider time-varying lead time parameters within a single planning run, so this line of research is not directly relevant. Time-varying lead times are also of interest for due-date assignment (Ioannou and Dimitriou 2012), since the

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_6

state of the shop at the time an order is placed will impact its planned finish date. This is again somewhat different from our problem since we use lead times as input parameters to an order release model that determines the release dates for all orders simultaneously, as opposed to predicting the cycle time of a particular order introduced into the shop at a particular time.

Flow factors or flow allowances, which estimate the lead time associated with an order at a workcenter as a multiple of its processing time, have been widely used for estimating lead times (Keskinocak and Tayur 2004). This approach appears to have originated in the literature on due date setting for make-to-order shops (Keskinocak and Tayur 2004) and has since been widely used in production planning and scheduling. Morton and Pentico (1993) extend this concept to suggest a load-dependent proportionality factor that can be estimated from historical data for lightly, moderately, and heavily loaded shops (p. 218), or by regression from historical data. However, at high levels of resource utilization, cycle times will consist mainly of waiting time in the queues, rendering a proportional relationship between processing and cycle time common to all orders in the shop unlikely except under specific conditions, such as lot sizes that depend strongly on resource utilization, or a sequencing rule that prioritizes jobs with short operation times. Ozturk et al. (2006) apply data mining based on regression tree techniques to this problem.

In Sects. 6.1–6.3 we discuss the representation and modeling of time-varying lead times. Sections 6.4–6.5 then present improved methods to adjust the lead times to the order release plan. In particular, Sect. 6.5 presents methods for iterative adjustment of order releases and time-varying lead times, an approach that has also been proposed for other production planning problems as discussed in Sect. 6.6.

## 6.2 Relaxing the Fixed Lead Time Constraint: Conceptual Issues

In discrete manufacturing systems, the cycle time of production orders at bottleneck resources consists mainly of waiting time in the queues and usually follows a probability distribution with substantial variance. The moments of this distribution, notably its mean, are highly nonlinear functions of the resource utilization as shown in Chap. 2. Planned lead times, which are parameters of the planning system, are derived from these cycle times or their distribution. In MRP, this is accomplished by treating the cycle times as a quantity to be forecast or predicted. In most order release models, planned lead times are obtained by specifying target lead times and controlling the WIP level to ensure that observed cycle times are consistent with these targets via Little’s Law. The definition of “consistent” depends on how cycle time uncertainty is handled in the planning system—this uncertainty is (hopefully) reduced, but not eliminated by load-based order release. If the estimated average cycle time is used as the planned lead time, safety stock or a downstream time buffer can help to manage the uncertainty. The alternative is safety lead time achieved, for example, by setting planned lead times equal to the historical mean cycle time plus a specified safety lead time (Hopp and Sturgis 2000). This approach amounts to setting the planned lead time to some percentile of the underlying cycle time distribution. A number of authors have addressed the problem of determining optimal lead times for different production and inventory systems, including Ben-Daya and Raouf (1994) for inventory systems and Gong et al. (1994) and Milne et al. (2015) for MRP systems.

If the relationship between the cycle time distribution and lead times can be specified, lead time parameters that remain constant over time can be consistent with the steady-state behavior of the production unit. When the aggregate demand faced by the production unit, and hence the average utilization of its bottleneck resources, exhibit little variation over time, this approach is likely to be quite satisfactory. However, if demand varies widely over time, even if the release model has some load-leveling capability, the releases, and thus the work input to the resources and their utilization, may also vary over time, and the constant lead times will not match the actual cycle times. This issue can arise due to both the total demand for all products varying over time and the time-varying demand for individual products with different production routings and resource requirements.

Inconsistency between constant lead times and load-dependent cycle time distributions causes two distinct difficulties. On the one hand, the lead times must allow high bottleneck utilization, which requires high WIP levels, high average cycle times, and thus a high value of the planned lead time. A temporary decrease in demand will lead to reduced releases, work input, and resource utilization, resulting in shorter cycle times. Material will be released earlier than is necessary to meet demand, causing unnecessarily high FGI levels. On the other hand, temporarily increasing releases, and hence WIP levels between workcenters, to improve load smoothing is not possible since this would raise realized lead times above the planned lead time. Directly addressing the latter issue within the release model requires estimates of lead times in the face of time-varying demand, either through a separate planning module that estimates the lead time parameters to be used in the release model or within the release procedure itself. The latter requires representation of time-varying lead times in the order release model, either explicitly as decision variables or implicitly as time-varying WIP, leading to additional complications discussed in the next two chapters.

To illustrate the issues that arise when considering time-varying lead times, consider the following example.

**Example 6.1** The fixed lead times associated with the orders released in each period are given in Table 6.1 for 12 consecutive planning periods. We make no pretense that these lead times are realistic in any way; our purpose is to illustrate the issues that arise in selecting time-varying lead time estimates. The reader will note that the lead times increase and decrease by substantial jumps, with some being fractional and others integer.

Table 6.2 shows the loading factors that represent the fraction of material released in period $\tau$ that will emerge in period $t$ based on these lead time estimates. The lower diagonal is, as expected, empty since a positive entry in this area would imply a

**Table 6.1** Lead time parameters for Example 6.1

| Period | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lead time | 1 | 2 | 2.5 | 2.5 | 3 | 4.5 | 3 | 2.5 | 2.5 | 2 | 1.5 | 1 | 1 |

**Table 6.2** Loading fractions for Example 6.1

| Period | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 13 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

negative lead time, which, although it might be welcomed by many manufacturing managers, is difficult to achieve. There are, however, several areas of interest above the diagonal. No output emerges at all in period 8, due to the long lead times in preceding periods. Material released in period 4 emerges in period 7, but material released in period 5 emerges in periods 9 and 10. All materials released in period 6 emerge in period 9. However, half the material released in period 5 emerges in period 9 and the other half in period 10, indicating that the releases from period 5 are being overtaken by those from period 6.

This example illustrates that unless time-varying lead times are selected with some care, they can lead to quite unrealistic behavior in a planning model. It is thus useful to seek conditions to impose on lead time estimates that will ensure reasonable behavior of the planning models in which they are deployed. One such requirement would seem to be that of no-passing, or first-in-first-out (FIFO): material released in earlier periods should not emerge from the system before material that is released later. In the dynamic traffic assignment literature (Peeta and Ziliaskopoulos 2001), this implies no overtaking: vehicles entering a road segment at a point in time cannot exit before those entering earlier.

Carey (1992) examines several such conditions in the context of the dynamic traffic assignment problem, focusing on the need to preserve the FIFO property, in our context to ensure that material that is released earlier does not emerge after material released later. He first considers the case of a single product where $x_{ts}$ represents the amount of the product arriving at the resource in period $t$ and completing its processing in period $s$. Thus the amount of material $x_{ts}$ will remain at the resource for $(s - t)$ periods, and the average time a unit of work arriving in period $t$ will spend at the resource will be given by

$$\bar{m}_t = \frac{\sum_{\tau \ge t} x_{t\tau}(\tau - t)}{\sum_{\tau \ge t} x_{t\tau}} \tag{6.1}$$

Note that since it represents an average, the value of $\bar{m}_t$ need not be an integer. To maintain FIFO on the basis of the average flows, material that arrives at the resource in period $t$ must exit by period $t + \bar{m}_t$. Thus, to ensure that on average material arriving later exits later, material entering in a later period $s > t$ must exit in period $s + \bar{m}_s$. Thus to preserve FIFO on average, we must have

$$\bar{m}_t \le \bar{m}_s + (s - t), \quad \text{for} \quad s \ge t \tag{6.2}$$

yielding

$$t + \bar{m}_t \le s + \bar{m}_s, \quad \text{for all} \quad s \ge t \tag{6.3}$$

Since $(s - t) \ge 1$, this implies that the constraints

$$\bar{m}_t \le \bar{m}_{t+1} + 1 \tag{6.4}$$

are necessary and sufficient to ensure FIFO for the average flows, although, as he shows by a counterexample, necessary but not sufficient for the individual components $x_{ts}$. Note that in this representation, the planned lead times are not represented explicitly as a parameter, but through the definition of the decision variables $x_{ts}$, with $\bar{m}_t$ defined as in (6.1). The explicit inclusion of condition (6.4) in an optimization formulation with a planning horizon of $T$ periods requires $O(T^2)$ non-convex constraints, resulting in a model that is significantly more difficult to solve. He goes on to show that analogous conditions are necessary and sufficient for the average flow in the presence of multiple vehicle classes, analogous to multiple products in our context, and necessary but not sufficient to maintain FIFO at the level of individual items. This necessary condition plays an important role in the formulation of the Allocated Clearing Function model in Chap. 7 and will be revisited in that context. However, Carey's findings are, in general, discouraging: they show that a variety of approaches to maintain the FIFO property all lead to planning models with non-convex feasible regions.

## 6.3 Modeling Time-Varying Lead Times

We can distinguish two different types of planned lead times for a single workcenter using a continuous representation of time and orders as seen in Fig. 6.1. The *forward* lead time $L^f(t)$ represents the lead time of an order that arrives at time $t$, i.e., the estimated time spent in the workcenter by an order arriving at time $t$. Similarly, the *backward* lead time $L^b(t)$ represents the planned amount of time spent in the

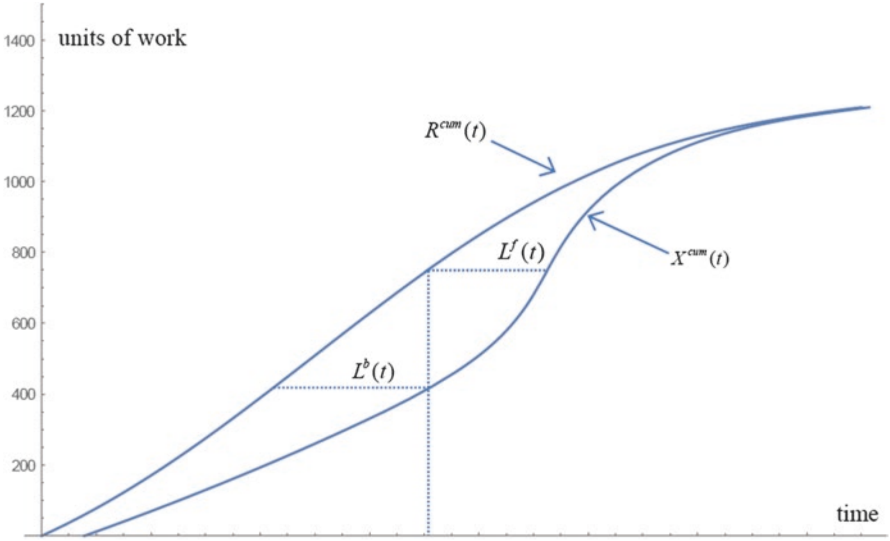


**Fig. 6.1** Evolution of forward and backward lead time over time

workcenter by an order leaving the workcenter at time $t$. In other words, a unit of work arriving at the workcenter at time $t$ departs at time $t + L^{\text{f}}(t)$, while one departing at time $t$ must have arrived at time $t - L^{\text{b}}(t)$.

Following our previous notation, let $R(t)$ denote the rate of material release into the workcenter at time $t$, and $X(t)$ its output rate at time $t$. We shall denote the cumulative releases and output up to $t$ by $R^{\text{cum}}(t)$ and $X^{\text{cum}}(t)$, respectively, and let $W(t)$ denote the planned WIP level at time $t$. If we require the FIFO or no-passing property, under which work released at time $t$ cannot complete before work released at any time $s < t$ and production orders (or work particles in the continuous representation) depart the workcenter in the same sequence as they arrive, the cycle times are determined by the evolution of WIP over time. Based on Fig. 6.1, we have the material balance relations

$$W(0) + \int_0^t R(\tau)\,d\tau = \int_0^{t+L^{\text{f}}(t)} X(\tau)\,d\tau \tag{6.5}$$

$$W(0) + \int_0^{t-L^{\text{b}}(t)} R(\tau)\,d\tau = \int_0^t X(\tau)\,d\tau \tag{6.6}$$

which calculate the time-dependent output of the workcenter from its time-dependent input, constituting a *dynamic production function* (Hackman 2008). Equation (6.5) states that all materials entering the system by time $t$ must, by the definition of $L^{\text{f}}(t)$, have been converted into output by time $t + L^{\text{f}}(t)$. Similarly, (6.6) states that all materials leaving the system by time $t$ must have entered by time $t - L^{\mathrm{b}}(t)$. Hence $L^{\mathrm{f}}(t)$ and $L^{\mathrm{b}}(t)$ are related as

$$\begin{aligned} L^{\mathrm{b}}(t) &= L^{\mathrm{f}}\left(t - L^{\mathrm{b}}(t)\right) \\ L^{\mathrm{f}}(t) &= L^{\mathrm{b}}\left(t + L^{\mathrm{f}}(t)\right) \end{aligned} \tag{6.7}$$

Extending this logic to discrete-time models is not straightforward. The simplest analogy is period-based, integer lead times $L_t^{\mathrm{f}}$ and $L_t^{\mathrm{b}}$ representing the lead times of orders arriving or departing in period $t$, respectively. Thus the fixed lead time formulation in Chap. 5 represents a backward lead time implying

$$R_{t-L_t^{\mathrm{b}}} = X_t \tag{6.8}$$

This is perfectly adequate when $L_t^{\mathrm{b}} = L_{t+1}^{\mathrm{b}}$ for all periods $t = 1,..., T - 1$ in the planning horizon; each unit of work emerging as output at any time within period $t$ was released exactly $L_t^{b}$ time units earlier. However, if the planned lead time at the workcenter increases by 1 period from period $t$ to period $t$+1 such that $L_{t+1}^{\mathrm{b}} = L_t^{\mathrm{b}} + 1$, (6.8) implies that the output of two or more consecutive periods was released in the same period, as was the case for period 9 in Example 5.1. Hence, (6.8) must be formulated as an inequality constraint of the form

$$\sum_{k=1}^{t-L_t^{\mathrm{b}}} R_k \geq \sum_{k=1}^{t} X_k \tag{6.9}$$

for all $t$, which only gives a lower bound on the releases or, expressed in terms of time, the latest possible release period for given output over time. Thus it represents time-varying lead time parameters only in the context of a release model that delays releases as much as possible, usually due to positive WIP holding costs in the objective function (as in the release models in Chap. 5). This is the first shortcoming of representing lead times directly as parameters $L_t^{\mathrm{f}}$ or $L_t^{\mathrm{b}}$.

A second problem is that this representation cannot express lead time distributions. Empirical cycle time distributions often exhibit high coefficients of variation as seen in Fig. 2.3, and an effective planning model should be able to represent this. One approach to representing lead time distributions is the use of *loading factors* $w_{\tau t}$ defined as the fraction of the work released in period $\tau$ that emerges as output in period $t$.

A backward lead time $L_t^{\mathrm{b}}$ can be converted into a loading factor by noting that

$$w_{\tau t} = \begin{cases} 1, & \text{if } L_t^{\mathrm{b}} = t - \tau \\ 0, & \text{otherwise} \end{cases} \tag{6.10}$$

yielding the relationship between releases and output as

$$R_t = \sum_{\tau=t}^{T} X_\tau w_{t\tau} \tag{6.11}$$

The loading factors can be interpreted as the expected fraction of work released in a certain period that leaves the workcenter after a certain time, representing a discrete probability distribution for lead times. In (6.11), and most of the iterative approaches discussed below, this expectation is treated as a deterministic fraction, resulting in a set of linear constraints.

## 6.4 Epoch-Based Lead Times

Until now we have assumed period-based lead times such that lead times are associated with specified planning periods, implying that that all releases (for forward lead times) or output (for backward lead times) associated with a period is subject to the same lead time. Hung and Leachman (1996) suggest the use of epoch-based lead times defined at the period boundaries, which permits a more general representation of fractional lead times. We now describe this approach since it forms the basis for many of the iterative approaches in Sect. 6.5.

The basic formulation is derived from the model discussed in Chap. 5, which requires lead time estimates $L_{jk}$ representing the time required for a unit of product $j$ to reach the $k$'th resource in its product routing after being released into the plant. However, instead of fixed lead times that remain constant over the entire planning horizon, Hung and Leachman (1996) associate lead time parameters with the start of each planning period. In the following we shall assume unit-length planning periods such that period $t$ starts at time $t-1$, i.e., $t=0$ is the start of period 1, $t=1$ the start of period 2, etc. Equivalently, this can be viewed as period $t$ ending at time $t$. The lead time parameters $L_{jkt}$, which may take on fractional values, represent the lead time after its release required for an order of product $j$ to reach the $k$'th resource on its routing if the order reaches that resource at the end of period $t$. This definition of epoch-based lead time parameters is depicted in Fig. 6.2. The key assumption is that the releases associated with a planning period take place at a uniform rate over the planning period, as discussed in Chap. 5 and in Hackman and Leachman (1989).

Given these lead times, the loading of the production resource in period $t$ is defined by releases occurring in the time interval $Q_t = [(t{-}1) - L_{j,k,t\text{-}1}, t{-}L_{jkt}]$, recalling that planning period $t$ starts at time $(t{-}1)$ and ends at time $t$. There are two cases to consider here. In the first, simpler case, the time interval $Q_t$ lies within a single planning period $\lceil (t-1) - L_{j,k,t-1} \rceil = \lceil t - L_{jkt} \rceil$ where $\lceil x \rceil$ denotes the smallest integer greater than or equal to $x$. In this case the entire amount released in period $\lceil (t-1) - L_{j,k,t-1} \rceil$ arrives at workcenter $k$ in period $t$. Hence the amount $Y_{jkt}$ of product $j$ loading workcenter $k$ in period $t$ is given by

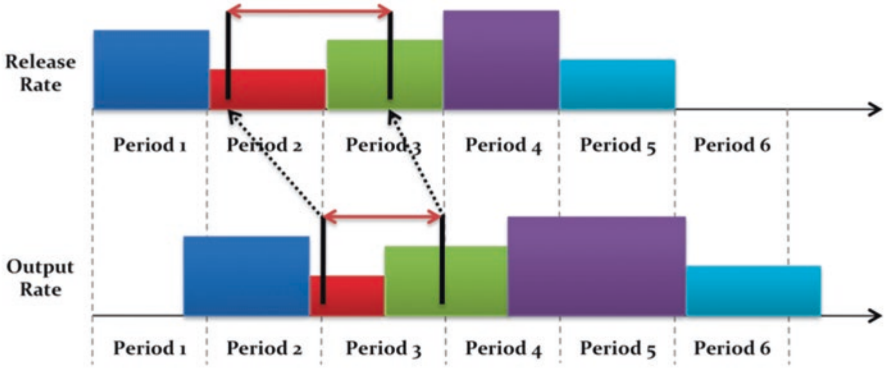


**Fig. 6.2** Relationship between releases and loading with time-dependent lead times. Adapted from Hung and Leachman (1996)

$$Y_{jkt} = \left(\frac{\left(t-L_{jkt}\right)-\left(\left(t-1\right)-L_{j,k,t-1}\right)}{\Delta}\right)e_{jk}R_{j,\left\lceil (t-1)-L_{j,k,t-1}\right\rceil} \tag{6.12}$$

where Δ denotes the period length (set to 1 by definition) and $e_{ij}$ the average fraction of the release quantity of product $j$ that will arrive at resource $k$.

If, on the other hand, the time interval $Q_t$ spans multiple planning periods, we allocate the load due to releases in that period in proportion to the fraction of that period's total duration included in the interval $Q_t$ assuming uniform release rates over the planning periods, yielding

$$\begin{aligned} Y_{jkt} = & \left(\frac{\left(\left\lceil \left(t-1\right)-L_{j,k,t-1}\right\rceil-\left(t-1\right)-L_{j,k,t-1}\right)}{\Delta}\right)e_{jk}R_{j,\left\lceil (t-1)-L_{j,k,t-1}\right\rceil} \\ & + \sum_{\tau=\left\lceil (t-1)-L_{j,k,t-1})\right\rceil+1}^{\left\lceil t-L_{jkt}-1\right\rceil} e_{jk}R_{j\tau} \\ & + \left(\frac{\left(t-L_{jkt}\right)-\left\lceil t-L_{jkt}-1\right\rceil}{\Delta}\right)e_{jkt}R_{j,\left\lceil t-L_{jkt}\right\rceil} \end{aligned} \tag{6.13}$$

The operation of this approach is illustrated in Fig. 6.2. The upper part of the figure shows the uniform release rates in each planning period, and the lower portion the resource loading resulting from these releases arriving at the resource after the specified fixed lead times. Releases in periods 2 and 3 contribute to the work input in period 3 at resource $k$ corresponding to the first and the third term in (6.13); the second term is not relevant here because the release interval only spans the two periods 2 and 3. Due to the use of backward lead times, the lead times are associated with the boundary points between periods at the workcenter, not those between the release periods, and hence the lead time at the start of a period may not be the same as that at the end. The coloring indicates the correspondence between the releases and the arrival of the material at the resource.

The loading factors $w_{j\tau t}$ that denote the fraction of releases of product $j$ in period $\tau$ that contribute to output in period $t$ follow immediately from (6.12) or (6.13), depending on the case. The amount $Y_{jt}$ of product $j$ arriving at the workcenter in period $t$ is, analogously to (6.11), given by the linear expression

$$Y_{jt} = \sum_{\tau=1}^{t} R_{j\tau} w_{j\tau t} \tag{6.14}$$

If we could obtain the correct values of the loading factors $w_{j\tau t}$ efficiently, we would no longer need an explicit capacity constraint since the loading factors would reflect the ability of the resource to produce output over time. However, formulating and solving a model that encompasses both order release planning and estimation of the $w_{j\tau t}$ values turns out to be challenging, as we discuss below.

## 6.5 Lead Time Estimation Within the Order Release Procedure

The previous section described different ways to represent time-varying lead times in an order release model, assuming these were treated as exogenous parameters. We now turn to the crucial question of how to specify values for these lead times, that is, how to represent the functional relationship between capacity loading and lead times. This can be handled in two fundamentally different ways:

- Time-varying lead times can be treated as exogenous parameters whose values are determined based on information known prior to order release, such as historical flow times, capacity, and demand. Orders are then released based on these lead time parameters.
- The order release model, or the order release procedure of which it is a part, can treat the lead times as functionally related to the release schedule and hence must represent this functional relationship. This can, in turn, be accomplished in two ways:
    - The lead times can be defined as decision variables *endogenous* to the optimization model, which optimizes releases and lead times simultaneously.
    - The problem can be decomposed into two related subproblems: one that determines an optimal release schedule given the estimates of lead times, and another that estimates lead times based on a given release schedule. An iterative procedure then solves these subproblems in sequence until some convergence criterion is satisfied.

The first approach, that of setting lead times prior to order release, has been treated extensively in the MRP literature. “MRP treats lead times as attributes of the part and possibly the job, but *not* of the status of the shop floor” (Hopp and Spearman 2008: 124). Planned lead times “serve as a proxy for dealing with capacity constraints; a longer planned lead time leads to a longer planned queue that permits more production smoothing" (Graves 2011: 93). As described in Chap. 2, using this approach, lead times are usually estimated from historical cycle times and are updated only infrequently. Due to the importance of planned lead times for manufacturing performance, there has been extensive research on improving lead time estimation (Milne et al. 2015). Since the role of the lead times in this framework is to coordinate the various planning levels of the PPC system, their values are determined by a parameter-setting function that seeks to ensure this coordination (see Chap. 1). Since the cycle times are closely related to capacity utilization, they must be coordinated with the production smoothing decisions made at the master production scheduling level, but we are not aware of any research on jointly determining the master production schedule and time-dependent lead times.

The principal difficulty in determining time-varying lead times derives from the nonlinear relation between cycle time and resource utilization described in Chap. 2. Since cycle times depend on resource utilization, and resource utilization on the release decisions, determining time-varying lead times for use in an order release model requires knowledge of capacity loading over time, and hence of the order releases, at least at an aggregate level. If the aggregate level of capacity loading, and hence resource utilization, remains largely constant over time, this may not be a major issue. However, even this may be moot at high utilization levels, where small changes in resource utilization may lead to large changes in cycle times. Directly addressing this interdependence between lead time estimates and release decisions requires models that simultaneously determine time-varying lead times and order releases, which we discuss in the following section.

### *6.5.1 Models Without WIP Evolution*

The evolution of cycle times over time is closely related to the evolution of WIP over time as expressed in (6.5) and (6.6); Hackman (2008: 309ff.) gives a more detailed discussion. For complex manufacturing systems modeled as networks of queues that need not be in steady state, models that accurately anticipate the evolution of WIP and cycle time over time generally involve some type of simulation, either of the discrete-event type (Law and Kelton 2004) or continuous-time models based on ordinary or partial differential equations (Armbruster and Uzsoy 2012), which are difficult to incorporate into a tractable mathematical programming model. Therefore, a number of approaches that estimate load-dependent lead times within the order release model without explicitly considering the evolution of WIP over time have been proposed.

Since in steady state the average cycle time increases nonlinearly with utilization, it is intuitively appealing that this pattern also holds within each planning period of an order release model. A number of authors have developed models that select an appropriate lead time for each planning period based on the resource loading in that period. These are closely related to those developed for dynamic traffic

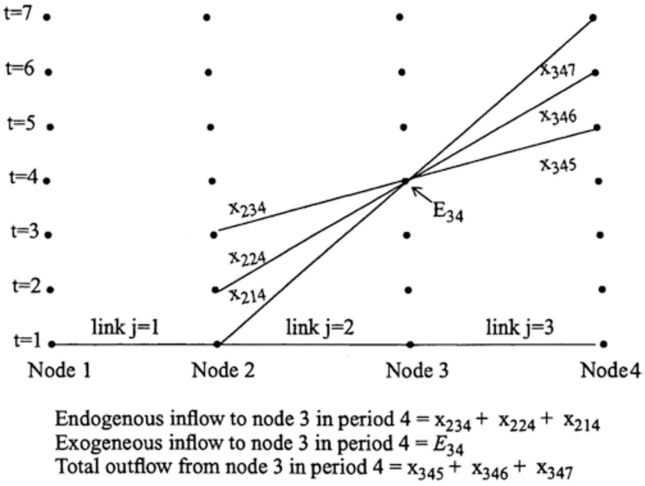

t=7
t=6
t=5
t=4
t=3
t=2
t=1
x347
x346
x345
E34
x234
x224
x214
link j=1
link j=2
link j=3
Node 1
Node 2
Node 3
Node4
Endogenous inflow to node 3 in period 4 = x234 + x224 + x214
Exogeneous inflow to node 3 in period 4 = E34
Total outflow from node 3 in period 4 = x345 + x346 + x347

**Fig. 6.3** Conservation of flows on a time-expanded network (Carey and Subrahmanian 2000)

assignment models (Peeta and Ziliaskopoulos 2001) that seek to determine the routing of vehicles through a road network to optimize some measure of performance. Since individual traffic links (road segments) are subject to congestion, considerable effort has been devoted to developing models that capture the relationship between the volume of flow on a traffic link and the velocity of that flow.

One way to model congestion in traffic networks is through the use of *time–space links* (Carey and Subrahmanian 2000). If two nodes *i* and *j* of a traffic network are connected by a spatial link (in other words, a road segment), this two-node network can be expanded over time to yield a network of time–space nodes as seen in Fig. 6.3. The flow on a time–space link represents the number of vehicles that pass the nodes at the times corresponding to the nodes at the end points of the link and hence requires the associated (integer) traversal time.

The impact of congestion is manifested as a link traversal time that increases with the volume of flow on the link and can be represented by having the upper bounds on the flow through the time–space links leaving node *i* at time *t* depend on the flow through node *i* at time *t*, i.e., the inflow to the time–space links leaving node (*i,t*). In the model of Carey and Subrahmanian (2000), the capacities of at most two neighboring time–space links leaving node (*i,t*) are positive and the other time–space links are closed for the given inflow. As the inflow increases, the time–space links with positive capacities move to higher traversal times, implying a flow-dependent traversal time distribution that is stationary over time for a given inflow. The relationship between the flow *x* through the time–space link and the time *s* it takes to traverse the link, referred to as the travel time function *s* = *f(x)*, is assumed to be convex, increasing, and piecewise linear, which allows the breakpoints of the function to be mapped onto the time–space links as in Fig. 6.4. If the inflow *x* is exactly at a breakpoint, only the corresponding time–space link is active. Otherwise

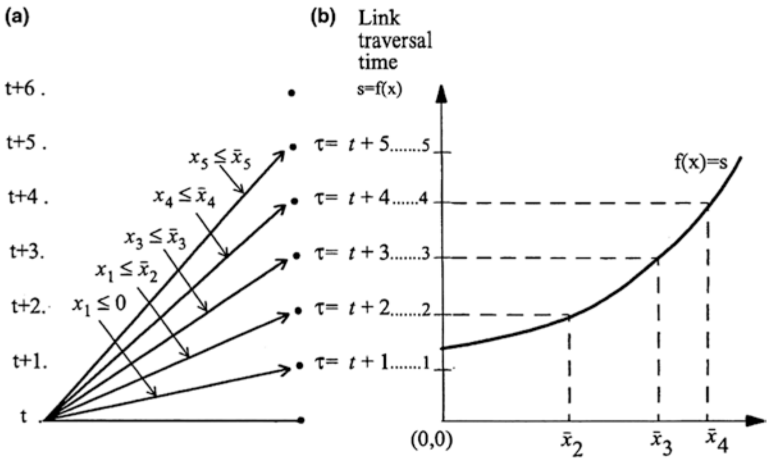


**Fig. 6.4** Relationship between time–space link capacities (left) and travel time function (right) (Carey and Subrahmanian 2000)

the respective fractions of the inflow are assigned to the two adjacent time–space links (Carey and Subrahmanian 2000: 163). The authors consider a piecewise linear convex objective function and develop two alternative formulations based on this representation. Using concepts from separable programming (Bazaraa et al. 1979), they show that under these assumptions at most two adjacent time–space links will carry positive flow. They also show that as long as there is no holding back behavior, where traffic that has entered a link is not allowed to exit in order to alleviate congestion in later periods, the solutions will satisfy FIFO unless a sharp increase in inflow is followed by a sharp decrease. When holding back occurs, however, solutions may violate the FIFO property.

The traversal time of a spatial link in a traffic network is analogous to the cycle time at a workcenter, and models with similar structure have been developed for order release planning in manufacturing. Voss and Woodruff (2003) assume a steady-state relationship between workcenter utilization and the expected cycle time at that workcenter. They then discretize this curve using integer variables to ensure that only one segment of the discretized curve is active in a given time period. The relationship between utilization and expected lead time is evaluated at discrete utilization levels (breakpoints) $BP_q$, $q = 1,\ldots,U$ where $L^q$ denotes the expected lead time value associated with the $q$'th utilization level $BP_q$. Thus the expected lead time of the resource is assumed to be $L^q$ when its utilization level is between $BP_q$ and $BP_{q-1}$. The authors suggest setting the breakpoints $BP_q$ such that each lead time $L^q$ corresponds to an integer number of periods. If $a_j$ denotes the fraction of the available resource capacity required for one unit of product $j$, $j = 1,\ldots,P$, and $R_{jt}$ the amount of product $j$ released in period $t$, the utilization of the resource in period $t$ is given by

$$u_t = \sum_{j=1}^{P} a_j R_{jt} \tag{6.15}$$

We now define binary variables $y_{tq}$ that select a particular lead time value $L^q$ to be active in a given period $t$ as follows:

$$L_t = \sum_{q=1}^{U} y_{tq} L^q \text{ for all } t \tag{6.16}$$

$$\sum_{q=1}^{U} y_{tq} = 1, \text{for all } t \tag{6.17}$$

Additional constraints of the form

$$\sum_{q=1}^{U} BP_q y_{tq} \geq \sum_{j=1}^{P} a_j R_{jt}, \text{for all } t \tag{6.18}$$

ensure that the lead time selected is consistent with the workload. In addition, for any given period $t$, we require $L_t - L_{t+1} \leq 1$, giving

$$\sum_{q=1}^{U} y_{tq} L^q - \sum_{q=1}^{U} y_{t+1,q} L^q \leq 1, \quad \text{for all} \quad t \tag{6.19}$$

This latter constraint is interesting in that it restricts the changes in lead time from one period to the next to at most one period to avoid overtaking, i.e., material released into the system at a later time emerging before material released earlier. Note that (6.19) enforces the condition (6.4) shown by Carey (1992) to be necessary for the flow through a node to satisfy the first-in-first-out (FIFO) condition.

To complete the formulation, the authors present an objective function that includes an explicit holding cost for WIP, based on Little's Law (Hopp and Spearman 2008), leading to

$$\min \sum_{t=1}^{T} \sum_{j=1}^{P} h_{jt} \sum_{q=1}^{U} y_{tq} L^q R_{jt} \tag{6.20}$$

This objective function is nonlinear due to the product of the $y_{tr}$ and $R_{jt}$, leading to a formulation that is computationally hard to solve.

Lautenschläger (1999) describes a similar approach. In order to consider load-dependent lead times for master production scheduling, this model determines the fraction of the planned production available in a period $t$ that has to be started one period ahead in period $(t - 1)$ assuming the rest is produced in period $t$. This fraction is a function of the planned utilization. Thus production on a resource can be performed in two modes, one with lead time of zero periods and the other with lead time of one period, essentially the same idea as the time-expanded network in Fig. 6.3. The maximum production volumes that can be realized in each mode are limited, leading to a utilization-dependent lead time distribution. Short-term oscillations in capacity utilization over time, which are considered undesirable due to considerations not explicitly represented in the model, are reduced by a low-pass filter (Lautenschläger 1999: 114ff.). Many factory managers consider large variations in utilization to be detrimental to performance, perhaps due to their impact on staffing and other support services such as material procurement (Lautenschläger 1999: 114f). However, the high-frequency oscillations may also be due to the simplifications in the flow time modeling. Orcun and Uzsoy (2011) have shown that inconsistencies between the lead times used in a planning model and the cycle times in the production system can lead to significant oscillating behavior when the planning model is implemented in a rolling horizon environment, supporting the latter conclusion.

### *6.5.2 Critique of Lead Time Estimation Without WIP Evolution*

While the models in the previous section address the load-dependent nature of lead times, they ignore the relationship between time-dependent lead times and WIP evolution over time expressed in (6.5) and (6.6) and formulated more generally in transient versions of Little's Law (Bertsimas and Mourtzinou 1997; Riaño 2003) in order to obtain a tractable mathematical programming formulation. As such, they must be viewed as approximations that exhibit several shortcomings:

- All the models described above assume a well-defined relationship between the workload or utilization of a resource in a planning period and its expected cycle time in that period. The form of this relationship is generally posited assuming steady-state is reached by all related queues during the planning period. However, since planning models assume discrete planning periods of a fixed length and work releases vary over time, planning models inherently operate in a transient regime, and the cycle time of work released in a given period may deviate quite substantially from the long-run steady-state average.
- If the amount of work released decreases sharply from period $t$ to period $t$+1, the estimated lead time for the orders can decrease by more than one period from $t$ to $t$+1, implying overtaking (Voss and Woodruff 2003: 165; Carey and Subrahmanian (2000)). This is unlikely to occur in practice—although it may be accomplished to a limited extent by expediting, which has its own disadvantages (Ehteshami et al. 1992; Narahari and Khan 1997)—and violates the assumption that the released work must be processed first-in-first-out. This suggests that these models can lead to unrealistic results. Voss and Woodruff (2003) add a constraint that keeps the lead time from decreasing by more than one period from $t$ to $t$+1, which Carey (1992) has shown is a necessary condition for the preservation of the FIFO property.

Several researchers have sought to address these problems by using either a discrete-event simulation model or a transient queueing model to model the joint evolution of lead times and WIP levels. This leads to computationally intractable optimization models, requiring lead time estimation to be performed outside the optimization model. This approach will be discussed in the next section.

## 6.6 Lead Time Estimation Outside the Optimization Model: Iterative Multi-Model Approaches

### *6.6.1 Overview*

Modeling the joint evolution of lead times and WIP levels in a transient setting usually leads to computationally intractable order release models even in simple cases. This can be seen from (6.5) and (6.9) where the lead times are elements of the integration or summation limits. However, this structure can be addressed by decomposing the order release problem into two separate subproblems: one that computes a release plan given a set of time-varying lead time estimates and another that computes the expected lead times or output associated with each period, or boundary between periods, for a given release plan. These are usually deployed within an iterative framework that seeks convergence to a pair of consistent subproblem solutions. A review of multi-model approaches combining optimization and simulation is given by Figueira and Almada-Lobo (2014).

The central difficulty of multi-model approaches that decompose the release planning problem into separate release planning and lead time estimation problems is that of any decomposition procedure: that of efficiently achieving a solution simultaneously satisfying the constraints of both subproblems. In isolation, both subproblems can be addressed satisfactorily with well-known techniques. The release planning subproblem can be solved directly by the LP models described in Chap. 5, whose mathematical structure easily accommodates time-varying lead time estimates as long as reasonable estimates can be obtained as discussed in Sect. 6.1. The lead time estimation subproblem can be addressed by queueing or simulation models. What is required is a coordination mechanism that leads to mutually consistent solutions to the two subproblems that are at least feasible, and hopefully near-optimal, to the overall problem. In order to preserve the tractability of the release planning subproblem, its parameters (capacities and lead times) must be exogenous to whatever model is used to solve it, i.e., unaffected by the release schedule it produces. Similarly, the lead time/output estimation subproblem must treat the release schedule as an exogenous input. Hence these procedures combine mathematical programming and simulation or queueing models such that each model determines estimated values of parameters required by the other. Since the primary optimization mechanism is embedded in the mathematical programming model, the simulation or queueing model used for lead time estimation is subordinate

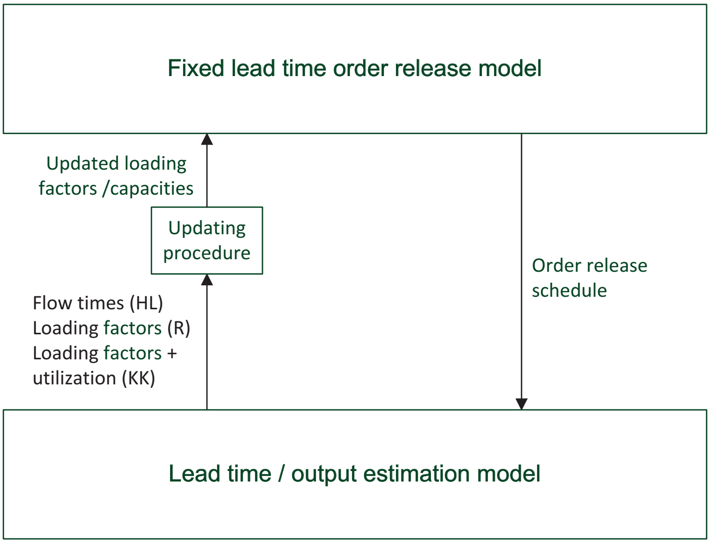

Fixed lead time order release model
Updated loading factors /capacities
Updating procedure
Order release schedule
Flow times (HL)
Loading factors (R)
Loading factors + utilization (KK)
Lead time / output estimation model

**Fig. 6.5** Iterative simulation—LP approach for order release planning: the generic mechanism. HL: Hung/Leachman, KK: Kim/Kim, R: Riaño

to the optimization model per the taxonomy of hybrid simulation/analytic models by Shanthikumar and Sargent (1983). The procedure is outlined in Fig. 6.5.

How the parameters of each model are updated based on the results of the other is likely to have significant impact on both the convergence of the procedure and the quality of the solution to which it converges. The communication from the release planning model to the lead time estimation model is generally straightforward: the quantity of each product released in each period. The information passed from the lead time/output estimation model to the release planning model usually consists of the estimated mean cycle times associated with each period or epoch, while some approaches also consider average resource utilization levels in each period. The cycle times observed by the lead time estimation model may be represented in the release planning model as exogenous lead times or loading ratios as described in Sect. 6.2.

At each iteration, the current estimates of cycle times or loading ratios and utilization are used to update the lead times and capacities that constitute the parameters of the order release model, and the order release model is re-run. This iterative procedure is repeated until convergence, which can be defined as reaching a *fixed point* of the iterative mechanism, a solution where the parameters of the order release model lead to an order release schedule that results in the same cycle time and output estimates by which the order release schedule was produced. Thus, once the algorithm arrives at this solution, it remains there. Ideally, the optimal solution should be a fixed point, but there is as yet no rigorous proof that this is the case in general. There is considerable experimental evidence that the solution spaces of some formulations of this problem are non-convex, leading to the procedure converging to different points from different initial solutions. Experimental evidence discussed later in the chapter suggests that even quite subtle differences in implementation may produce qualitative differences in computational behavior.

### 6.6.2 Iterative Simulation-LP Algorithms

Zaepfel (1984) was the first author to formulate such an iterative mechanism and its associated order release model. In his procedure only the estimated lead times are communicated from the lead time estimation (simulation) model to the order release model which assumes fixed lead times and unlimited capacities. The reasoning is that since overloading of capacities leads to higher flow times, information on capacity overload (excessive releases in certain periods) is captured in the revised lead time estimates in the feedback from the simulation model. No numerical results are provided.

Hung and Leachman (1996) were the first to provide numerical tests of this type of iterative scheme. Their order release model modifies the step-separated formulation of Leachman and Carmon (1992) to represent the lead times as loading factors per Sect. 6.2, with epoch-based backward lead times defined at the period boundaries as in Fig. 6.2. Updating these lead time parameters during the iterations requires observing the simulated flow times at the period boundaries, which are interpolated from the flow times of orders arriving at the workcenters immediately before and immediately after the boundary epoch (Hung and Leachman 1996: 262). The order release model includes capacity constraints and assumes that capacity is consumed at the end of the planned lead time. The release period determines the period in which the work is processed and capacity is required. Hung and Leachman (1996) examine the rate of convergence of the flow time estimates to the flow times observed in the simulation and find that convergence to the correct expected flow time values can be quite rapid but that the procedure can fail to converge in some cases which are not fully understood. Subsequent numerical tests by other authors (Irdem et al. 2010; Kacar et al. 2012) confirm that the convergence behavior of the general procedure is not well understood, as will be discussed further in Sect. 6.6.3.

Hung and Hou (2001) use the same basic procedure as Hung and Leachman (1996) but replace the simulation model with an analytical queueing model. The queueing model proceeds by dividing each planning period into a number of shorter subperiods and assumes steady-state behavior within the subperiods. The lead times applicable at the boundaries of the subperiods are obtained using the epoch-based lead time estimates obtained at a previous iteration. The *M/M/s* queueing model is used to predict average cycle times at individual workcenters, which are then composed into estimates of cycle times from the beginning of the process to each operation. They terminate the iterations when the percentage mean absolute deviation between the flow time estimates at successive iterations is sufficiently small. However, they find that especially at high utilization levels, the *M/M/s* queueing model predicts extremely high flow times, rendering the cycle time predictions inaccurate. They also find that the method has difficulty in converging (specifically in Fig. 7 of Hung and Hou (2001)). Hence they develop an empirical approach that uses historical data to develop a model relating expected cycle times to workload at individual workcenters, similar to the function used by Voss and Woodruff (2003). They report short computation times and good convergence for longer sub-periods, but this issue is only described briefly.

Riaño (2003) proposes a rather different iterative technique in which loading factors $w_{st}$ that describe the fraction of total releases in period $s$ that will emerge by period $t \geq s$ are estimated using a transient model of a queueing network. To present the basic idea, we shall consider its application to a single-server workcenter; the extension to multiple stages and servers is discussed in Riaño et al. (2006). A job released to the workcenter at time $s$ will see $Q(s)$ jobs ahead of it in the queue or in process. Hence the cycle time of that job will be given by

$$W(s) = \sum_{k=1}^{Q(s)} S_k + S \tag{6.21}$$

where $S_k$, $k = 2,\ldots,Q(s)$ denote the processing times of jobs ahead of this job in the queue, $S_1$ the residual (remaining) processing time of the job currently in process and $S$ the processing time for the new arrival. The distribution function of the cycle time of the job introduced into the system at time $s$ is then given by

$$G(s,t) = \sum_{n=0}^{\infty} F_1 * F^{n*}(t) P\{Q(s) = n\} \tag{6.22}$$

where $F_1$ denotes the distribution function of the residual processing time of the job currently in process, $*$ the convolution operation, and $F^{n*}$ the $n$-fold convolution of the processing time distribution $F$ at the server. $G(s,t)$ thus describes a state-dependent cycle time distribution that depends on the number of jobs $Q(s)$ in the system at the time $s$ the job was released. We seek an approximation of this function that will allow us to calculate approximate values of the loading factors $w_{st}$. To develop this approximation, the author assumes that this time-dependent delay distribution of an arriving order will have the same form as the steady-state distribution of the waiting time for an *M/G/1* queue, which is given by Shortle et al. (2018: 273), as

$$(1-\rho)\sum_{n=0}^{\infty} \rho^n F_e^{n*}(t) \tag{6.23}$$

where $F_e$ is the steady-state residual processing time distribution derived assuming that the time a new job enters the system is uniformly distributed over the duration of the current service time. This suggests an approximation of the form

$$G(s,t) = F_1 * (1-\beta(s)) \sum_{n=0}^{\infty} \beta(s)^n F_e^{n*}(t) \tag{6.24}$$

where $\beta(s)$ denotes a time-dependent traffic intensity. Noting that for a phase-type service time distribution (Neuts 1981), $G(s,t)$ will also be of phase type, the author proposes heuristic estimates of $\beta(s)$, obtaining an approximation for $G(s,t)$ that depends only on the expected WIP level at time $s$, denoted by $\phi(s)$, and its time derivative $\phi'(s)$. Hence, to obtain an approximation to $G(s,t)$, we now need a viable technique for estimating $\phi(s)$ and $\phi'(s)$. These quantities are clearly linked to the evolution of WIP over time, which, in turn, depends on the pattern of releases into the production system, suggesting a recursive technique. Given a release pattern, we can compute estimates of $\phi(t)$ for every planning period $t$ in a recursive manner, starting from period $t = 1$ and moving forward in time. If the processing time distribution at the server is phase-type, these computations can be performed efficiently. The resulting approximation to $G(s,t)$ yields approximate values of the $w_{st}$, which can be interpreted as the probability that a job released in period $s$ will complete in period $t$. The author suggests a successive approximation method to compute the $w_{st}$, where for a given release pattern estimates of the $w_{st}$ are developed after which a planning problem is solved to estimate WIP levels over time. These new WIP levels are used to estimate new loading factors until the estimates of weights converge.

The larger pattern of the iteration procedure is now clear: we begin with an initial release pattern, and calculate initial estimates of the $w_{st}$. We then calculate a new release pattern using these weights, and repeat until, hopefully, convergence is achieved. As in Zaepfel (1984), the model does not include separate capacity constraints because the load factors $w_{st}$ reflect how the input is transformed into output. “If correctly computed, they will ensure the output is actually bounded. If too much input is placed into the system the weights will reflect these longer lead times” (Riaño 2003: 72).

As with the approach of Hung and Leachman (1996), the convergence behavior of this procedure is not well understood; when it converges, it converges quite rapidly to a solution that does not depend on the initial solution used, but in other cases, it can cycle through a limited number of solutions (Riaño 2003: 83). Further experimental and theoretical work is necessary to understand this convergence issue (see Sect. 6.5.3), but the overall approach stands as a very interesting and novel approach to modeling workload-dependent lead times in production planning, with a strong theoretical underpinning. Interesting discussions in this direction are given by Hackman (2008).

The iterative mechanisms discussed so far iterate solely on the lead times or on the loading factors. Byrne and Bakir (1999) iterate between a conventional multiperiod LP production planning model that determines the optimal production levels for given capacity constraints and a simulation model that is used to update the available capacities if the production levels obtained from the initial optimization run turn out to be infeasible in the simulation. Lead times are not considered. Byrne and Hossain (2005) provide some extensions to this mechanism, again without considering lead times in the production planning model.

Kim and Kim (2001) also use loading factors to express lead times and include capacity constraints in their release model. Simulation is used to obtain estimates of the effective loading factors and resource utilization that are used to update the lead times and the capacities in the release model within an iterative mechanism. The authors do not report convergence problems in their numerical tests. Irdem et al. (2010) report good convergence of this approach under both high and low levels of resource utilization. They conclude that "the convergence behavior of the KK (Kim and Kim 2001) procedure is qualitatively different from that of the HL (Hung and Leachman 1996) procedure" (452f.). Albey and Bilge (2014) conduct extensive experiments with the KK procedure and find that the procedure converges to different solutions from different initial release plans. They also observe that when the release planning model proposes a release plan that results in low capacity utilization, agreement with the lead time estimation model is often achieved fairly quickly, which may result in the procedure converging to a suboptimal solution. Once capacity estimates have been revised downwards and passed to the release planning model, they are implemented in a hard constraint that does not permit them to be revised upwards again at a subsequent iteration. They also find that combining the values of estimates from successive iterations using a smoothing constant improves performance and that convergence in aggregate convergence criteria such as total throughput over all periods and products is much easier to obtain that agreement for each product in each period. These authors also examine the performance of the KK procedure in the presence of routing flexibility and find that increasing flexibility improves its performance.

Bang and Kim (2010) formulate an iterative procedure using an aggregate production planning model designed for semiconductor wafer fabrication that uses a separate disaggregation stage to obtain the release quantities over time. Based on an extended (compared to Hung and Leachman 1996) simulation model, not only is cycle time information updated but also product types are regrouped for the next run of the aggregate production planning model. The authors report improvements compared to Hung and Leachman (1996) and good convergence for both methods in all problem instances tested, although convergence cannot be guaranteed. Kim and Lee (2016) propose an iterative scheme where the production planning level determines production and WIP levels (or the deviations from target values, respectively). These target values are updated based on the simulated cycle times, number of setups, and available WIP. The convergence of the procedure seems to depend on the variable used to specify the convergence criterion.

### *6.6.3 Critique of Iterative Simulation-LP Algorithms*

The iterative simulation-LP approach to order release planning combines two familiar, off-the-shelf modeling techniques, linear programming, and simulation, in an iterative scheme that addresses the complex interdependency of releases and lead times. However, the simulation model requires large amounts of engineering effort and data to construct, validate, and maintain and increases run time significantly. The computational burden can be reduced by limiting the level of detail of the model to what is necessary for the specific purpose (Law and Kelton 2000: 267ff.), e.g., by focusing on highly utilized workcenters and replacing operations at low-utilization workstations with fixed time lags (Hung and Leachman 1999). The ongoing increase in computational power alleviates this problem somewhat, but does not eliminate it. The overall procedure—starting with reasonable cycle time estimates that are refined based on the simulated dynamics of the material flow—is intuitive and easy to explain. Modeling the flow time dynamics outside the optimization model allows complex system dynamics to be embedded in the simulation or queueing models used to estimate lead times, permitting realistic modeling of the system within the limits of the available computational resources.

However, the behavior of this type of order release mechanism is not well understood. There is no guarantee of optimality and hardly any insight into its deviation from the optimum. Although convergence is ergodic in some numerical experiments (Riaño 2003), there is no proof of this property. The approaches often converge within a reasonable number of iterations (five or six in Hung and Leachman (1996)), but can frequently fail to converge, in which case it does not reach a feasible solution. This is not acceptable in real-life situations and largely precludes practical application. However, Kim and Kim (2001) do not report convergence problems in their numerical tests, which might indicate that including the capacities in the iterative mechanism makes a substantial difference. Note that updating the capacities changes the right-hand side of the order release model, while updating the lead times changes the coefficient matrix. However, it is not clear how this difference is related to the mechanism that coordinates the order release and lead time estimation models.

Irdem et al. (2008, 2010) and Kacar et al. (2012) perform numerical studies that explore both the convergence of the HL (Hung and Leachman 1996) and the KK (Kim and Kim 2001) method and, in the latter paper, their performance relative to a clearing function model of the type described in the next chapter. All three papers use the same simulation testbed, a scaled-down wafer fabrication facility first studied by Kayton et al. (1997). Irdem et al. (2008) find substantial convergence problems for the HL method, especially under high bottleneck utilization, which are confirmed in Irdem et al. (2010). This behavior is qualitatively different from the KK procedure for which they report good convergence (Irdem et al. 2010; Kacar et al. 2012). Kacar et al. (2012) compare a clearing function model with two parameter settings to the KK and the HL procedure using the same testbed. They find that for the KK method convergence is achieved after four iterations in most test cases, while the HL method is "consistently outperformed by the clearing function model" (p. 116). They also conclude that the dynamic behavior of the HL method is problematic due to large swings in releases from one period to the next. The KK procedure is mostly outperformed by the clearing function model, at least for the better of the two parameter settings.

The convergence issue highlights the fact that the theory behind the iterative simulation-LP approach is largely unclear, making it difficult to explain their behavior and the nature of the solution to which they converge. Missbauer (2020) analyzes a simplified version of the HL procedure assuming a production unit with a single workcenter. He shows that in the order release model the lead times, which are time-varying parameters, act as prices for producing an item in a certain period. This is because the WIP holding costs are assigned to the production period due to the use of backward lead times and are proportional to the lead time assigned to this period. Similar insights arise in the analysis of fixed lead time models in Chap. 5, notably Equation (5.19). Hence an iterative order release procedure that iterates on the lead times behaves like a price coordination mechanism. Missbauer (2020) shows that the price coordination mechanism implied by the iterative order release mechanism does not meet the theoretical requirements for an effective price coordination mechanism, so a reasonable solution can only be expected under very specific conditions. This argument clearly does not extend, e.g., to the KK procedure that iterates on the capacities as well, suggesting different theoretical underpinning for different variants of the iterative mechanisms. These issues are largely unexplored, and more research is needed. Future research should link the design of iterative LP-simulation algorithms to the theory of mathematical decomposition and coordination that is available in the mathematical programming literature.

A comparison to the widely used techniques of simulation optimization (Fu 2002; Zapata et al. 2011) provides additional perspective on the performance of these iterative approaches. Simulation optimization is used when the objective function and constraints of the system of interest do not admit of a tractable mathematical representation but instead can be represented in a discrete-event simulation model. Thus the performance measure of interest cannot be computed directly, but must be estimated based on samples obtained from replications of the simulation. If we denote the vector of decision variables by $\theta$ and the estimate of the performance measure to be optimized obtained from the simulation replication $w$ by $L(\theta, w)$, the general statement of a simulation optimization problem is then

$$\min_{\theta \in \Theta} J\left(\theta\right) \tag{6.25}$$

where $J(\theta) = E_w[L(\theta, w]$ where $\Theta$ denotes the set of all acceptable decision variable vector $\theta$. The decision variables $\theta$ can be discrete or continuous. Fu (2002), Henderson and Nelson (2006), and Zapata et al. (2011) provide extensive reviews of this area. A wide variety of such algorithms exist, including genetic algorithms that use the simulation model to compute a fitness measure for different solutions and stochastic approximation methods for continuous state spaces. The latter methods start with an initial solution $\theta_0$ that is updated iteratively using an estimate of the gradient $\nabla J(\theta)$ of $J(\theta)$. The general form of the stochastic approximation algorithm is as follows:

*Step 1:* Choose an initial solution $\theta_0$. Set $n = 0$.

*Step 2:* Compute a new solution $\theta_{n+1} = \Pi_\Theta(\theta_n + a_n \nabla J(\theta_n))$ where $\theta_n$ is the variable set at the $n$'th iteration, $a_n$ a step size, and $\Pi_\Theta$ denotes a projection onto $\Theta$ such that if $\theta_{n+1}$ lies outside the feasible region, $\Pi_\Theta$ returns it to the feasible region; one such projection is setting $\theta_{n+1} = \theta_n$. If a specified stopping criterion is satisfied, stop and return $\theta_{n+1}$ as the estimated optimal solution. Otherwise set $n = n+1$ and return to Step 2.

The quality of the solution obtained and the speed of convergence to that solution depend on the choices of the step sizes $a_n$ and the manner in which the gradient $\nabla J(\theta_n)$ is computed. There are four general gradient estimation techniques: finite differences, likelihood ratio, perturbation analysis, and frequency domain experimentation. The finite difference technique estimates the gradient by running multiple simulations to obtain an approximation of the gradient. One version of finite differences is $\hat{\nabla}\left(J\left(\theta_n\right)\right)=\left[\hat{\nabla}_1 J\left(\theta_n\right)\cdots\hat{\nabla}_p J\left(\theta_n\right)\right]^T$ where $p$ denotes the number of decision variables and

$$\hat{\nabla}_i J\left(\theta_n\right)=\frac{\hat{J}\left(\theta_n+c_i e_i\right)-\hat{J}\left(\theta_n-c_i e_i\right)}{2c_i} \tag{6.26}$$

Convergence requires that $c_i \to 0$. Here $e_i$ denotes the $i$'th unit vector and the $c_i$ difference parameters whose values represent a trade-off between too much noise (small values) and too much bias (large values). This gradient estimation technique is broadly applicable, but requires $2p$ simulation runs at each iteration.

The direct application of simulation optimization to release planning would treat the release quantities $R_{it}$ of each product $i$ in each period $t$ as the decision variables and seek to optimize some objective function. Although simulation optimization is generally employed in the presence of random variables such as processing times, machine failures and yields, the basic approach can be implemented in completely deterministic simulations. Although models based on this approach have been developed and shown to yield good solutions (Liu et al. 2011; Kacar and Uzsoy 2015), their computational requirements are usually very high due to the time required to run multiple independent replications of a large simulation model. Some recent work attempts to reduce the computational burden of these procedures by replacing the simulation model with a metamodel based on extensive offline simulation experiments, with promising results (Li et al. 2016).

The iterative multi-model approaches can be viewed as simulations of a particular decision process: initial estimates of planning parameters such as lead times and resource utilizations are obtained, the release planning model is run, and the resulting release pattern is simulated. This perspective provides some insight into their performance. First, most multi-model iterative approaches do not consider the objective function value in their convergence criteria; instead they focus on achieving consistency in the flow time estimates obtained from successive iterations. Hence there is no *a priori* evidence that these procedures will converge to even a locally optimal solution with respect to the objective function of concern, as implemented in the release planning model; the best that can be hoped for is a feasible solution. Although the primary concern is the reduction of the differences in lead time estimates obtained at successive iterations, this is never explicitly formulated as an objective function to be reduced from one iteration to the next, nor is any information on the gradient of this quantity used. Simulation optimization methods that explicitly consider the gradient of the objective function generally yield good solutions, although their computational burden is very high.

Viewing these techniques as applications of fixed point iteration also raises concerns. The basic fixed point iteration procedure, common in numerical analysis, generates a sequence of solutions $x_{n+1} = f(x_n)$, $n = 0, 1, \ldots$. In the context of the iterative multi-model methods, the solution $x_n$ at iteration $n$ represents a vector of lead time estimates, while the function $f(x_n)$ represents the simulation of the decision process by which a release schedule is obtained by the LP model from the previous iteration's lead time estimates $x_{n-1}$. This release schedule is then simulated to obtain revised lead time estimates. Per the Banach Fixed Point Theorem (O'Regan et al. 2001), the existence of a fixed point in general requires the existence of a contraction mapping such that for any two points $x_i$ and $x_j$ there exists a constant $0 \leq q < 1$ such that $||f(x_i) - f(x_j)|| \leq q||x_i - x_j||$. In the current iterative methods, no conditions of this type are considered, let alone satisfied.

Our discussion of simulation optimization and fixed point iteration in relation to the iterative multi-model procedures is clearly heuristic in nature and provides no mathematically rigorous evidence. However, these considerations do suggest that most existing iterative methods are, in mathematical terms, ill-posed and require the imposition of additional conditions to ensure reliable performance in terms of solution quality and convergence.

## 6.7 Iterative Methods for Production Planning and Scheduling

The iterative methods described in Sect. 6.5.2 represent a small and rather specialized research direction in order release planning. However, a closer look at the literature reveals that this is a special case of a more general problem: Order release planning—as a subproblem of production planning—requires information on lead times and maximum possible production which, in turn, depend on the detailed schedule within the production unit. While it is true that, as stated in the optimized production technology (OPT) approach, “lead times are the result of a schedule and can’t be predetermined” (Vollmann et al. 1997: 797), the monolithic approach to production planning and control, at least for the bottleneck workcenters, that results from this view is not always applicable, motivating the hierarchical approaches described in Chap. 1. Planned lead times allow decomposition of the complex planning problem into planning and scheduling levels (Graves 2011: 93) and thus are necessary within this planning concept, but both lead times and capacities should anticipate the outcomes of the scheduling level reliably (Kanet and Sridharan 1998).

It is thus not surprising that iteration between the planning and scheduling levels has also been proposed for other production planning tasks. Integrating the planning and scheduling levels is particularly important in lot sizing. This can be achieved, e.g., by anticipating the queuing effects of lot sizes using stochastic models and determining lot sizes accordingly, as discussed in Chap. 9, or by lot streaming, that is, splitting up production lots into smaller transfer batches whose processing on different workcenters can be overlapped in time (Cheng et al. 2013). Dauzere-Peres and Lasserre (2002) present an integrated model for lot sizing and scheduling and an algorithm that iterates between a lot-sizing module that assumes a fixed production sequence and a scheduling module that sequences the given lots. In this approach the lead time acts as a capacity constraint (p. 789). Negenman (2000) presents an algorithm that iterates between an LP model that calculates the production plan for a production network and a flexible flow shop scheduling model that is solved by a heuristic. The feedback information provided by the scheduling level is the completion times of the orders. If the planned lead times are exceeded, the planning level reduces the available capacities of the workcenters in the next iteration. A detailed analysis of the convergence behavior is not provided. Albey and Bilge (2011) present a hierarchical production planning and control system framework for a Flexible Manufacturing System that consists of three levels: aggregate planning, loading, and detailed planning. The behavior of the shop floor for a given production plan is anticipated using simulation. The simulation result is used to update capacity coefficients in the upper-level modules. Again, convergence is not analyzed in detail, but the authors indicate that capacity updating is complex due to the special problem and decision structure.

Planned lead times can also be required for dispatching decisions when the dispatching rule compares a job's current slack time to its remaining lead time. In this case the lead time estimate must be consistent with the schedule that is based on this estimate (Vepsalainen and Morton 1988). The lead time iteration method (Vepsalainen and Morton 1988; Morton and Pentico 1993) updates initial lead time estimates used for scheduling using the actual flow times obtained from the scheduling algorithm using exponential smoothing (Morton and Pentico 1993: 218f). Convergence is not guaranteed, and "it is then an empirical question whether such a procedure obtains good results or not" (Morton and Pentico 1993: 219). Lu et al. (1994) provide an interesting illustration of the lead time iteration procedure in a semiconductor wafer fabrication facility. Note that the role of planned lead times in scheduling algorithms is different from that in order release models, and thus the relationship of these results to the convergence issue of the LP-simulation approaches discussed in the previous section is not straightforward. A unifying view of algorithms that iterate between a production planning model, independent of its formulation, and a scheduling model, independent of the scheduling algorithm, is a challenge for future research.

## 6.8 Conclusions

The various models discussed in this chapter highlight the difficulty of the central problem addressed in this volume: how to anticipate the behavior of the scheduling level in planning models in a manner that is both sufficiently accurate and computationally tractable. The linear programming models presented in Chap. 5 can be extended easily to handle time-varying exogenous lead times, but this begs the question of how to obtain such estimates since lead times are determined by utilization and utilization by the release decisions the model seeks to address. Work in traffic modeling suggests that optimization models with lead times as an endogenous decision variable are often non-convex and hence hard to solve. Attempts to preserve computational tractability have led to the use of multi-model approaches that separate the problems of release planning and lead time estimation, but the convergence behavior of these is not well understood, and the use of a simulation model to construct the planning solution (as opposed to estimating its parameters offline, outside the planning run) result in high computational burden for large production systems. What is needed is a way of representing the behavior of the scheduling level within the release planning model that is consistent with the queueing view of production resources in Chap. 2, but which yields tractable optimization models. The clearing functions discussed in the next two chapters seek to provide such a model.

## References

Albey E, Bilge Ü (2011) A hierarchical approach to FMS planning and control with simulation-based capacity anticipation. Int J Prod Res 49(11):3319–3342

Albey E, Bilge U (2014) An improved iterative linear programming-simulation approach for production planning. Department of Industrial Engineering, Ozyegin University, Istanbul

Armbruster D, Uzsoy R (2012) Continuous dynamic models, clearing functions, and discrete-event simulation in aggregate production planning. INFORMS Tutorials in Operations Research

Bang JY, Kim YD (2010) Hierarchical production planning for semiconductor wafer fabrication based on linear programming and discrete-event simulation. IEEE Trans Autom Sci Eng 7(2):326–336

Bazaraa MS, Sherali HD, Jarvis J (1979) Nonlinear programming: theory and algorithms. Wiley, New York

Ben-Daya M, Raouf A (1994) Inventory models involving lead time as a decision variable. J Oper Res Soc 45(5):579–582

Bertsimas D, Mourtzinou G (1997) Transient laws of non-stationary queueing systems and their applications. Queue Syst 25:115–155

Byrne MD, Bakir MA (1999) Production planning using a hybrid simulation-analytical approach. Int J Prod Econ 59:305–311

Byrne MD, Hossain MM (2005) Production planning: an improved hybrid approach. Int J Prod Econ 93-94:225–229

Carey M (1992) Nonconvexity of the dynamic traffic assignment problem. Transport Res B 26B(2):127–133

Carey M, Subrahmanian E (2000) An approach to modelling time-varying flows on congested networks. Transport Res B 34:157–183

Cheng M, Mukherjee NJ, Sarin SC (2013) A review of lot streaming. Int J Prod Res 51(23/24):7023–7046

Dauzere-Peres S, Lasserre JB (2002) On the importance of sequencing decisions in production planning and scheduling. Int Trans Oper Res 9:779–793

Ehteshami B, Petrakian R, Shabe P (1992) Trade-offs in cycle time management: hot lots. IEEE Trans Semicond Manuf 5(2):101–106

Figueira G, Almada-Lobo B (2014) Hybrid simulation–optimization methods: a taxonomy and discussion. Simul Model Pract Theor 46:118–134

Fu MC (2002) Optimization for simulation: theory vs practice. INFORMS J Comput 14(3):192–215

Gong L, de Kok T, Ding J (1994) Optimal leadtimes planning in a serial production system. Manag Sci 40(5):629–632

Graves SC (2011) In: Kempf KG, Keskinocak P, Uzsoy R (eds) Uncertainty and Production Planning. Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook, Volume 1, International Series in Operations Research and Management Science, vol 151. Springer, New York and Heidelberg, pp 83–101

Hackman S (2008) Production economics: integrating the microeconomic and engineering perspectives. Springer, Berlin

Hackman ST, Leachman RC (1989) A general framework for modeling production. Manag Sci 35(4):478–495

Henderson SG, Nelson BL eds (2006) Simulation. In: Handbooks in operations research and management science. North-Holland, Amsterdam

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Hopp WJ, Sturgis MLR (2000) Quoting manufacturing due dates subject to a service level constraint. IIE Trans 32(9):771–784

Hung YF, Hou MC (2001) A production planning approach based on iterations of linear programming optimization and flow time prediction. J Chin Inst Indus Eng 18(3):55–67

Hung YF, Leachman RC (1996) A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. IEEE Trans Semicond Manuf 9(2):257–269

Hung YF, Leachman RC (1999) Reduced simulation models of wafer fabrication facilities. Int J Prod Res 37(12):2685–2701

Ioannou G, Dimitriou S (2012) Lead time estimation in MRP/ERP for make-to-order manufacturing systems. Int J Prod Econ 139(2):551–563

Irdem DF, Kacar NB, Uzsoy R (2008) An experimental study of an iterative simulation-optimization algorithm for production planning. In: Mason SJ, Hill R, Moench L, Rose O (eds) 2008 Winter Simulation Conference, Miami, FL

Irdem DF, Kacar NB, Uzsoy R (2010) An exploratory analysis of two iterative linear programming-simulation approaches for production planning. IEEE Trans Semicond Manuf 23:442–455

Jonsson P, Matsson SA (2006) A longitudinal study of material planning applications in manufacturing companies. Int J Oper Prod Manag 26(9):971–995

Kacar NB, Uzsoy R (2015) Estimating clearing functions for production resources using simulation optimization. IEEE Trans Autom Sci Eng 12(2):539–552

Kacar NB, Irdem DF, Uzsoy R (2012) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. IEEE Trans Semicond Manuf 25(1):104–117

Kanet JJ, Sridharan V (1998) The value of using scheduling information in planning material requirements. Decis Sci 29(2):479–496

Kayton D, Teyner T, Schwartz C, Uzsoy R (1997) Focusing maintenance improvement efforts in a wafer fabrication facility operating under theory of constraints. Prod Invent Manag 38(Fourth Quarter):51–57

Keskinocak P, Tayur S (2004) Due-date management policies. In: Simchi-Levi D, Wu SD, Shen ZM (eds) Supply chain analysis in the e-business era: handbook of quantitative supply chain analysis. Kluwer Academic, Dordrecht

Kim B, Kim S (2001) Extended model for a hybrid production planning approach. Int J Prod Econ 73:165–173

Kim SH, Lee YH (2016) Synchronized production planning and scheduling in semiconductor fabrication. Comput Indus Eng 96:72–85

Lautenschläger M (1999) Mittelfristige Produktionsprogrammplanung mit auslastungsabhängigen Vorlaufzeiten. Peter Lang, Frankfurt am Main

Law AM, Kelton WD (2000) Simulation modeling and analysis, 3rd edn. McGraw Hill, New York

Law AM, Kelton WD (2004) Simulation modeling and analysis. McGraw-Hill, New York

Leachman RC, Carmon TF (1992) On capacity modeling for production planning with alternative machine types. IIE Trans 24(4):62–72

Li M, Yang F, Uzsoy R, Xu J (2016) A metamodel-based monte carlo simulation approach for responsive production planning of manufacturing systems. J Manuf Syst 38:114–133

Liu J, Li C, Yang F, Wan H, Uzsoy R (2011) Production planning for semiconductor manufacturing via simulation optimization. In: Jain S, Creasey RR, Himmelspach J, White KP, Fu R (eds) Winter simulation conferemce. IEEE, Piscataway, NJ

Lu S, Ramaswamy D, Kumar PR (1994) Efficient scheduling policies to reduce mean and variance of cycle time in semiconductor plants. IEEE Trans Semicond Manuf 7:374–388

Milne RJ, Mahapatra S, Wang C-T (2015) Optimizing planned lead times for enhancing performance of MRP systems. Int J Prod Econ 167:220–231

Missbauer H (2020) Order release planning by iterative simulation and linear programming: theoretical foundation and analysis of its shortcomings. Eur J Oper Res 280:495–507

Morton TE, Pentico D (1993) Heuristic scheduling systems: with applications to production systems and project management. Wiley, New York

Narahari Y, Khan LM (1997) Modeling the effect of hot lots in semiconductor manufacturing systems. IEEE Trans Semicond Manuf 10(1):185–188

Negenman EG (2000) Material coordination under capacity constraints. Industrial engineering. Eindhoven University of Technology, Eindhoven

Neuts MF (1981) Matrix-geometric solutions in stochastic models. Johns Hopkins University Press, Baltimore, MD

O'Regan D, Meehan M, Agarwal RP (2001) Contractions. In: Fixed point theory and applications. Cambridge University Press, Cambridge, pp 1–11

Orcun S, Uzsoy R (2011) The effects of production planning on the dynamic behavior of a simple supply chain: an experimental study. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning in the extended enterprise: a state of the art handbook. Springer, Berlin, pp 43–80

Ozturk A, Kayaligil S, Ozdemirel NE (2006) Manufacturing lead time estimation using data mining. Eur J Oper Res 173:683–700

Peeta S, Ziliaskopoulos AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. Netw Spat Econ 1(3-4):233–265

Riaño G (2003) Transient behavior of stochastic networks: application to production planning with load-dependent lead times. School of Industrial and Systems Engineering. Georgia Institute of Technology, Atlanta, GA

Riaño G, Hackman S, Serfozo R (2006) Transient behavior of queueing networks. School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA

Shanthikumar JG, Sargent RG (1983) A unifying view of hybrid simulation/analytic models and modeling. Oper Res 31(6):1030–1052

Shortle JF, Thompson JM, Gross D, Harris CM (2018) Fundamentals of queueing theory. Wiley, Hoboken, NJ

Vepsalainen AP, Morton TE (1988) Improving local priority rules with global lead-time estimates: a simulation study. J Manuf Oper Manag 1:102–118

Vollmann T, Berry W, Whybark D (1997) Manufacturing planning and control systems. Irwin, Boston

Voss S, Woodruff DL (2003) Introduction to computational optimization models for production planning in a supply chain. Springer, New York

Zaepfel G (1984) Systemanalytische Konzeption der Produktionsplanung und –steuerung für Betriebe der Fertigungsindustrie. In: Zink C (ed) Sozio-Technologische Systemgestaltung als Zukunftsaufgabe, (in German). Carl Hanser Verlag, Munich

Zapata JC, Pekny J, Reklaitis GV (2011) Simulation-optimization in support of tactical and strategic enterprise decisions. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extemnded enterprise: a state of the art handbook, vol 1. Springer, New York, pp 593–628

# Chapter 7
# Univariate Clearing Functions

In this chapter, we introduce the concept of the clearing function (CF), a metamodel of a production resource that relates the expected output of a resource to some measure of the work available to it in the planning period. We focus on clearing functions with a single state variable and examine a variety of functional forms that have been proposed in the production and traffic literature. We then formulate release planning models using these functions and show that while single-product models yield tractable convex optimization problems, the presence of multiple products competing for capacity at a shared resource creates significant difficulties. The allocated clearing function formulation is presented to address these issues and shown to yield more informative dual prices for resource capacity than conventional LP models.

## 7.1 Preliminaries

The models in the previous two chapters anticipate the performance of the production units using exogenous, workload-independent lead times that are assumed to remain valid as long as a maximum capacity loading is not exceeded. These lead time estimates may take different forms based on how capacity is consumed during the lead time, as discussed in Chap. 5, and can be specific to individual planning periods as discussed in Chap. 6.

The combination of fixed, exogenous planned lead times with a maximum capacity limit as an anticipation function yields computationally tractable linear programming (LP) models as long as lot sizing is not a consideration; the presence of lot sizing requires the introduction of integer variables and yields considerably more challenging models (Pochet and Wolsey 2006). As long as the production unit operates at approximately constant utilization over time, historical data can be used to estimate planned lead times that are consistent with observed cycle times, for example by setting the planned lead times to a specified fractile of the observed cycle

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_7

time distribution. However, if the resource utilization level, the product mix, or both vary over time, the distribution of the cycle time will also change over time. This, in turn, may cause the cycle times observed on the shop floor to deviate significantly from the lead times used in the planning models, adversely affecting the performance of the production units trying to execute these plans.

In contrast to these LP models where the output of the system is determined by the combination of planned lead times and a maximum capacity loading, the models in this chapter express the expected output of the production unit in a planning period as a function of the workload available to the resource for processing in that period. Models of this type have arisen in the context of queueing systems, in the management of traffic networks and as representations of particular production control policies. We shall refer to models of this type as *clearing functions*, following the terminology of Karmarkar (1989).

We define a clearing function as a functional relationship that specifies the expected output $X_t$ of a production resource in a planning period $t$ of duration $\Delta$ as

$$X_t = f\left(\Delta, \Omega_t\right) \tag{7.1}$$

where $\Omega_t$ denotes a set of state variables that collectively describe the amount of work available to the resource in period $t$. The specific set of state variables to include in the set $\Omega_t$ is not immediately obvious. From a queueing perspective, the state of the resource at time $t$ potentially depends on the entire past history of the relevant stochastic processes (interarrival times, service times, machine failures, setups, number of available machines, etc.) up to that instant in time. It is also apparent that the clearing function must depend on the length $\Delta$ of the planning period for which it is being constructed. Finally, the amount of work available to the resource and the distribution of its arrival over time depend on the model used by the planning level to determine releases over time. In queueing terms, the release decisions made by the planning level affect both the mean interarrival time of orders to the resource and its variance.

The purpose of the clearing function is to represent the behavior of the resource to an acceptable degree of accuracy while still yielding tractable optimization models for the planning problem. The extremely high dimensionality and complex functional forms required by general methods, such as queueing approaches considering the entire history of the process or a large portion of it, make it very difficult to obtain clearing functions leading to tractable optimization models. Even simple functional forms for clearing functions can yield non-convex optimization models. Hence most clearing functions proposed to date have used a single state variable; we shall see that even in this case formulations involving multiple products can become challenging. In this chapter, we discuss various single-variable clearing functions, the difficulties that arise when multiple products compete for capacity at a resource, and solutions to these difficulties. We also show that planning models using clearing functions can produce meaningful dual prices for resources at any level of utilization, which is not the case for the models discussed in Chaps. 5 and 6.

## 7.2 Single-Variable Clearing Functions

### 7.2.1 Average WIP-Based Clearing Functions

This family of clearing functions, the motivation for which was sketched in Sect. 2.2, uses the set of state variables $\Omega_t = \left\{\bar{W}_t\right\}$, where $\bar{W}_t$ denotes the time-average WIP level, measured in number of units or lots, at the production resource over the planning period *t*. Specifically, if planning period *t* spans the time interval *(t−Δ, t]* and *W(t)* denotes the amount of WIP at the resource at time *t*, we have

$$\bar{W}_t = \frac{1}{\Delta}\int_{t-\Delta}^{t} W(\tau)\,d\tau \tag{7.2}$$

The advantage of $\bar{W}_t$ as a workload metric is its straightforward relation to the well-known steady-state analyses of queues such as the *M/G/1* and *G/G/1* (Buzacott and Shanthikumar 1993; Curry and Feldman 2000), from which exact or approximate expressions relating the expected WIP, expected cycle time and utilization can be derived. As discussed in Sect. 2.2, the expected WIP level of the *G/G/1* queue in steady state is given by

$$\bar{W} = \frac{T}{t_{\mathrm{a}}} = \left(\frac{c_{\mathrm{a}}^2 + c_{\mathrm{e}}^2}{2}\right)\left(\frac{u}{1-u}\right)\frac{t_{\mathrm{e}}}{t_{\mathrm{a}}} + \frac{t_{\mathrm{e}}}{t_{\mathrm{a}}} = \left(\frac{c_{\mathrm{a}}^2 + c_{\mathrm{e}}^2}{2}\right)\left(\frac{u^2}{1-u}\right) + u \tag{7.3}$$

dropping the time subscript since this is a steady-state relation. Solving for *u* in terms of $\bar{W}$ yields a quadratic equation in $\bar{W}$ whose nonnegative solution is

$$u = \frac{-\left(\bar{W}+1\right) + \sqrt{\left(\bar{W}+1\right)^2 + 4\left(\psi-1\right)\bar{W}}}{2\left(\psi-1\right)} \tag{7.4}$$

where $\psi = \left(c_{\mathrm{a}}^2 + c_{\mathrm{e}}^2\right)/2$; recall from Sect. 2.2 that $\psi = 1$ represents the special case of the *M/M/1* queue. Intuitively, the higher the average WIP level $\bar{W}$ at the resource, the lower the probability (1 − *u*) that the resource will be idle due to lack of work; hence maintaining a planned average throughput rate of *X* in a planning period requires maintaining a certain average WIP level at the resource. The average utilization can be interpreted as the fraction of the planning period during which the resource will be producing usable output. Thus the expected number of units produced over the planning period is given by:

$$X_t = \frac{u\Delta}{t_{\mathrm{e}}} = \frac{\Delta}{t_{\mathrm{e}}}\left[\frac{-\left(\bar{W}+1\right) + \sqrt{\left(\bar{W}+1\right)^2 + 4\left(\psi-1\right)\bar{W}}}{2\left(\psi-1\right)}\right] \tag{7.5}$$

Incorporating this state variable into optimization models requires some additional considerations. At the point in time the optimization model is solved to determine releases for the next $T$ periods, the average WIP values $\bar{W}_t$ are not known with certainty; they are in fact random variables whose distribution is determined by the release decisions made by the planning model. Hence the state variables $W_t$ representing WIP in the optimization models actually represent the planned state of the resource at the end of period $t$ and do not capture the evolution of the WIP level throughout the planning period. However, many different WIP trajectories $W(t)$ may give the same beginning and ending WIP levels $W_{t-1}$ and $W_t$. Optimization models using this type of clearing function must estimate the planned value of $\bar{W}_t$ for a given planning period $t$ using the planned values of $W_{t-1}$ and $W_t$. The most obvious approach is to use the arithmetic average to obtain

$$\bar{W}_t = \frac{\left(W_{t-1} + W_t\right)}{2} \tag{7.6}$$

However, this has implications for the behavior of the resulting optimization models. Note that if $W_{t-1}$ is increased by a certain amount in (7.6) and $W_t$ reduced by the same amount, $\bar{W}_t$ remains unchanged. Depending on the structure of the optimization model, this can lead to oscillating WIP levels at the period boundaries due to the presence of alternative optimal solutions, which is undesirable (Missbauer (1998): 413 ff.).

Given their origin in steady-state queueing analysis, clearing functions of this type are more appropriate for longer planning periods, where the transient behavior of the resource at the start of the period due to changes in releases can safely be neglected. Note that it is possible to have $X_t \geq \bar{W}_t$ using a clearing function of this type; at low utilization levels, the average queue length will be very small, while the total output will be approximately equal to the number of arrivals during the period.

Although it is not explicitly stated as such, the practical worst case model of production lines given in Chap. 7 of Hopp and Spearman (2008) also represents an average WIP-based clearing function. This model considers a balanced serial production line operating under the CONWIP policy discussed in Chap. 4. They define the system state as a vector whose components represent the number of jobs in front of each machine in the line. Assuming all such states to be equally likely, they note that for a total WIP level of $w$ jobs in the line, a new job entering the system will see on average

$$W_i = \frac{\left(w-1\right)}{N} \tag{7.7}$$

jobs ahead of it at each of the $N$ machines in the system, implying an average cycle time of

$$T = T_0 + \frac{w-1}{r_{\mathrm{b}}} \tag{7.8}$$

where $T_0$ denotes the raw processing time of the line, the average time in system a job will encounter if it enters an empty line, and $r_b$ the processing rate of the bottleneck machine. Substituting (7.8) into Little's Law yields an average throughput rate of

$$X = \frac{wr_b}{r_b T_0 + w - 1} \tag{7.9}$$

Since in a CONWIP system the average WIP level will be equal to the total WIP level $w$ permitted in the system, this represents an average WIP-based clearing function that can be shown to be concave and monotonically non-decreasing in the average WIP level $w$. The assumption of equally likely system states is exact only for a balanced line with a single exponential server at each stage, but provides a WIP level that is unlikely to be exceeded in systems with more general structures.

### 7.2.2 Initial WIP-Based Clearing Functions

This family of clearing functions assumes that the expected output of the resource in a planning period is determined solely by the amount of work available to it at the start of the planning period; work arriving during the period will have no effect on expected output. Hence the set of state variables considered in each period $t$ is $\Omega_t = \{W_{t-1}\}$. Under this model either the probability of new work arriving during the planning period is negligible, the scheduling policy only allows work to be released at the start of a period (which coincides with the end of the previous one), or the planning interval is sufficiently short that work available at the start of the period will fully occupy the resource until the next period.

Clearing functions of this type have been discussed extensively in the context of traffic assignment problems (Dafermos and Sparrow 1969; Carey 1987; Peeta and Ziliaskopoulos 2001) where they are used to model the behavior of a section of highway in a given time period. In these networks, which bear considerable similarity to those studied in this volume, a traffic system is modeled as a network with node set $N$ and directed arc set $A$. The arcs $(i,j) \in A$ correspond to specific segments of roadway whose starting and ending points are represented by nodes $i, j \in N$, respectively. The amount of traffic $X_{ij}(t)$ that can exit the arc $(i,j)$ over a planning period $t$ is expressed as a concave, non-decreasing function $g_{ij}(W_{ij}(t-1))$ of the amount of traffic $W_{ij}(t-1)$ present on the arc at the start of the period. These exit functions are used in discrete-time optimization models very similar to those developed later in this chapter.

The exit functions used in the dynamic traffic assignment work are derived from flow-density functions, which are discussed in detail in Carey and Bowers (2012). The basic resource considered in these models, analogous to the machine or work-center in production units, is a segment of road whose characteristics such as width, surface quality, visibility, and signage are assumed to be known. For ease of exposition we shall assume the road segment to be of unit length, and will drop the time subscript to discuss a generic time period, as in the discussion of steady-state clearing functions in Sect. 7.2.1. The progress of individual vehicles along the road segment is represented as a continuous flow, in much the same manner as the LP models of Chap. 5 treat the processing of discrete orders at the production resources. The traffic density $k$ represents the number of vehicles occupying the road segment of unit length being considered. This quantity is analogous to the average WIP $\bar{W}_t$ or workload in production contexts. The flow rate $q$, the number of vehicles passing a particular point on the road per unit time, is analogous to the throughput rate $X$ of a workcenter or production resource. Hence the exit function captures the rate at which vehicles pass the end point of the road segment, either entering another segment or exiting the system. The space mean speed $v$ of the traffic along the unit road segment is given by the length of the road segment divided by the average time to traverse it. The relation between flow rate $q$, speed $v$, and traffic density $k$ is thus

$$q = kv \tag{7.10}$$

Noting that $v = 1/T$, where $T$ denotes the average time to traverse the road segment, we obtain

$$q = \frac{k}{T} \tag{7.11}$$

which can be rewritten as

$$k = qT \tag{7.12}$$

Replacing each term with its counterpart in the production context ($k$ with $\bar{W}$ and $q$ with $X$) and noting that the interpretation of $T$ as the average time to traverse the system under consideration is the same in both traffic and production contexts, we recover Little's Law (Hopp and Spearman 2008):

$$W = XT \tag{7.13}$$

Flow-density functions $q = f(k)$ are intended to be empirical relations whose parameters are estimated from appropriately collected data. However, most flow-density functions $f(.)$ used in traffic research have been derived using a limited set of parameters:

- The free-flow velocity $V_0$ of the road segment, representing the flow of traffic at very low density, analogous to the raw process time $T_0$ discussed in the previous section. Since by (7.10) the average velocity $v = q/k = f(k)/k$, we have

$$V_0 = \lim_{k \to 0^+} \frac{f(k)}{k} = \left.\frac{\partial f}{\partial k}\right|_{k=0} \tag{7.14}$$

- The jam density $k_j$, the density at which $v = q = 0$, i.e., traffic comes to a stop.

- The wave speed at jam density $c_\mathrm{j}$, the rate at which flow decreases as density increases to the jam density $k_\mathrm{j}$, given by

$$c_\mathrm{j} = \lim_{k \to k_\mathrm{j}} \frac{df\left(k\right)}{dk} \tag{7.15}$$

- The maximum flow rate $q_\mathrm{c}$. The density at which the maximum flow rate occurs is referred to as the critical density $k_\mathrm{c}$, analogous to the critical WIP concept of Hopp and Spearman (2008).

Carey and Bowers (2012) propose several desirable properties for a flow-density function. These include unimodality, appropriate finite values of the free-flow speed $V_0$, jam density $k_\mathrm{j}$, and the ratio $k_\mathrm{j}/k_\mathrm{c}$, as well as an appropriate negative value of $c_\mathrm{j}$ and the possibility of convexity as $k \to k_\mathrm{j}$. A generic flow-density function *f(k)* satisfying these conditions would appear as shown in Fig. 7.1.

Production systems research has generally assumed an infinite jam density $k_\mathrm{j} = \infty$, under the assumption that as the work available to a queueing system in a planning period increases its output rate $X$ will eventually level off at $1/t_\mathrm{e}$, but will never decrease. In environments where jobs do not interfere with each other through sequence-dependent setup times or scheduling policies, this assumption appears reasonable. Hence most clearing functions proposed by production system researchers have taken the form of monotonically non-decreasing concave functions that asymptotically approach the maximum production rate as workload or WIP approach infinity. Clearing functions for environments where this assumption is not valid, such as those with significant sequence-dependent setup times, are discussed in the next two chapters. Clearing functions that decrease beyond a certain WIP level like the flow-density function in Fig. 7.1, due to e.g., reduced worker efficiency when workload is too high or by excessive material shuffling which reduces capacity, are rare in the literature (Van Ooijen and Bertrand 2003).

While a wide range of flow-density functions have been discussed in the traffic research community, we will use two examples to illustrate the types of models

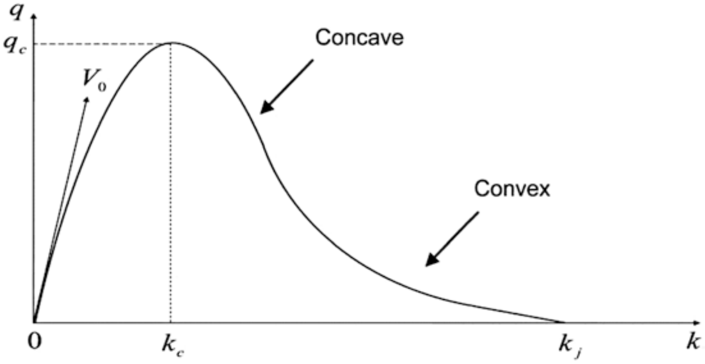


**Fig. 7.1** A generic flow-density function (Carey and Bowers 2012)

considered. The output function proposed by Newell (1961) and Franklin (1961) takes the form

$$f(k) = kV_0\left(1 - \exp\left[\frac{|c_{\mathrm{j}}|}{V_0}\left(1 - \frac{k_{\mathrm{j}}}{k}\right)\right]\right) \tag{7.16}$$

Carey and Bowers (2012) note that this flow-density function satisfies more of the desirable properties they propose than any other function; however, it is concave everywhere, not admitting convexity as the jam density is approached. They also point out that the function is defined by three parameters ($k_{\mathrm{j}}$, $V_0$, and $c_{\mathrm{j}}$) that give the behavior of the function at the origin and at jam density unduly high influence on its overall shape. This function appears to have motivated the clearing function of Srinivasan et al. (1988) discussed below. Another class of flow-density functions proposed by Van Aerde and Rakha (1995) takes the form

$$f(k) = \frac{1 - (c_1 - c_3V_0)k - \sqrt{\left[(c_1 + c_3V_0) - 1\right]^2 + 4c_3c_2k^2}}{2c_3k} \tag{7.17}$$

where $c_1$, $c_2$, and $c_3$ are constants computed from $V_0$, $k_{\mathrm{j}}$, $q_{\mathrm{c}}$, and $v_{\mathrm{c}}$, where the latter denotes the average speed at critical density $q_{\mathrm{c}}$. The resemblance to (7.4) is striking.

### 7.2.3 Workload-Based Clearing Functions

The discrete-time nature of production planning models creates difficulties for average WIP-based clearing functions due to the fact that multiple combinations of values for $W_t$ and $W_{t-1}$ can yield the same $\overline{W}_t$ value for any period $t$. Initial WIP-based clearing functions assume that the expected output $X_t$ of the resource in period $t$ cannot exceed the initial WIP $W_{t-1}$ available at the start of the period, ignoring the possibility that work released during the period might be completed during the period. Workload-based clearing functions address this issue by using a state variable $\Lambda_t$ that represents the total amount of work made available to the resource during period $t$, given by

$$\Lambda_t = W_{t-1} + R_t \tag{7.18}$$

where $W_{t-1}$ denotes the amount of WIP carried over from the previous period $t - 1$ and $R_t$ the amount of work released to the resource during period $t$. Clearing functions of this form must have

$$\frac{\partial f}{\partial \Lambda_t} \leq 1 \tag{7.19}$$

for all $\Lambda_t \geq 0$, implying that the resource can never convert more material into output in a period than becomes available to it over the period.

Missbauer (2002) proposes a clearing function of this form for a resource that can be represented as an *M/G/1* queue in steady state. We present here the same development for a *G/G/1* queue. Recall from (7.5) that the expected throughput of a *G/G/1* queue in steady state can be approximated as

$$X_t = \frac{u\Delta}{t_\mathrm{e}} = \frac{\Delta}{t_\mathrm{e}}\left[\frac{-\left(\bar{W}+1\right)+\sqrt{\left(\bar{W}+1\right)^2+4\left(\psi-1\right)\bar{W}}}{2\left(\psi-1\right)}\right] \tag{7.20}$$

Now consider a *G/G/1* queue in steady state where at the start of some planning period $t$ there are $W_{t-1}$ units of work remaining on hand from the previous period and $R_t$ units are released into the production unit. Recalling that the workload $\Lambda_t = W_{t-1} + R_t$, we have $W_{t-1} = \Lambda_t - R_t$. Since the queue is assumed to be in steady state, we must have $X_t = R_t$ and $W_{t-1} = \bar{W}_t$. Substituting $\bar{W}_t = \Lambda_t - X_t$ into (7.20) and solving for $X_t$, we obtain

$$X_t = \frac{\Delta}{t_\mathrm{e}}\left[\frac{\left(\Delta+t_\mathrm{e}\left(1+\Lambda_t\right)\right)-\sqrt{\left[\Delta+\left(1+\Lambda_t\right)\right]^2-4\Lambda_t t_\mathrm{e}\left[\Delta+t_\mathrm{e}\left(\psi-1\right)\right]}}{2\left[\Delta-t_\mathrm{e}\left(\psi-1\right)\right]}\right] \tag{7.21}$$

The basic form of this expression is quite similar to that derived for the average WIP case in (7.20); most notably, it retains the concave saturating form and guarantees that $X_t \leq \Lambda_t$. Its drawback is the assumption of steady state, which is not generally valid under the conditions of time-varying demand and finite period length under which we wish to use the release planning models we study. Again, we note in passing the similarity to (7.17).

### *7.2.4 The Constant Cycle Time Clearing Function*

Graves (1986) proposes a discrete-time model of a production resource whose expected output $X_t$ in period $t$ is given by the clearing function

$$X_t = \alpha W_{t-1} \tag{7.22}$$

where $W_{t-1}$ denotes the amount of WIP available to the resource at the start of period $t$, i.e., the end of period $t-1$. Since Graves assumes that work can only arrive at or depart from the resource at the start of a planning period, this can also be viewed as a workload-based clearing function in our terminology. The resource will always process a fraction $\alpha$ of the WIP $W_{t-1}$ available to it at the start of the period, no matter how large $W_{t-1}$ may be. Equivalently, the model assumes that the resource is managed to maintain an average cycle time of $1/\alpha$ periods; as the amount of available work $W_{t-1}$ increases, the resource can work faster. Hence this linear clearing function is best viewed as describing the behavior of the production resource under a specified production control policy, where the processing rate can be varied to maintain the planned lead time of $1/\alpha$ periods. The clearing function will, naturally, only be valid over the range of operating conditions that satisfy this condition.

The author uses clearing functions of this type to analyze the performance of a job shop by computing the mean and variance of performance measures such as throughput, queue length, and backlog. In particular, he examines the tradeoff between production smoothing (which requires long planned lead times and hence low values of $\alpha$) and reducing cycle times and WIP levels (which requires high values of $\alpha$) by simulating a job shop environment. The author uses this model in several subsequent papers to examine the issue of setting safety stocks in such systems (Graves 1988), planning in multistage production-inventory systems (Graves et al. 1998), and setting planned lead times in make-to-order systems (Teo et al. 2011; Teo et al. 2012). Parrish (1987) extends the model to a network of workcenters in a transient regime.

### 7.2.5 Empirically Based Single-Variable Clearing Functions

These are functional forms that have been used to fit clearing functions empirically to data obtained from either industrial data or simulation. One or another of the clearing function families discussed above is used to postulate a basic functional form whose parameters are then fitted to empirical data gathered from either direct observation of the production unit or, more frequently, a simulation model.

Karmarkar (1989) proposes a workload-based clearing function of the form

$$X_t = \min\left\{\Lambda_t, \frac{K_1\Lambda_t}{K_2+\Lambda_t}\right\} \tag{7.23}$$

motivated by the clearing function for an *M/M/1* queue. Here $K_1$ represents the maximum expected output of the resource assuming unlimited workload and $K_2$ a user-determined parameter governing the curvature of the clearing function. In general, $K_2$ is increasing in the amount of variability in the system as described by the coefficients of variation of the service times and interarrival times. The clearing function is given as the minimum of two quantities to ensure that output does not exceed the total workload available to the resource; this can also be achieved by selecting the value of $K_2$ such that $\left.\frac{\partial X_t}{\partial \Lambda_t}\right|_{\Lambda_t=0} = 1$. This function is concave and monotonically non-decreasing, with $\lim_{\Lambda_t \to \infty} X_t = K_1$.

The functional form $X_t = K_1\Lambda_t/(K_2 + \Lambda_t)$ in (7.23) originates from the functional relationship between average WIP (in contrast to the workload $\Lambda_t$) and output; therefore, it can exceed the available workload in period $t$. Missbauer (2002) shows that for the *M/G/1* model in equilibrium, the expected output $X_t$ and expected load $\Lambda_t$ of a workcenter are related as follows:

$$X_t = \frac{1}{2}\left(K_1 + K_2 + \Lambda_t - \sqrt{K_1^{\ 2} + 2K_1K_2 + K_2^{\ 2} - 2K_1\Lambda_t + 2K_2\Lambda_t + \Lambda_t^{\ 2}}\right) \tag{7.24}$$

with $K_1$ the maximum expected output of the resource (capacity) as above, $K_2 = \dfrac{\sigma^2}{2t_{\mathrm{e}}} + \dfrac{t_{\mathrm{e}}}{2}$ and $\sigma^2$ the variance of the service times. This function is analogous to (7.21) and can be parameterized using empirical or simulated data.

Srinivasan et al. (1988) suggest an initial WIP-based clearing function similar to the flow-density function (7.16), given by

$$X_t = K_1\left[1 - \exp\left(-K_2W_{t-1}\right)\right] \tag{7.25}$$

Here $K_1$ again represents the maximum expected output of the resource with unlimited WIP, and $K_2$ a user-defined parameter governing the curvature of the clearing function. Once again we have $\lim\limits_{\Lambda_t \to \infty} X_t = K_1$.

Concave, saturating functional forms of clearing functions derived from queueing models usually approach their limit (the maximum possible expected output) asymptotically because the underlying assumptions of renewal processes usually allow arbitrarily long interarrival and service times. In reality, this is often not the case since the order release system will try to prevent very long interarrival times and service times can be controlled by lot sizing. Nyhuis and Wiendahl (2009) suggest defining threshold values $\overline{W}^{\mathrm{u}}$ and $\overline{W}^{\mathrm{o}}$ with $\overline{W}^{\mathrm{u}} < \overline{W}^{\mathrm{o}}$ for the average WIP $\overline{W}$ where for $\overline{W} < \overline{W}^{\mathrm{u}}$ output is proportional to $\overline{W}$, as in the "Best Case" clearing function of Hopp and Spearman (2008), and for $\overline{W} > \overline{W}^{\mathrm{o}}$ the workcenter is fully utilized. Appropriate functional forms are derived. In order to apply this logic to a period-based clearing function with the workload $\Lambda_t$ as state variable, threshold values $\Lambda_t^{\mathrm{u}}$ and $\Lambda_t^{\mathrm{o}}$ with $\Lambda_t^{\mathrm{u}} \le \Lambda_t^{\mathrm{o}}$ are defined for the workload, leading to different clearing functions for different regimes of operation such that

$$X_t = \begin{cases} \Lambda_t,\ 0 < \Lambda_t \le \Lambda_t^{\mathrm{u}} \\ f\left(\Lambda_t\right),\ \Lambda_t^{\mathrm{u}} \le \Lambda_t \le \Lambda_t^{\mathrm{o}}. \\ C_t,\ \Lambda_t \ge \Lambda_t^{\mathrm{o}} \end{cases} \tag{7.26}$$

In this case, the problem of estimating the clearing function is essentially that of estimating its deviation from the ideal shape $X_t =$ Min $(\Lambda_t; C_t)$.

## 7.3 Piecewise Linear Single-Variable Clearing Functions

Many authors using single-variable clearing functions in optimization models have chosen to approximate the concave clearing function by outer linearization. This approach has several benefits: it allows the overall production planning model to take the form of a linear program, which is computationally tractable and scalable. In addition, piecewise linearization of a univariate clearing function proves extremely useful in the development of clearing function models for multiple-item systems. We shall present the ideas in this section using the workload-based clearing function as our vehicle, but the basic issues are relevant to all concave single-variable clearing functions.

It is well known in convex analysis, as a consequence of the Fenchel-Young Theorem (that any convex region can be represented as the supremum of its affine minorants) (Boyd and Vandenberghe 2009), that any convex function can be approximated to any desired degree of accuracy by the convex hull of a set of affine functions of the form

$$f^q\left(\Lambda_t\right)=\alpha^q\Lambda_t+\beta^q,\quad q=1,\ldots,Q \tag{7.27}$$

In order to reflect the concavity of the original clearing function $f(\Lambda_t)$, we assume that the segments have slopes such that $\alpha^1 \geq 1 > \alpha^2 > \ldots > \alpha^Q = 0$, and intercepts $0 = \beta^1 < \beta^2 < \ldots < \beta^Q$. The intercept $\beta^Q$ of the final segment represents the maximum possible expected output from the production unit in a time period, while the slope $\alpha^1$ of the first segment is bounded above by 1, since even at very low workloads there may be a nonzero probability of some work remaining incomplete at the end of the period if, for example, a large fraction of the workload arrives very late in the period.

Given a concave clearing function of whatever specific functional form, the problem of determining the best piecewise linear approximation can be formulated as an optimization problem in several different ways. We shall describe one such formulation described by Turkseven (2005), which we shall refer to as the trapezoidal formulation, to illustrate the basic approach. Imamoto and Tang (2008) present an alternative formulation that minimizes the maximum error of the piecewise linear approximation for a given number of segments.

For illustrative purposes, we shall consider the problem of obtaining the best piecewise linear approximation to a concave non-decreasing clearing function $f(\Lambda_t)$ using three linear segments of the form (7.27) as seen in Fig. 7.2. Let $t_q$, $q = 1, \ldots, Q$ denote the value of $\Lambda_t$ at which segments $q$ and $q$+1 intersect, and $a_q$, $q = 1, \ldots Q$ the value of $\Lambda_t$ at which the $q$th linear segment is tangent to the concave clearing function. Additionally we define $t_0 = 0$ and $t_{Q+1} = \Lambda_{\max}$, an upper limit on the workload considered. For given values of $\alpha^q$ and $\beta^q$ straightforward geometry gives

$$a_q=\beta_q\frac{\left(\alpha_{q+1}-\alpha_q\right)}{\left(\beta_{q+1}-\beta_q\right)}+\alpha_q \tag{7.28}$$

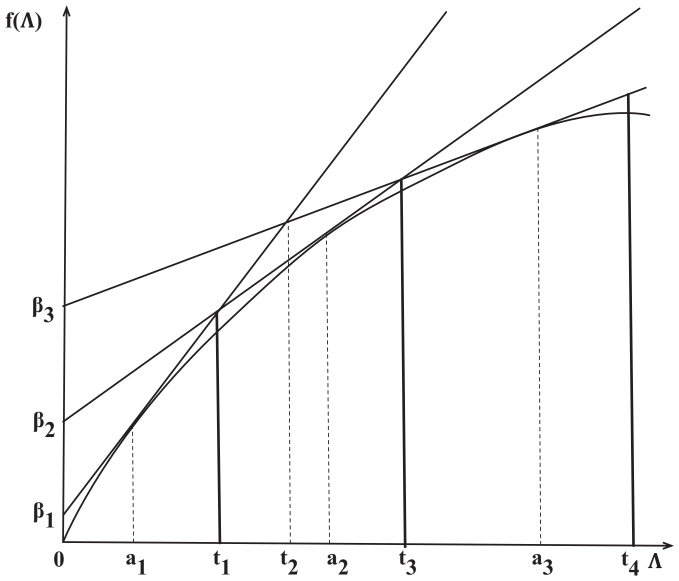


**Fig. 7.2** Illustration of trapezoid formulation for piecewise linearization of a concave clearing function

$$t_q = \frac{\alpha_{q+1} - \alpha_q}{\beta_{q+1} - \beta_q} \tag{7.29}$$

The decision variables in the optimization formulation are the slopes $\alpha^q$ and intercepts $\beta^q$ of the linear segments $q = 1, ..., Q$. The objective function to be minimized is given by the difference in the areas under the convex clearing function and its piecewise linear approximation, which is equivalent to minimizing the area under the piecewise linear approximation when the segments $q$ are constrained to be tangent to the original clearing function. For $Q$ linear segments, the area under the piecewise linear approximation will consist of $Q$ trapezoids, with the area of the trapezoid formed by segments $q$ and $q + 1$ given by

$$A_q = \beta^q + \frac{1}{2}\alpha^q\left(t_{q-1} + t_q\right) \tag{7.30}$$

The optimization model can then be written as

$$\min \sum_{q=1}^{Q} A_q \tag{7.31}$$

subject to

$$\alpha^q t_q + \beta^q = f\left(t_q\right), \quad q = 1, \ldots, Q \tag{7.32}$$

$$\beta^q = \frac{df(\Lambda)}{d\Lambda}\Big|_{\Lambda=a_q}, \quad q=1,\ldots,Q \tag{7.33}$$

$$\alpha^q, \beta^q \geq 0, \quad q=1,\ldots,Q \tag{7.34}$$

The fractional structure of (7.28) and (7.29) generally results in a non-convex nonlinear formulation for which a global optimum is hard to obtain in reasonable CPU times. Imamoto and Tang (2008) provide an exact recursive algorithm for their minimax formulation, while Turkseven (2005) proposes an alternative heuristic for the trapezoid formulation. Asmundsson et al. (2009) solve the trapezoid formulation with a standard convex nonlinear solver, obtaining a local optimum that appears satisfactory in most cases, although some instances where the solver failed to converge were also encountered.

We make no claim as to the originality of the trapezoid formulation; it is one of several fairly obvious approaches to the problem, and has almost certainly been formulated before, although we have been unable to find the original reference. We provide it here for the sake of completeness. However, recent work by Gopalswamy and Uzsoy (2019) suggests that rather than fitting a nonlinear functional form to data and then piecewise linearizing this concave function, directly fitting a piecewise linear concave function to the data using convex regression (Toriello and Vielma 2012; Hannah and Dunson 2013; Gopalswamy et al. 2019) yields considerably better results.

## 7.4 Optimization Models for a Single Production Resource

The clearing functions presented above all represent the system state in an aggregate manner; the workload $\Lambda_t$, the initial WIP $W_{t-1}$, or the time-average WIP $\overline{W}_t$ are aggregated over the different products in the system, in a manner similar to that used by queueing models of multi-item systems: the mix of different items arriving randomly at the resource over time results in the effective service times following a probability distribution whose first and second moments can be used to derive a clearing function. However, any useful production planning model must determine the mix of products to be released into the system in each planning period $t$, requiring disaggregation if an aggregate single-variable clearing function is used. The development of clearing function models for multiple-item systems presents a number of challenges; similar issues are encountered in traffic modeling with multiple vehicle classes or origin-destination pairs (Carey 1992). These difficulties have proven to be persistent in both research areas, and merit detailed discussion since a fully satisfactory solution remains elusive.

To illustrate the issues, we first present a model of a simple single-product problem, closely following the development of Karmarkar (1989). For ease of exposition, we assume a time-stationary workload-based clearing function $f(\Lambda_t)$ and time-stationary cost parameters. We also assume no backlogging of unmet demand is allowed; if present, it can be incorporated easily (Johnson and Montgomery 1974). We define the following notation:

**Indices:**

$t$: planning period, $t = 1,\ldots,T$. $t = 0$ will be used to denote the initial state of the system at the start of period 1, i.e., the end of period 0.

**Parameters:**

$c$: unit production cost
$h$: unit finished goods inventory holding cost
$w$: unit WIP holding cost
$r$: unit cost of raw materials, incurred upon release of the material to the production unit
$f(\Lambda_t)$: clearing function representing the behavior of the production unit, which we assume to be a concave monotonically non-decreasing function of $\Lambda_t$
$D_t$: demand in period $t$
$I_0$: amount of product in finished goods inventory at the start of period 1
$W_0$: amount of product in WIP at the start of period 1

*Decision Variables*:

$X_t$: output of production unit in period $t$, in units of product
$R_t$: amount of product released into production unit in period $t$
$I_t$: amount of product in finished goods inventory at the end of period $t$
$W_t$: amount of product remaining in WIP at the end of period $t$

In the fixed lead time models of Chap. 5, material released at the start of period $t$ subject to a fixed lead time $L$ emerges as finished product at the start of period $t+L$. Thus the output of the production unit is simply the time-shifted release schedule. However, in clearing function models the output of the resource in a given period $t$ is driven only indirectly by the releases $R_t$. In a given period $t$, the resource is assumed to have $W_{t-1} \geq 0$ units of WIP remaining from the previous period. $R_t$ units of product are released to the resource, resulting in a workload of $\Lambda_t = W_{t-1} + R_t$ units. The output of the resource during this period $t$ is then determined by the clearing function as $X_t = f(\Lambda_t)$. These dynamics yield the following single-product clearing function (SPCF) model:

$$\min \sum_{t=1}^{T} \left[ rR_t + cX_t + hI_t + wW_t \right] \tag{7.35}$$

subject to

$$W_t = W_{t-1} + R_t - X_t, \quad t = 1,\ldots,T \tag{7.36}$$

$$I_t = I_{t-1} + X_t - D_t, \quad t = 1,\ldots,T \tag{7.37}$$

$$X_t \leq f\left(\Lambda_t\right), \quad t = 1,\ldots,T \tag{7.38}$$

$$R_t, X_t, I_t, W_t \geq 0, \quad t = 1,\ldots,T \tag{7.39}$$

The objective function (7.35) minimizes the sum of raw material, production, finished goods holding and WIP holding costs over the planning horizon of $T$ periods. Constraints (7.36) are material balance equations for the WIP, and constraints (7.37) those for finished goods inventory. Constraints (7.38) limit the output in each period by the clearing function, while (7.39) ensure nonnegativity of the decision variables. Like most of the LP models discussed in Chap. 5, the SPCF model can be represented as a network flow model on a time-replicated network as shown in Fig. 7.3.

Several differences from the models of Chap. 5 are worth highlighting. First of all, no lead times appear in the formulation; the delay between material being released and its emergence as finished product capable of meeting demand is implied by the clearing function constraints (7.38). Since the argument of the clearing function depends on the WIP variables $W_t$, material balance constraints (7.36) are required to keep track of these variables. This distinction between WIP and finished goods inventory is intuitive, since in practice these inventories serve different purposes. Production is made possible by having sufficient WIP in the system, while finished goods inventory, represented by the $I_t$ variables, allows inventories to be built up in anticipation of future demand peaks.

While appearing deceptively simple, the SPCF model already involves a number of subtleties. The reader will have noticed that the output constraints (7.38) are written in inequality form; this is because writing them as equalities results in a non-convex feasible region (Merchant and Nemhauser 1978) as seen in the following example:

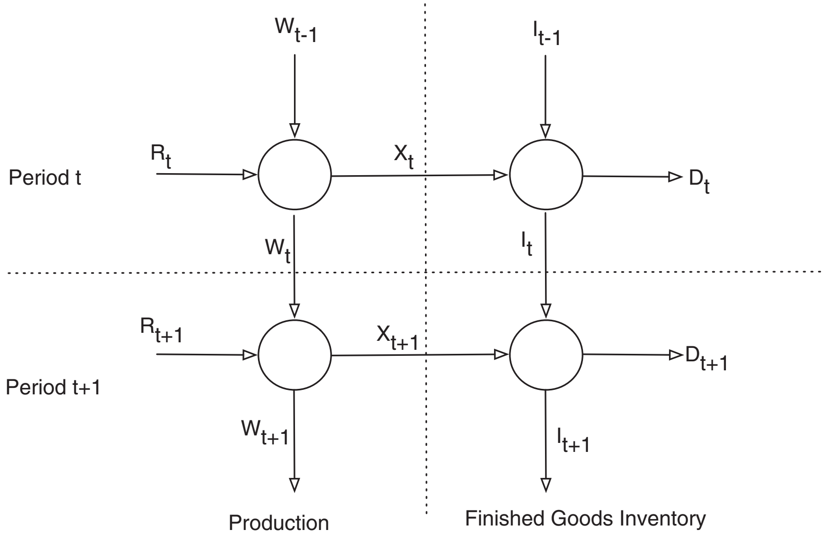


**Fig. 7.3** Material Flows in Single-Product CF Model (Karmarkar 1989)

**Example 7.1** Consider a two-period production planning problem with $D_1 = 3$, $D_2 = 9$, $I_0 = W_0 = 0$ and a clearing function

$$f\left(\Lambda_t\right) = \frac{10\Lambda_t}{10+\Lambda_t} = \frac{10\left(W_{t-1}+R_t\right)}{10+W_{t-1}+R_t}, \quad \Lambda_t \geq 0 \tag{7.40}$$

Consider the two solutions $Y^1$ and $Y^2$ summarized in Table 7.1.

Now consider a solution $Y^3 = 0.3Y^1 + 0.7Y^2$. The reader can easily verify that $Y^3$ satisfies the material balance constraints (7.36) and (7.37). However, $X_3^1 = 0.3X_1^1 + 0.7X_1^2 = 5.67 < f\left(0.3\left(R_1^1 + W_0\right) + 0.7\left(R_1^2 + W_0\right)\right) = f\left(15.5\right) = 6.07$. Similarly, $X_2^3 = 7.51 < f\left(0.3\left(90\right)+0.7\left(40\right)\right) = 7.84$. Thus $Y^3$ is not a feasible solution when the clearing function constraints (7.38) are enforced at equality, indicating a non-convex feasible region.

The slack variables associated with constraints (7.38) represent a situation where the resource is not producing the maximum output it is capable of given the workload available to it; it is holding back some WIP that it is capable of converting into output because of adverse consequences in future periods. The following example illustrates this behavior of the SPCF model.

**Example 7.2** Consider a two-period instance of the SPCF model with $r = 1$, $w = 2$, $h = 3$, $c = 1$, $D_1 = 9$, and $D_2 = 3$. Assume the same workload-based clearing function used in Example 7.1. The optimal solution to this instance is illustrated in Fig. 7.4.

**Table 7.1** Data for Example 7.1

| Solution | $R_1^i$ | $R_2^i$ | $W_1^i$ | $W_2^i$ | $\Lambda_1^i$ | $\Lambda_2^i$ | $X_1^i$ | $X_2^i$ | $I_1^i$ | $I_2^i$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $Y^1$ | 5 | 65 | 1.67 | 57.97 | 5 | 66.67 | 3.33 | 8.70 | 0.33 | 0.03 |
| $Y^2$ | 20 | 10 | 13.33 | 16.33 | 20 | 23.33 | 5.67 | 7.51 | 2.67 | 1.17 |

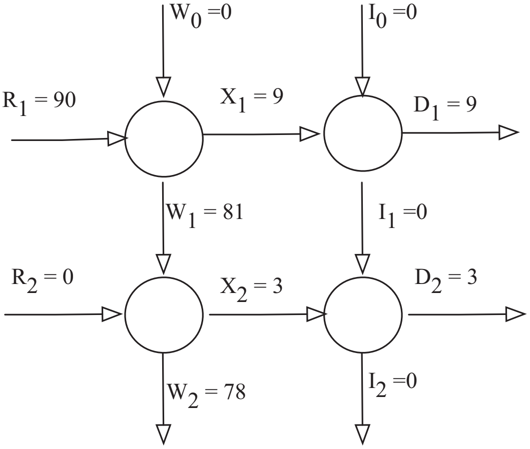


**Fig. 7.4** Example of optimal solution with slack in CF constraints for Example 7.2

The high demand in period 1 requires the release of a large amount of work in that period to raise the workload to a level allowing output to meet the demand. The concave shape of the clearing function results in a large amount of WIP remaining at the end of period 1. However, the low demand in period 2 can be met with only three units of production. Since processing a unit of WIP to pass it into finished goods inventory incurs a total unit cost of $c = 1$ for production and $h = 3$ for holding the resulting finished inventory, it is cheaper to hold the excess material as WIP, resulting in a production of $X_2 = 3 < f(81) = 8.9$ units. Note that the behavior remains the same even if $c = 0$; simply having $h > w$ is sufficient.

Carey (1987) shows that constraints (7.38) will be satisfied as equalities as long as the marginal cost $c + (h - w)$ of moving material from WIP to finished goods in the absence of demand for it is nonpositive. To see this, note that in the network representation in Fig. 7.4, there are exactly two arcs incident out of the node corresponding to the WIP balance equation (7.36). Material in WIP at the start of period $t$ is either retained in WIP in the next period (the vertical arc), or produced and moved to FGI (the horizontal arc). When a unit of WIP is converted into output and remains in finished inventory at the end of the period, the production cost of $c$ is incurred, and the total holding cost in that period increases by $(h - w)$. An item for which there is external demand in the period will not incur the FGI holding cost $h$, and will be produced even if $c > w$ since otherwise demand will not be met, resulting in an infeasible solution in the absence of backlogging.

This holding back behavior can be explained in the context of traffic modeling as avoiding the release of traffic from one road segment to prevent congestion at downstream segments, say using traffic lights to regulate the flow of traffic. However, in production systems it is uncommon to hold WIP within the production process (as opposed to inventory points where intermediate products can be stored) without processing it if the capacity to process it is available, unless it is on hold due to quality or engineering problems. Thus this holding back behavior needs to be considered when implementing clearing function based planning models. The simplest approach is to set WIP holding costs sufficiently high ($w > c + h$ in this example) to ensure it is cheaper to move material downstream rather than retain it in the queue for a given process as WIP; after all, this is how production managers seem to behave in practice. However, this contradicts conventional cost accounting practice under which the holding cost of an item increases as it moves towards completion, due to the increasing value added during production. While it can be justified in some situations, such as semiconductor wafer fabrication where the high cost and limited availability of clean room space makes holding strategic inventory inside the factory undesirable, the manipulation of costs in this manner needs to be considered carefully in the context of the economics of the production system under study.

We have just seen that the SPCF model exhibits interesting behavior when restricted to a single-stage production system. We now explore its obvious extensions to multistage single-product and single-stage multiple product systems.

## 7.5 Multistage Single-Product Systems

The SPCF model (7.35)–(7.39) can be extended to multistage single-product environments in a straightforward manner by defining an index $n$ denoting the stage of the production process. Thus a product is assumed to require a total of $n = 1, 2, ..., N$ operations whose sequence, or routing, is known and deterministic. However, this requires addressing the issue raised above of whether strategic inventory can be held between production stages or only at the output of the final stage. We shall first examine the model assuming such inventory cannot be held at intermediate locations, and then briefly discuss the case where such inventory can be held. For simplicity of exposition, we shall assume that each stage of the production process or routing corresponds to a distinct resource, each represented by its own clearing function. The extension to reentrant flows, where the product may undergo multiple operations at the same workcenter, is straightforward and can be addressed in exactly the same manner used for conventional models with fixed lead times (Leachman 2001; Kacar et al. 2016).

The parameters and decision variables remain the same as those in the SPCF model, except for the addition of an index $n$ denoting the stage of the production process to which they refer. Demand can only be met with the output of stage $N$, and we shall again assume no backlogging of missed demand.

Our first model assumes that no inventory can be held within the production unit for tactical purposes such as anticipation of a future demand peak; such inventory is only held after stage $N$ and consists of finished goods that can be used to meet demand. In this situation, work is released into the system at stage $n = 1$; the input $Y_{nt}$ to stages $n > 1$ in a period $t$ is given by the output of the previous stage in that period, i.e., $Y_{nt} = X_{n-1,t}$ in the notation of Chap. 6. This single-product multistage clearing function model (SPMCF) can be written as follows:

$$\min \sum_{t=1}^{T} \left[ rR_t + hI_t + \sum_{n=1}^{N} \left( c^n X_t^n + w^n W_t^n \right) \right] \tag{7.41}$$

subject to

$$W_t^1 = W_{t-1}^1 + R_t - X_t^1, \quad t = 1, \ldots, T \tag{7.42}$$

$$W_t^n = W_{t-1}^n + X_t^{n-1} - X_t^n, \quad n = 2, \ldots, N, \quad t = 1, \ldots, T \tag{7.43}$$

$$I_t = I_{t-1} + X_t^N - D_t, \quad t = 1, \ldots, T \tag{7.44}$$

$$X_t^n \le f^n \left( \Lambda_t^n \right), \quad n = 1, \ldots, N, \quad t = 1, \ldots, T \tag{7.45}$$

$$I_t, W_t^n, X_t^n, R_t \ge 0, \quad n = 1, \ldots, N, \quad t = 1, \ldots, T \tag{7.46}$$

The decision variables and constraints in this model are analogous to those in the single-stage SPCF model (7.35)–(7.39). The WIP balance constraint (7.43) is written for stages 2, …, $N$, where the output from the previous stage $n - 1$ provides the input of new work entering the stage. The WIP balance constraint (7.42) for Stage 1 is written using the release variables $R_t$ representing external releases of new work into the production unit. Constraints (7.44) represent the material balance equations for the finished goods inventory held after the final stage $N$.

The objective function of this model is straightforward; our interest lies in the constraint set which attempts to model the behavior of a multistage production unit. The following example illustrates the behavior of these constraints.

**Example 7.3** To illustrate the behavior of the constraint set (7.42)–(7.46), consider a serial production system consisting of five identical stages. Each stage is modeled by the workload-based clearing function $f(\Lambda_t) = 10\Lambda_t/(10 + \Lambda_t)$ used in the previous examples. Assuming $W_0^n = 0$ for all $n = 1, \ldots, N$, we release $R_1 = 10$ units of work into the first stage in the first period, with $R_t = 0$ for all remaining $t > 1$. Table 7.2 shows the evolution of the system state and output over time, while Fig. 7.5 illustrates the output of each stage.

Several interesting observations emerge from Table 7.2. 16.7% of the material released at the start of period 1 exits the overall system in the period in which it is released, traversing all five stages in a single period. This is analogous to Equation (4.6) in the discussion of load-oriented order release that estimates the fraction of the workload released in a certain period that traverses the first $n$ workcenters on its routing within the same period. Given our assumption of instantaneous material transfer between stages and the fact that all stages are empty at the start of period 1, this behavior seems reasonable. It requires slightly more than six periods for all material released to exit the system. The small ending WIP levels at stages 4 and 5 and the end of period 6 are due to the fact that $f(\Lambda_t) < \Lambda_t$ for the CF in the example.

In order to estimate the average cycle time at each stage in each period, we shall use the expression

$$T^n = \begin{cases} 0.1\Lambda_t^n, \text{if } \Lambda_t^n < 0.25 \\ \dfrac{\left(W_{t-1}^n + W_t^n\right)}{2X_t^n}, \text{otherwise} \end{cases} \tag{7.47}$$

Assuming that all quantities are given in units of product, the clearing function implies that the maximum possible output from each stage in a planning period is 10 units, for an average processing time per unit of 0.1 periods. The first expression represents the situation where the workload is sufficiently low that the entire available workload $\Lambda_t^n$ can be converted into output in the same period. This is a slight approximation, since the slope of the CF at the origin is equal to 1 and is decreasing in $\Lambda_t^n$; however, at $\Lambda_t^n = 0.25$ the clearing function posits an output of 0.2439 units of product, an error of 2.5%. The second term estimates cycle time using Little's Law, where the time-average WIP level in a period is estimated as the arithmetic

**Table 7.2** Response of Empty System to Step Input

| $t$ | $R_t$ | $\Lambda_1$ | $X_1$ | $W_1$ | $\Lambda_2$ | $X_2$ | $W_2$ | $\Lambda_3$ | $X_3$ | $W_3$ | $\Lambda_4$ | $X_4$ | $W_4$ | $\Lambda_5$ | $X_5$ | $W_5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | 0 | | | 0 | | | 0 | | | 0 | | | 0 |
| 1 | 10 | 10.000 | 5.000 | 5.000 | 5.000 | 3.333 | 1.667 | 3.333 | 2.500 | 0.833 | 2.500 | 2.000 | 0.500 | 2.000 | 1.667 | 0.333 |
| 2 | 0 | 5.000 | 3.333 | 1.667 | 5.000 | 3.333 | 1.667 | 4.167 | 2.941 | 1.225 | 3.441 | 2.560 | 0.881 | 2.894 | 2.244 | 0.649 |
| 3 | 0 | 1.667 | 1.429 | 0.238 | 3.095 | 2.364 | 0.732 | 3.589 | 2.641 | 0.948 | 3.522 | 2.605 | 0.917 | 3.254 | 2.455 | 0.799 |
| 4 | 0 | 0.238 | 0.233 | 0.006 | 0.964 | 0.879 | 0.085 | 1.827 | 1.545 | 0.282 | 2.462 | 1.976 | 0.487 | 2.775 | 2.172 | 0.603 |
| 5 | 0 | 0.006 | 0.006 | 0.000 | 0.090 | 0.090 | 0.001 | 0.372 | 0.359 | 0.013 | 0.845 | 0.779 | 0.066 | 1.382 | 1.214 | 0.168 |
| 6 | 0 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.000 | 0.014 | 0.014 | 0.000 | 0.080 | 0.079 | 0.001 | 0.247 | 0.241 | 0.006 |
| 7 | 0 | 0.000 | 0.000 | 0.0000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.000 | 0.007 | 0.007 | 0.000 |
| 8 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

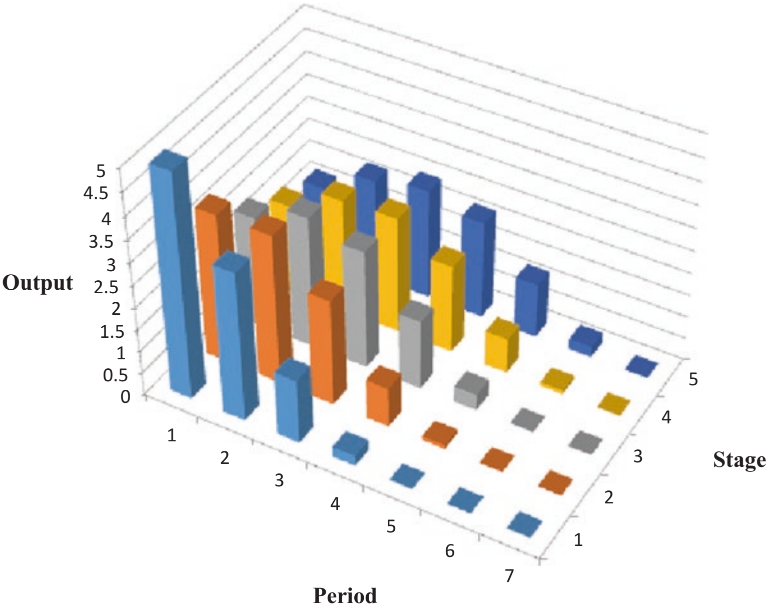


**Fig. 7.5** Output by Stage and Period for Empty System in Example 7.3

average of its beginning and ending WIP levels. These estimates of cycle time are clearly crude averages; in particular, the use of Little's Law implies that the queue representing each stage in each period is in steady state, which requires, at the very least, long planning periods. If required, a relationship analogous to (6.5) can be used.

However, within these limitations, the results are still interesting, as shown in Fig. 7.6. The cycle time estimates at each stage increase in the early periods, as material arrives, and then decrease as the released material flows out of the system and is not replaced. The cycle time at each stage varies over time, highlighting the difficulties of using exogenous lead times in planning models. If we were to assume that each stage had a fixed lead time of 1 period and a maximum production capacity of 10 units per period—compatible with the clearing function used in the example—each stage *n* would produce an output of 10 units in period *n,* a completely different profile from that illustrated in Fig. 7.5.

For comparison, consider the situation illustrated in Table 7.3 and Fig. 7.7, where we again release 10 units into the system at the start of period 1, but each stage has 10 units in WIP at the start of that period. The combination of previous WIP and new releases results in a workload of $\Lambda_n = 20$ units at each stage *n* at the start of period 1. It now requires 12 periods for all work to exit the system. The output of all stages decreases over time, since the material released at the start of period 1 increases the output of all stages in that period, and hence moves material downstream to all stages in the subsequent periods. The additional 10 units of input at the

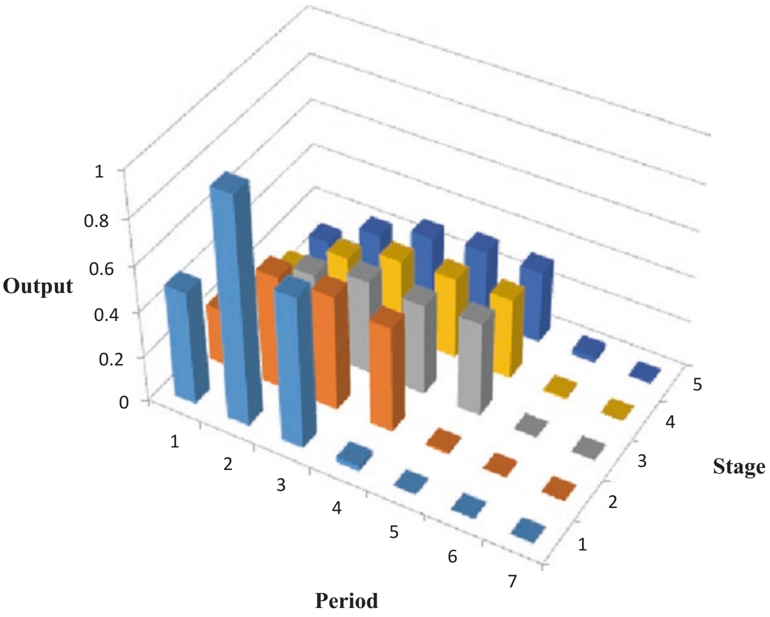


**Fig. 7.6** Estimated Cycle Times for Example 7.3

start of period 1 increase the output of each subsequent stage in period 1 by less than 2 units; note that if there were no new releases, the output of each stage in period 1 would be 5 units. However, this does not imply that 16% of the newly released material completes all its processing in period 1. If we assume first-in-first-out processing at each stage, no new material is processed at Stage 1 in period 1; there are 10 units of WIP at the start of the period, of which only 6.667 units are converted into output. This is due to the relatively flat clearing function, which requires $\Lambda = 1000$ units to achieve an output of 9.9 units (Fig. 7.8).

The cycle times are now substantially higher than was the case with an empty system. The relatively slow decrease in the cycle times at all stages in periods 1 through 3 is noteworthy; after period 5, though, as the workload decreases the cycle time decreases rapidly. The relative stability of the cycle times in the early periods provides some insight into why fixed lead time models can work well under many situations: as long as the workload does not vary greatly from period to period, cycle times may remain stable, allowing a fixed lead time to provide a sufficiently accurate solution, especially if fractional lead times as suggested by Hackman and Leachman (1989) are used (Kacar et al. 2016).

This example provides a qualitative illustration of the behavior of the constraints (7.42)–(7.46) that represent the behavior of the production system using clearing functions, particularly the strong differences from the fixed lead time models in Chap. 5. We now extend this model to the case of multiple products competing for capacity at the resource, which proves to be treacherous territory.

**Table 7.3** Response of Loaded System to Step Input

| $t$ | $R_t$ | $\Lambda_1$ | $X_1$ | $W_1$ | $\Lambda_2$ | $X_2$ | $W_2$ | $\Lambda_3$ | $X_3$ | $W_3$ | $\Lambda_4$ | $X_4$ | $W_4$ | $\Lambda_5$ | $X_5$ | $W_5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | 10 | | | 10 | | | 10 | | | 10 | | | 10 |
| 1 | 10 | 20.000 | 6.667 | 13.333 | 16.667 | 6.250 | 10.417 | 16.250 | 6.190 | 10.060 | 16.190 | 6.182 | 10.009 | 16.182 | 6.181 | 10.001 |
| 2 | 0 | 13.333 | 5.714 | 7.619 | 16.131 | 6.173 | 9.958 | 16.233 | 6.188 | 10.045 | 16.197 | 6.183 | 10.014 | 16.184 | 6.181 | 10.003 |
| 3 | 0 | 7.619 | 4.324 | 3.295 | 14.282 | 5.882 | 8.400 | 15.926 | 6.143 | 9.784 | 16.157 | 6.177 | 9.980 | 16.180 | 6.180 | 10.000 |
| 4 | 0 | 3.295 | 2.478 | 0.817 | 10.879 | 5.210 | 5.668 | 14.994 | 5.999 | 8.995 | 15.979 | 6.151 | 9.828 | 16.150 | 6.176 | 9.974 |
| 5 | 0 | 0.817 | 0.755 | 0.062 | 6.423 | 3.911 | 2.512 | 12.906 | 5.634 | 7.272 | 15.463 | 6.073 | 9.390 | 16.047 | 6.161 | 9.886 |
| 6 | 0 | 0.062 | 0.061 | 0.000 | 2.573 | 2.047 | 0.527 | 9.318 | 4.824 | 4.495 | 14.213 | 5.870 | 8.343 | 15.756 | 6.117 | 9.639 |
| 7 | 0 | 0.000 | 0.000 | 0.0000 | 0.527 | 0.501 | 0.026 | 4.995 | 3.331 | 1.664 | 11.675 | 5.386 | 6.288 | 15.025 | 6.004 | 9.021 |
| 8 | 0 | 0.000 | 0.000 | 0.000 | 0.026 | 0.026 | 0.000 | 1.690 | 1.446 | 0.244 | 7.734 | 4.361 | 3.373 | 13.382 | 5.723 | 7.659 |
| 9 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.244 | 0.239 | 0.006 | 3.612 | 2.653 | 0.958 | 10.313 | 5.077 | 5.236 |
| 10 | 0 | 0.000 | 0.000 | 0000 | 0.000 | 0000 | 0.000 | 0006 | 0.006 | 0.000 | 0.964 | 0.879 | 0.085 | 6.115 | 3.795 | 2.320 |
| 11 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.085 | 0.084 | 0.001 | 2.404 | 1.938 | 0.466 |
| 12 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.000 | 0.467 | 0.446 | 0.021 |
| 13 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.021 | 0.021 | 0.000 |
| 14 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

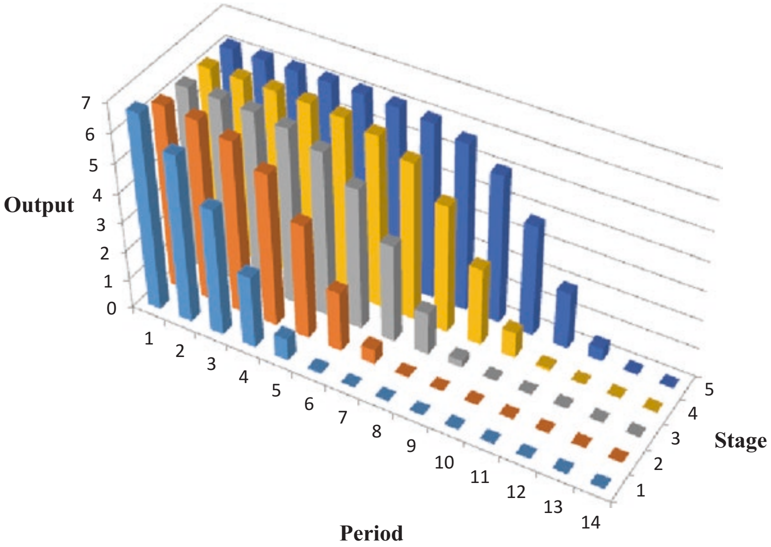


**Fig. 7.7** Output of Loaded System in Example 7.3

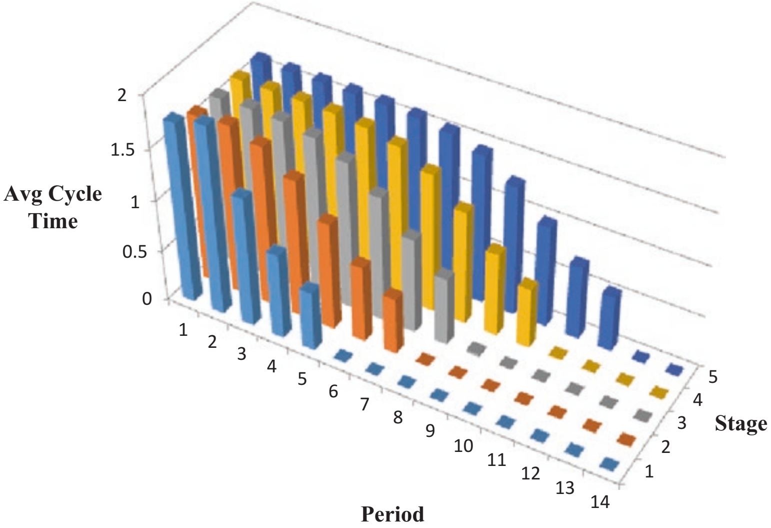


**Fig. 7.8** Estimated Cycle Time of Loaded System Under Step Input

## 7.6 Single-Variable Clearing Functions with Multiple Products

The presence of multiple items brings the need to allocate the output of the resources among the different items. The use of a single-variable clearing function implies that the output of a resource is determined by its total workload, and hence that the amount of output of each item must depend in some way on the workload of the other items. On the shop floor, the mix of items produced for a given workload of each item is determined by events on the shop floor, such as the arrival times of specific jobs at specific machines as well as scheduling and dispatching decisions. Since these operational policies are internal to the production unit, and hence not transparent to the planning level, it is difficult to model these directly in the planning problem. Even if the performance induced by these policies in the production unit could be described in computationally tractable models, it is not evident that it would be beneficial to do so, since local management will have the most current information on the status of the shop floor, and is responsible for managing the production unit in the face of changing local conditions. Hence a reasonable objective for the planning problem is to produce a release plan for each production unit that does not violate the basic constraints viewed by management as essential for the release plan to be usable.

One such set of constraints that has been widely discussed in the context of both production planning and traffic modeling is the maintenance of basic continuity conditions on the material flow. In both the traffic and production contexts, these can be expressed as a requirement that material entering the production unit earlier ought to exit earlier. Some deviation from this condition at the level of individual orders is clearly possible, and even desirable, in practice due to the ability of local management to expedite the processing of some jobs over others. Hence it ought to be sufficient for planning models to satisfy this requirement on average, while avoiding gross violations. Several sets of necessary and sufficient conditions for this first in first out (FIFO) property derived by Carey (1992) were discussed in Chap. 6, noting that they all lead to non-convex feasible regions.

We shall begin our discussion of multi-item models with single-variable clearing functions by presenting a naive extension of the SPCF model (7.35)–(7.39), to illustrate the difficulties that arise. We then discuss several solution approaches, most suggested in the context of traffic modeling (Carey 1992; Carey and Subrahmanian 2000a, b) which result in non-convex formulations. Finally, we present the allocated clearing function (ACF) model of Asmundsson et al. (2006, 2009), which provides a workable solution to these difficulties in the limited context of a single-variable clearing function.

### *7.6.1 Difficulties with Multiple Items*

At first sight, extending the SPCF model to multiple items appears quite straightforward: we should add an item index $i$, write WIP balance and finished goods inventory balance equations for each item and add a clearing function constraint shared across all items. We use the following notation in addition to that already defined:

*Indices***:**

$i$: item index, $i = 1,\ldots, I$

*Parameters*:

$c_i$: unit production cost of item *i*
$h_i$: unit finished goods inventory holding cost for item *i*
$w_i$*:* unit WIP holding cost for item *i*
$r_i$: unit cost of raw materials for item *i*, incurred upon release of the material to the resource
$a_i$: amount of time required on the resource to produce one unit of item *i*
$f(\Lambda_t)$*:* the clearing function, which we assume to be a concave, monotonically non-decreasing function of the total workload $\Lambda_t$
$D_{it}$*:* demand for item *i* in period *t*
$I_{i0}$: number of units of item *i* in finished goods inventory at start of period 1
$W_{i0}$: number of unprocessed units of item *i* in WIP at the start of period 1

*Decision Variables:*

$X_{it}$: output of item *i* in period *t*, in units of product
$R_{it}$: number of units of item *i* released to the resource in period *t*
$I_{it}$: number of units of item *i* remaining in finished goods inventory at the end of period *t*
$W_{it}$: number of units of item *i* in WIP at the end of period *t*
$\Lambda_{it}$ : workload due to item *i* in period *t*, given by $a_i(R_{it}+W_{i,t\text{-}1})$ in units of time
$\Lambda_t$*:* total workload available to resource at the start of period *t*, given by

$$\Lambda_t = \sum_{i=1}^{I} a_i \Lambda_{it} \tag{7.48}$$

As in the previous examples, we assume time-stationary values of all parameters for simplicity of exposition. The Naive extension of the SPCF model to multiple items, which we shall refer to as the NSPCF model, can now be written as:

$$\min \sum_{t=1}^{T} \sum_{i=1}^{I} \left[ r_i R_{it} + w_i W_{it} + c_i X_{it} + h_i I_{it} \right] \tag{7.49}$$

subject to:

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.50}$$

$$W_{it} = W_{i,t-1} + R_{it} - X_{it}, \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.51}$$

$$\sum_{i=1}^{I} a_i X_{it} \le f\left(\Lambda_t\right), \quad t = 1,\ldots,T \tag{7.52}$$

$$R_{it}, X_{it}, I_{it}, W_{it} \ge 0, \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.53}$$

**Table 7.4** Parameter values for Example 7.4

| Item $i$ | $a_i$ | $r_i$ | $c_i$ | $w_i$ | $h_i$ | $W_{i0}$ | $I_{i0}$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 6 | 0 | 0 | 0 |

**Table 7.5** Demand data for Example 7.4

| Item $i$ | Period 1 | Period 2 | Period 3 | Period 4 | Period 5 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 8 | 8 | 8 | 8 | 8 |

The decision variables and constraints of this model are completely analogous to those in the SPCF model. However, the treatment of the clearing function causes some interesting difficulties, which we illustrate in the following example.

**Example 7.4** Consider a problem instance with two items, five periods, the parameter values shown in Table 7.4, and the demand data in Table 7.5. We again use the clearing function $f(\Lambda_t) = 10\Lambda_t/(10 + \Lambda_t)$ used in the previous examples.

Solving the model (7.49)–(7.53) yields the solution in Table 7.6. The problem is evident upon inspecting the results. There is no demand for item 1 in any period, and yet 29.93 units of Item 1 are released into the system, none of which is converted into output. The total workload generated by both products is used to meet the demand for item 2 with minimum WIP holding cost. Note that in periods 1 and 2, the model elects to produce less output than the clearing function allows: 8.18 units in period 1 as opposed to the 8.35 units the clearing function allows for the available workload of $\Lambda_t = 50.59$ units.

The problem is now clear: the WIP of item 1 is being held stationary in the system to artificially raise the available workload and permit the expensive item 2 to pass through the system rapidly. The planned releases of item 2 cannot, on their own, generate sufficient workload to produce the planned output of item 2. Simply parking idle WIP on the shop floor is increasing the output capacity of the system!

The reason for this behavior is also apparent. There is nothing in the model that links the output of the system to the composition of the workload enabling that output. This can be interpreted as a violation of the no-passing condition mentioned above—we are allowing the new releases of item 2 to constantly overtake the material of item 1 released in period 1. While in any practical production system some overtaking will arise naturally through the operation of shop-floor scheduling and dispatching systems, the idea that holding inert, idle WIP in the system increases the capability of the resources is clearly unrealistic.

As seen in Example 7.1, the non-convexity of clearing function models for single-item formulations enforcing the clearing function constraints as equalities was identified by Merchant and Nemhauser (1978). Carey (1987) demonstrated that implementing the clearing (exit) functions as inequalities results in a convex optimization problem for the single-item case, and discusses the issue of flow controls, where the clearing function constraints may hold as strict inequalities. He shows that the holding back behavior discussed in Example 7.2 will be avoided if the

**Table 7.6** Optimal solution for Example 7.4

| Period | $R_1$ | $R_2$ | $W_1$ | $W_2$ | $\Lambda_1$ | $\Lambda_2$ | $\Lambda$ | $f(\Lambda)$ | $X_1$ | $X_2$ | $aX$ | $I_1$ | $I_2$ | $D_1$ | $D_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | 0 | 0 | | | | | | | | 0 | 0 | | |
| 1 | 29.93 | 20.65 | 29.93 | 12.47 | 29.93 | 20.65 | 50.59 | 8.35 | 0.00 | 8.18 | 8.18 | 0.00 | 0.18 | 0 | 8 |
| 2 | 0.00 | 3.72 | 29.93 | 8.12 | 29.93 | 16.19 | 46.13 | 8.22 | 0.00 | 8.08 | 8.08 | 0.00 | 0.26 | 0 | 8 |
| 3 | 0.00 | 0.00 | 29.93 | 0.20 | 29.93 | 8.12 | 38.05 | 7.92 | 0.00 | 7.92 | 7.92 | 0.00 | 018 | 0 | 8 |
| 4 | 0.00 | 7.71 | 29.93 | 0.00 | 29.93 | 7.91 | 37.84 | 7.91 | 0.00 | 7.91 | 7.91 | 0.00 | 0.09 | 0 | 8 |
| 5 | 0.00 | 7.91 | 29.93 | 0.00 | 29.93 | 7.91 | 37.84 | 7.91 | 0.00 | 7.91 | 7.91 | 0.00 | 0.00 | 0 | 8 |

marginal cost of moving flow downstream (in our context, moving material from WIP to finished goods inventory) is lower than that of holding it at its current location. Carey discusses this issue in the context of modeling traffic flows and suggests a number of options. We digress briefly to discuss several of these, since they highlight a number of issues arising in optimization models involving flows through production networks. Our discussion follows that of Carey (1992), adapting the notation to the production planning models of interest in this work.

### 7.6.2 Enforcing Average No-Passing (FIFO) Behavior

Returning for a moment to the single-item problem, let $R_{st}$ denote the amount of material released in period $s$ that is converted into output in period $t$. Thus, in the notation of the SPCF model, we have

$$X_t = \sum_{s=1}^{t} R_{st} \tag{7.54}$$

One way of enforcing a no-passing condition is to ensure that material released earlier cannot be converted into output (i.e., transition from WIP to finished goods inventory) after material that is released later. This implies a condition that

$$R_{st} > 0 \rightarrow \sum_{\{s'<s,t'>t\}} R_{s',t'} = 0 \tag{7.55}$$

If we have multiple items $i = 1,...,I$, (7.55) must hold for each item $i$, as well as all pairs of items $i$ and $j$. The explicit enforcement of this condition requires the use of integer variables to represent what are effectively disjunction constraints, resulting in computationally demanding integer programming models.

Intuition suggests that the no-passing property is likely to be violated if there are large changes in cycle times from one period to the next. This would suggest that as long as the cycle times (flow times in the traffic terminology) are “smooth enough” over time, violations of no-passing ought to be at least mitigated. As seen in Chap. 6, we can calculate the average cycle time for material released into the system in period $s$ as

$$\overline{L}_s = \frac{\sum_{\tau=s}^{T} (\tau - s) R_{s\tau}}{\sum_{\tau=s}^{T} R_{s\tau}} \tag{7.56}$$

which is non-convex in the $R_{st}$ variables (Carey 1992). Thus the average unit of work emerging as finished goods inventory at time $t + \overline{L}_t$ must have entered the system at time $t$. Carey (1992) then shows that the condition

$$\overline{L}_t \leq \overline{L}_{t+1} + 1 \tag{7.57}$$

is necessary and sufficient to ensure no-passing of the average flows, and necessary but not sufficient to ensure no-passing on all individual work releases in the single-item case. This implies that the possibility of passing only arises when cycle times associated with the release periods are decreasing over time, i.e., the workload in the system is decreasing. As was the case for (7.55), in the presence of multiple items, this requires similar constraints to be written for each item $i$ and all pairs of items $i$, requiring $O(I^2)$ additional non-convex constraints in each period where $I$ is the number of items. In the presence of multiple items, we can enforce no-passing for all items by requiring that all pairs of items $i$ and $j$ have the same average cycle time, i.e.,

$$\overline{L}_{it} = \overline{L}_{jt} = \overline{L}_t, \quad \text{for all pairs of items} \quad i, j = 1, \ldots, I,\ i \neq j \tag{7.58}$$

where $\overline{L}_t$ denotes the average cycle time associated with material released in period $t$ over all items $i = 1,\ldots,I$, and then enforcing (7.58).

A third approach to ensuring no-passing solutions is to assume that the production unit selects work for processing from the available workload without systematically prioritizing any item over any other. In this case, the mix of items converted into finished goods inventory in a period should match the mix of the items in the available workload, i.e., for all items $i$ we should have

$$\frac{a_i X_{it}}{\sum_{i=1}^{I} a_i X_{it}} = \frac{\Lambda_{it}}{\sum_{i=1}^{I} \Lambda_{it}} \tag{7.59}$$

Although these conditions also result in non-convex constraints (Carey 1992), they form the basis for the allocated clearing function presented in the next section, which provides a computationally tractable approximate formulation that has proven effective for multi-item problems.

## 7.7 The Allocated Clearing Function (ACF) Model

The difficulties with the NSPCF formulation (7.49)–(7.53) arise because there is no constraint linking the output of each item in a period to the workload of that item available for processing in the period. This results in violation of the no-passing property, where workload of a cheap item is held immobile to allow rapid throughput of an expensive item without the need for high WIP levels of the latter. Clearly additional constraints of some sort are needed to address this situation, and we have discussed several possibilities in the previous section. However, all of these result in non-convex optimization models, which are computationally challenging to solve exactly for a proven global optimum. Hence some kind of approximation will be necessary.

To derive the ACF formulation, we consider the clearing function constraints (7.52) from the NSPCF model:

$$\sum_{i=1}^{I} a_i X_{it} \le f\left(\Lambda_t\right).$$

We wish to develop a set of constraints that relate the output $X_{it}$ of each item $i$ in period $t$ to the workload $\Lambda_{it}$ of that item in that period, while continuing to satisfy (7.52). To this end, we define a new set of variables $Z_{it}$ as the fraction of total system output in period $t$ allocated to item $i$ in that period, i.e.,

$$Z_{it} = \frac{a_i X_{it}}{\sum_{j=1}^{I} a_j X_{jt}} \tag{7.60}$$

The definition of the $Z_{it}$ implies that

$$\sum_{i=1}^{I} Z_{it} = 1 \tag{7.61}$$

The following constraint set is now equivalent to (7.52):

$$\begin{aligned} & a_i X_{it} \le Z_{it} f\left(\Lambda_t\right), \quad \text{forall} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \\ & \sum_{i=1}^{I} Z_{it} = 1, \quad t = 1,\ldots,T \end{aligned} \tag{7.62}$$

since summing the first set of constraints over all items $i$ recovers (7.52). We can now incorporate the no-passing conditions (7.59) suggested by Carey (1992) to obtain the constraints

$$\begin{aligned} & a_i X_{it} \le Z_{it} f\left(\Lambda_t\right), \quad \text{forall} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \\ & \sum_{i=1}^{I} Z_{it} = 1, \quad t = 1,\ldots,T \end{aligned} \tag{7.63}$$

$$\frac{a_i X_{it}}{\sum_{i=1}^{I} a_i X_{it}} = \frac{\Lambda_{it}}{\sum_{i=1}^{I} \Lambda_{it}} = Z_{it}, \quad \text{forall} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.64}$$

The last constraint implies that

$$\Lambda_t = \frac{\Lambda_{it}}{Z_{it}} \tag{7.65}$$

yielding the constraint set

$$\begin{aligned}
&a_i X_{it} \le Z_{it} f\left(\frac{\Lambda_{it}}{Z_{it}}\right), \quad \text{forall} \quad i = 1,\ldots,I \\
&\sum_{i=1}^{I} Z_{it} = 1 \\
&\frac{a_i X_{it}}{\sum_{i=1}^{I} a_i X_{it}} = \frac{\Lambda_{it}}{\sum_{i=1}^{I} \Lambda_{it}} = Z_{it}, \quad \text{forall} \quad i
\end{aligned} \tag{7.66}$$

The first constraint in (7.66) achieves our initial goal of obtaining a set of constraints that link the available workload $\Lambda_{it}$ of each item in the period to the output of that item in the period. However, it looks like we now have some seriously non-convex constraints. The situation is not as bad as it appears at first sight, however. A standard result in convex analysis states that for any convex function *f(x),* its perspective *zf(x/z)* is also convex (Boyd and Vandenberghe (2009): 89). Hence the two constraints in (7.66) define a convex feasible region. However, we have seen in Chap. 6 that the last constraint results in a non-convex feasible region.

To develop an approximate constraint set that may be more tractable than (7.66), we relax the third constraint set, which leaves us with the constraints

$$\begin{aligned}
&a_i X_{it} \le Z_{it} f\left(\frac{\Lambda_{it}}{Z_{it}}\right), \quad \text{forall} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \\
&\sum_{i=1}^{I} Z_{it} = 1, \quad t = 1,\ldots,T
\end{aligned} \tag{7.67}$$

The consequence of this relaxation is that the argument of the clearing function in the first constraints may not be accurate; we need not necessarily have $\Lambda_t = a_i\Lambda_{it}/Z_{it}$. This, in turn, introduces the possibility that the aggregate output constraint (7.52) may be violated if $a_i\Lambda_{it}/Z_{it} > \Lambda_t$ for some items *i*. To see that this is not the case, we need to show that the total output of all items *i* cannot exceed the aggregate output of the system implied by its total workload $\Lambda_t$, i.e.,

$$\sum_{i=1}^{I} Z_{it} f\left(\frac{\Lambda_{it}}{Z_{it}}\right) \le f\left(\Lambda_t\right) \tag{7.68}$$

We can write

$$\sum_{i=1}^{I} Z_{it} f\left(\sum_{i=1}^{I} \frac{\Lambda_{it}}{Z_{it}}\right) \le f\left(\sum_{i=1}^{I} Z_{it}\left[\frac{\Lambda_{it}}{Z_{it}}\right]\right) = f\left(\sum_{i=1}^{I} \Lambda_{it}\right) = f\left(\Lambda_t\right) \tag{7.69}$$

where the first inequality holds by the assumed concavity of *f(.)* and (7.61), the first equality from simple algebra and the second from the definitions of $\Lambda_t$ and $\Lambda_{it}$. Since this assumes only the concavity of the clearing function *f(.),* it holds for any concave clearing function. We can thus write the complete allocated clearing function formulation for a single-stage multi-item problem as follows:

$$\min \sum_{t=1}^{T}\sum_{i=1}^{I}\left[r_i R_{it} + c_i X_{it} + w_i W_{it} + h_i I_{it}\right] \tag{7.70}$$

subject to

$$W_{it} = W_{i,t-1} + R_{it} - X_{it}, \quad \text{for} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.71}$$

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad \text{for} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.72}$$

$$a_i X_{it} \le Z_{it} f\left(\frac{\Lambda_{it}}{Z_{it}}\right), \quad \text{for} \quad i = 1,\ldots,I, \quad T = 1,\ldots,T \tag{7.73}$$

$$\sum_{I=1}^{I} Z_{it} = 1, \quad t = 1,\ldots,T \tag{7.74}$$

$$R_{it}, X_{it}, W_{it}, I_{it}, Z_{it} \ge 0, \quad \text{for} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.75}$$

It is important to clarify the precise nature of the approximation being used here. The approximation arises from the fact that we estimate the total output of the system $f(\Lambda_t)$ by $f\left(\frac{\Lambda_{it}}{Z_{it}}\right)$ in the constraints (7.68). If we retain the no-passing constraints (7.64), the estimate of $f(\Lambda_t)$is exact; however, retaining these constraints results in a non-convex optimization model. By relaxing (7.64), we allow the mix of the output, defined by the ratios $a_i X_{it} / \sum_{i=1} a_i X_{it}$ to deviate from the mix of available workload, determined by the ratios $\Lambda_{it}/\Lambda_t$. Thus the ACF model may decide to process a larger fraction of the workload of one item *i* at the expense of another item *j*. However, there are limits to what is possible, as discussed in the next section. The primary insight is that despite their rather intimidating appearance, constraints (7.73) and (7.74) define a convex feasible region, resulting in a convex feasible region for the overall model (7.70)–(7.75) as long as the clearing function is concave.

### *7.7.1 ACF Model with Piecewise Linearized Clearing Function*

Piecewise linearizing the clearing function as in (7.27), we can approximate the convex constraints (7.73) with the linear constraints

$$a_i X_{it} \le Z_{it}\left(\alpha^q \frac{\Lambda_{it}}{Z_{it}} + \beta^q\right) = \alpha^q \Lambda_{it} + Z_{it}\beta_{it}^q, \\ \text{for} \quad i = 1,\ldots,I, \; q = 1,\ldots,Q, \; t = 1,\ldots,T \tag{7.76}$$

Now the ACF formulation has come into its own: the piecewise linear approximation of the single-variable clearing function has resulted in a set of linear constraints, yielding a linear program. However, this comes at the price of a substantial increase in the size of the model. The nonlinear model (7.70)–(7.75) has 5*IT* decision variables and *3IT* + *T* constraints, of which the *IT* constraints (7.73) are nonlinear. The piecewise linearized formulation requires *IQT* linear constraints to approximate the nonlinear constraints (7.73). As a point of reference, a model using exogenous lead times would require *2IT* decision variables representing releases and finished inventories and *IT* capacity and finished inventory balance constraints. As might be expected, the effort to model congestion more effectively increases the computational effort required to solve the models.

The following example illustrates the operation of the ACF formulation.

**Example 7.5** Consider a problem with $T = 14$ time periods and two products whose data is given in Table 7.7 below:

The $c_i$, $r_i$, and initial WIP and inventory levels for both products have been set to zero for simplicity of exposition. The demand data over the planning horizon is given in Table 7.8, and the data for the linear segments approximating the workload-based clearing function in Table 9.

The solution of the ACF model is summarized in Table 7.10. The reader can verify that there is no slack in the clearing function constraints in any period. The high inventory holding costs require the model to operate with as little finished inventory as possible. In periods 1 through 3 and periods 11 through 14, only one product is produced. The shaded cells for periods 4 through 10 represent periods in which both products are in production. In periods 8 and 10 a higher fraction of the aggregate output, represented by the $Z_{it}$ variables, is allocated to Product 1 than would be implied by the WIP fraction. For example, in period 10, Product 1 makes up 47% of the average WIP and yet is allocated 68% of the output capacity. This illustrates

**Table 7.7** Parameters for Example 7.5

| Product | $a_i$ | $c_i$ | $w_i$ | $h_i$ | $r_i$ | $W_{i0}$ | $I_{i0}$ |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 0 | 6 | 5 | 0 | 0 | 0 |
| 2 | 4 | 0 | 11 | 10 | 0 | 0 | 0 |

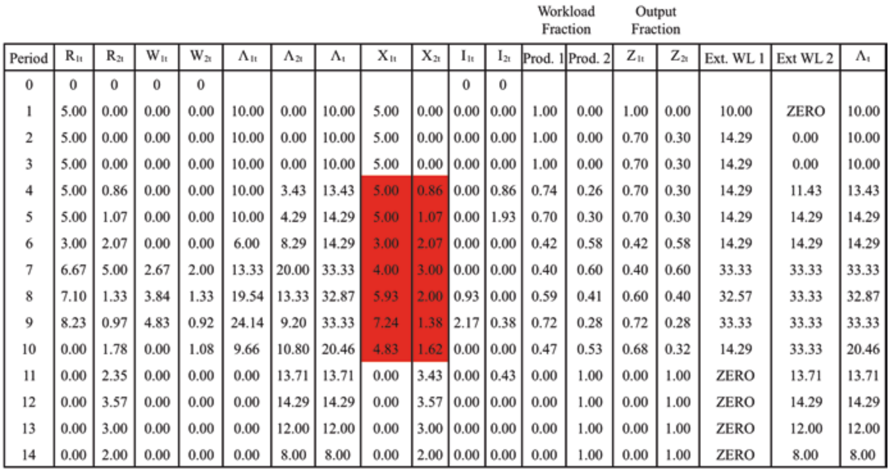
**Table 7.8** Demand Data for Example 7.5

| Period | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product 1 | 5 | 5 | 5 | 5 | 5 | 3 | 4 | 5 | 6 | 7 | 0 | 0 | 0 | 0 | 0 |
| Product 2 | 0 | 0 | 0 | 0 | 0 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 3 | 2 | 0 |

**Table 7.9** Clearing Function Parameters for Example 7.5

| Segment *c* | Slope $\alpha^q$ | Intercept $\beta^q$ |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 0.3 | 10 |
| 3 | 0 | 20 |

**Table 7.10** Solution to ACF Model of Example 7.5



| | | | | | | | | | | | | Workload Fraction | | Output Fraction | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period | $R_{1t}$ | $R_{2t}$ | $W_{1t}$ | $W_{2t}$ | $\Lambda_{1t}$ | $\Lambda_{2t}$ | $\Lambda_t$ | $X_{1t}$ | $X_{2t}$ | $I_{1t}$ | $I_{2t}$ | Prod. 1 | Prod. 2 | $Z_{1t}$ | $Z_{2t}$ | Ext. WL 1 | Ext WL 2 | $\Lambda_t$ |
| 0 | 0 | 0 | 0 | 0 | | | | | | 0 | 0 | | | | | | | |
| 1 | 5.00 | 0.00 | 0.00 | 0.00 | 10.00 | 0.00 | 10.00 | 5.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 | 10.00 | ZERO | 10.00 |
| 2 | 5.00 | 0.00 | 0.00 | 0.00 | 10.00 | 0.00 | 10.00 | 5.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.70 | 0.30 | 14.29 | 0.00 | 10.00 |
| 3 | 5.00 | 0.00 | 0.00 | 0.00 | 10.00 | 0.00 | 10.00 | 5.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.70 | 0.30 | 14.29 | 0.00 | 10.00 |
| 4 | 5.00 | 0.86 | 0.00 | 0.00 | 10.00 | 3.43 | 13.43 | 5.00 | 0.86 | 0.00 | 0.86 | 0.74 | 0.26 | 0.70 | 0.30 | 14.29 | 11.43 | 13.43 |
| 5 | 5.00 | 1.07 | 0.00 | 0.00 | 10.00 | 4.29 | 14.29 | 5.00 | 1.07 | 0.00 | 1.93 | 0.70 | 0.30 | 0.70 | 0.30 | 14.29 | 14.29 | 14.29 |
| 6 | 3.00 | 2.07 | 0.00 | 0.00 | 6.00 | 8.29 | 14.29 | 3.00 | 2.07 | 0.00 | 0.00 | 0.42 | 0.58 | 0.42 | 0.58 | 14.29 | 14.29 | 14.29 |
| 7 | 6.67 | 5.00 | 2.67 | 2.00 | 13.33 | 20.00 | 33.33 | 4.00 | 3.00 | 0.00 | 0.00 | 0.40 | 0.60 | 0.40 | 0.60 | 33.33 | 33.33 | 33.33 |
| 8 | 7.10 | 1.33 | 3.84 | 1.33 | 19.54 | 13.33 | 32.87 | 5.93 | 2.00 | 0.93 | 0.00 | 0.59 | 0.41 | 0.60 | 0.40 | 32.57 | 33.33 | 32.87 |
| 9 | 8.23 | 0.97 | 4.83 | 0.92 | 24.14 | 9.20 | 33.33 | 7.24 | 1.38 | 2.17 | 0.38 | 0.72 | 0.28 | 0.72 | 0.28 | 33.33 | 33.33 | 33.33 |
| 10 | 0.00 | 1.78 | 0.00 | 1.08 | 9.66 | 10.80 | 20.46 | 4.83 | 1.62 | 0.00 | 0.00 | 0.47 | 0.53 | 0.68 | 0.32 | 14.29 | 33.33 | 20.46 |
| 11 | 0.00 | 2.35 | 0.00 | 0.00 | 0.00 | 13.71 | 13.71 | 0.00 | 3.43 | 0.00 | 0.43 | 0.00 | 1.00 | 0.00 | 1.00 | ZERO | 13.71 | 13.71 |
| 12 | 0.00 | 3.57 | 0.00 | 0.00 | 0.00 | 14.29 | 14.29 | 0.00 | 3.57 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 | ZERO | 14.29 | 14.29 |
| 13 | 0.00 | 3.00 | 0.00 | 0.00 | 0.00 | 12.00 | 12.00 | 0.00 | 3.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 | ZERO | 12.00 | 12.00 |
| 14 | 0.00 | 2.00 | 0.00 | 0.00 | 0.00 | 8.00 | 8.00 | 0.00 | 2.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 | ZERO | 8.00 | 8.00 |

the impact of relaxing the FIFO constraints (7.64) on the solution of the ACF model: Product 1 is being prioritized over Product 2 due to its higher WIP holding cost. This is possible because of the structure of the linearized clearing function constraints.

It is of particular interest to compare the output fraction of each product, given by the values of the $Z_{it}$ variables in each period, with the workload fraction. The next to last two columns of the table show the extrapolated workload (Ext. WL) for each product, given by $\Lambda_{it}/Z_{it}$ which can be compared to the actual workload $\Lambda_t$ shown in the last column. In period 10, the extrapolated workload of Product 2 exceeds the actual workload by a considerable margin, while that of Product 1 falls below it. However, the weighted average of the two extrapolated workloads remains equal to the actual workload. In periods where the output fraction matches the workload fraction, extrapolated workload is equal to actual workload for both products. Thus the ACF model is increasing the output fraction of Product 2 by applying a smaller output fraction to larger extrapolated workload.

To see how this is accomplished, note that the output $a_iX_{it}$ of product $i$ in period $t$ can be decomposed into two components: one that is proportional to its workload WIP, given by $\alpha^q\Lambda_{it}$, and a portion of the intercept given by $Z_{it}\beta^q$. Unless $Z_{it} = 0$, the first component will always be produced in proportion to the available average WIP and the slope of the clearing function segment. However, the ACF model may distribute the $\beta^c$ units of output due to the intercept of the linear segment in the manner yielding the best objective function value. The amount of this discretionary output, which can be allocated among products subject only to the WIP balance constraints, increases as the resource is more heavily utilized, leading to higher workload $\Lambda_t$, whichever way the clearing function has been formulated. However, since the $Z_{it}$ variables must sum to 1, the total output in units of time allocated among the different products cannot exceed the disposable output $\beta^q$. Note that if the workload were sufficiently high that only segment $Q$ of the CF, with slope $\alpha^Q = 0$, were binding, the model could allocate output arbitrarily among products subject only to the WIP balance constraints, essentially replicating the N-SPCF model. However, it is easy to see that such a solution can never be optimal, as the same output can be achieved with lower total workload, and hence lower WIP holding costs.

In summary, the ACF formulation avoids the issues encountered with the N-SPCF formulation discussed in Sect. 7.6.1 by relaxing the constraint that output mix must exactly match the workload mix in each period. This allows flexibility in allocating output among the different products, but ensures positive production of all products with positive workload, avoiding the creation of capacity for one product by simply holding static WIP of another. It is by no means a fully satisfactory solution, but it has been extensively tested over more than a decade since its first introduction, and has consistently produced satisfactory, consistent solutions that have in many cases outperformed the fixed lead time models described in Chap. 5. Recent results (Gopalswamy and Uzsoy 2018) have shown that as long as the CF used is concave, the ACF model can be extended to a second-order conic programming formulation which preserves the structure of the dual solution described in the following section, and also significantly reduces the variability of releases across time periods frequently observed with linear programming models.

### 7.7.2 Dual Solution of the ACF Model

Recall from Chap. 5 that any resource with utilization below 1 in any period will have slack in its capacity constraint for that period, resulting in a zero value for the associated dual variable. We now develop and analyze the dual solution for the multistage clearing function model equivalent to that analyzed in Sect. 5.4. The analysis in this section is based on that in Kefeli and Uzsoy (2016), modified slightly for consistency with the discussion in Chap. 5. We shall consider the production system consisting of $K$ resources in series modeled in (5.42)–(5.55) for the case of fixed lead times. No strategic inventory of intermediate products is held between stages inside the production unit; the output of all stages $k = 1, ..., K-1$ except the final one moves directly into the WIP of the next stage $k$+1. Raw material is released into stage 1, and material completing processing at stage $K$ enters finished goods inventory from where it can be withdrawn to meet demand. We represent each stage with its own workload-based clearing function $f_k(\Lambda_{kt})$, where $\Lambda_{kt}$ denotes the planned workload at stage $k$ in period $t$. To implement the ACF model $f_k(\Lambda_{kt})$ will be approximated using the piecewise linearization (7.27) as

$$f_k\left(\Lambda_{kt}\right) = \alpha_k^q \Lambda_{kt} + \beta_k^q, \quad q = 1,\ldots,Q \tag{7.77}$$

To facilitate the sometimes extensive notation, we shall denote the set of all products by $I$ and the index set of all linear segments approximating the clearing function for resource $k$ as $Q$, assuming without loss of generality that all resources are approximated by the same number of linear segments. Using this notation, we write the ACF formulation as follows:

$$\min \sum_{t=1}^{T} \sum_{i \in I} \left( h_{it} I_{it} + r_{it} R_{it} + \sum_{k=1}^{K} \left( p_{it}^{k} X_{it}^{k} + w_{it}^{k} W_{it}^{k} \right) \right) \tag{7.78}$$

subject to

$$W_{it}^{1} = W_{i,t-1}^{1} + R_{it} - X_{it}^{1}, \quad i \in I,\ t = 1,\ldots,T \tag{7.79}$$

$$W_{it}^{k} = W_{i,t-1}^{k} + X_{it}^{k-1} - X_{it}^{k}, \quad i \in I, \quad k = 2,\ldots,K, \quad t = 1,\ldots,T \tag{7.80}$$

$$I_{it} = I_{i,t-1} + X_{it}^{K} - D_{it}, \quad i \in I, \quad t = 1,\ldots,T \tag{7.81}$$

$$a_{i}^{1} X_{it}^{1} \le \alpha_{q}^{1} a_{i}^{1} \left( R_{jt} + W_{i0}^{1} \right) + Z_{it}^{1} \beta_{q}^{1}, \quad i \in I, \quad q \in Q, \quad t = 1,\ldots,T \tag{7.82}$$

$$a_{i}^{k} X_{it}^{k} \le \alpha_{q}^{k} a_{i}^{k} \left( X_{it}^{k-1} + W_{i,t-1}^{k} \right) + Z_{it}^{k} \beta_{q}^{k}, \quad k = 2,\ldots,K, \quad i \in I, \quad q \in Q, \quad t = 1,\ldots,T \tag{7.83}$$

$$\sum_{i \in I} Z_{it}^{k} = 1, \quad k = 1,\ldots,K, \quad t = 1,\ldots,T \tag{7.84}$$

$$X_{it}^{k}, R_{it}, I_{it}, W_{jt}^{k}, Z_{it}^{k} \ge 0, \quad i \in I, \quad k = 1,\ldots,K, \quad t = 1,\ldots,T \tag{7.85}$$

Rewriting this in the cumulative form to eliminate the $I_{it}$ and $W_{it}^{k}$ variables and dropping constants from the objective function, we obtain

$$\min \sum_{t=1}^{T} \sum_{i \in I} \left[ \begin{pmatrix} \left( r_{it} + \sum_{\tau=t}^{T} w_{i\tau}^{1} \right) R_{jt} + \sum_{k=1}^{K-1} \left( p_{it}^{k} + \sum_{\tau=t}^{T} \left( w_{i\tau}^{k+1} - w_{i\tau}^{k} \right) \right) X_{it}^{k} \\ + \left( p_{it}^{K} + \sum_{\tau=t}^{T} \left( h_{i\tau} - w_{i\tau}^{J(i)} \right) \right) X_{it}^{K} \end{pmatrix} \right] \tag{7.86}$$

subject to

$$\sum_{\tau=1}^{t} R_{i\tau} - \sum_{\tau=1}^{t} X_{i\tau}^{1} \ge -W_{i0}^{1}, \quad i \in I, \quad t = 1,\ldots,T \tag{7.87}$$

$$\sum_{\tau=1}^{t} X_{i\tau}^{k-1} - \sum_{\tau=1}^{t} X_{i\tau}^{k} \ge -W_{i0}^{k}, \quad i \in I, \quad k = 2,\ldots,K, \quad t = 1,\ldots,T \tag{7.88}$$

$$\sum_{\tau=1}^{t} X_{it}^{K} - \sum_{\tau=1}^{t} D_{it} \ge -I_{i0}, \quad i \in I, \quad t = 1,\ldots,T \tag{7.89}$$

$$a_{i}^{1} X_{it}^{1} \le a_{i}^{1} \alpha_{q}^{1} \left( R_{it} + W_{i,t-1}^{1} \right) + Z_{it}^{1} \beta_{q}^{1}, \quad q \in Q, \quad i \in I, \quad t = 1,\ldots,T \tag{7.90}$$

$$a_i^k X_{it}^k \le a_i^k \alpha_q^k \left(X_{it}^{k-1} + W_{i,t-1}^k\right) + Z_{it}^k \beta_q^k, \quad k = 2,\ldots,K, \quad q \in Q, \quad i \in I, \quad t = 1,\ldots,T \tag{7.91}$$

$$\sum_{i=1}^{I} Z_{it}^k = 1, \quad k = 1,\ldots,K, \quad t = 1,\ldots,T \tag{7.92}$$

$$R_{it}, X_{it}^k \ge 0, \quad k = 1,\ldots,K, \quad i \in I, \quad t = 1,\ldots,T \tag{7.93}$$

Analysis of the ACF model is simplified by defining two dummy resources 0 and $K + 1$, where resource $K + 1$ represents the arrival of the material in the finished goods inventory. Resource 0, on the other hand, represents the release of the raw material of product $i$ into the line. Thus we define $p_{it}^0 = r_{it}$ for all products $i \in I$, and $w_{it}^0 = 0$. Similarly $w_{it}^{K+1} = h_{it}$ for all $i \in I$ and $t = 1,\ldots,T$, implying that $X_{it}^0 = R_{it}$ in the current notation. The formulation can now be written as follows:

$$\min \sum_{t=1}^{T} \sum_{i \in I} \sum_{k=0}^{K} \left( p_{it}^k + \sum_{\tau=t}^{T} \left( w_{i\tau}^{k+1} - w_{i\tau}^k \right) \right) X_{it}^k \tag{7.94}$$

subject to

$$\sum_{\tau=1}^{t} X_{i\tau}^K \ge \sum_{\tau=1}^{t} D_{i\tau} - I_{i0}, \quad i \in I, \quad t = 1,\ldots,T \qquad \left(\Gamma_{it}^{K+1}\right) \tag{7.95}$$

$$\sum_{\tau=1}^{t} X_{i\tau}^{k-1} - \sum_{\tau=1}^{t} X_{i\tau}^k \ge -W_{i0}^k, \quad i \in I, \quad t = 1,\ldots,T, \quad k = 1,\ldots,K \qquad \left(\Gamma_{it}^k\right) \tag{7.96}$$

$$\begin{gathered} -\alpha_q^k a_i^k \sum_{\tau=1}^{t-1} X_{i\tau}^k - a_i^k X_{it}^k + \alpha_q^k a_i^k \sum_{\tau=1}^{t} X_{i\tau}^{k-1} + Z_{it}^k \beta_q^k \ge -\alpha_q^k a_i^k W_{i0}^k, \qquad \left(\sigma_{itq}^k\right) \\ i \in I, \quad k = 1,\ldots,K+1, \quad t = 1,\ldots,T, \quad q = 1,\ldots,Q \end{gathered} \tag{7.97}$$

$$\sum_{i \in I} Z_{it}^k = 1, \quad k = 1,\ldots,K, \quad i \in I, \quad t = 1,\ldots,T \qquad \left(\lambda_t^k\right) \tag{7.98}$$

$$X_{it}^k \ge 0, \quad i \in I; \quad k = 0,\ldots,K+1, \quad t = 1,\ldots,T \tag{7.99}$$

$$Z_{it}^k \ge 0, \quad i \in I; \quad k = 1,\ldots,K, \quad t = 1,\ldots,T \tag{7.100}$$

with the Greek letters in parentheses denoting the dual variables associated with each constraint. The dual of the formulation (7.94)–(7.100) is given by:

$$\max \sum_{t=1}^{T} \left\{ \sum_{i \in I} \left[ \left( \sum_{\tau=1}^{t} D_{i\tau} - I_{i0} \right) \Gamma_{it}^{K+1} - \sum_{k=1}^{K} W_{i0}^k \Gamma_{it}^k - \sum_{k=1}^{K} \sum_{q} \alpha_q^k a_i^k W_{i0}^k \sigma_{itq}^k \right] + \sum_{k=1}^{K} \lambda_t^k \right\} \tag{7.101}$$

subject to

$$\sum_{\tau=t}^{T}\left(\Gamma_{i\tau}^{k+1}-\Gamma_{i\tau}^{k}\right)-\sum_{q\in Q}a_{i}^{k}\sigma_{itq}^{k}-\sum_{\tau=t+1}^{T}\sum_{q\in Q}\alpha_{q}^{k}a_{i}^{k}\sigma_{i\tau q}^{k}+\sum_{\tau=t}^{T}\sum_{q\in Q}\alpha_{q}^{k+1}a_{i}^{k+1}\sigma_{i\tau q}^{k+1}$$
$$\leq p_{it}^{k}+\sum_{\tau=t}^{T}\left(w_{i\tau}^{k+1}-w_{i\tau}^{k}\right), i\in I,\ t=1,\ldots,T-1,\ k=0,\ldots,K \quad \left(X_{it}^{k}\right) \tag{7.102}$$

$$\Gamma_{iT}^{k+1}-\Gamma_{iT}^{k}-\sum_{q\in Q(k)}a_{i}^{k}\sigma_{iTq}^{k}+\sum_{q\in Q(k+1)}\alpha_{q}^{k+1}a_{i}^{k+1}\sigma_{iTq}^{k+1}\leq p_{iT}^{k}+w_{iT}^{k+1}-w_{iT}^{k},$$
$$i\in I,\quad k\in K\quad \left(X_{iT}^{k}\right) \tag{7.103}$$

$$\lambda_{t}^{k}+\sum_{q\in Q}\beta_{q}^{k}\sigma_{itq}^{k}\leq 0,\quad i\in I,\quad k=1,\ldots,K,\quad t=1,\ldots,T\quad \left(Z_{it}^{k}\right) \tag{7.104}$$

$$\begin{gathered}\Gamma_{it}^{k}\geq 0\quad i\in I,\quad k=1,\ldots,K+1\\ \sigma_{itq}^{k}\geq 0\quad i\in I,\quad k=1,\ldots,K\\ t=1,\ldots,T,\quad q\in Q\end{gathered} \tag{7.105}$$

$$\lambda_{t}^{k}\quad \text{free in sign},\quad t=1,\ldots,T,\quad k=1,\ldots,K \tag{7.106}$$

with the associated primal variable indicated in parentheses next to each dual constraint. In the FLT model, the dual price of capacity is directly accessible as the dual variable $\hat{\sigma}_{kt}$ associated with the capacity constraints. Hence it is meaningful to refer to $\hat{\sigma}_{kt}$ as the dual price of capacity at workcenter *k*. The situation for the ACF model is more complex. The CF does not represent the "capacity" of the system in the sense of an upper limit on output; rather, it represents the relationship between expected workload and expected output at each workcenter *k*. Constraints (7.96) ensure that WIP is nonnegative in all periods, while (7.97) ensure that the output in each period is consistent with the capabilities of the workcenter described by its CF. Thus the dual variables $\Gamma_{it}^{k}$ associated with (7.96) will only be nonzero when these constraints are tight at optimality, i.e., when workcenter *k* has no WIP of product *i* on hand at the end of period *t*. This is achieved when the cumulative output of product *i* at resource *k* in period *t* and the cumulative input of that item to that workcenter differ by $W_{i0}^{k}$, the initial WIP of product *i* at resource *k* at the start of the planning horizon. Thus the $\Gamma_{it}^{k}$ can be interpreted as the cost impact in period *t* of a unit change in the initial WIP level $W_{i0}^{k}$. As implied by the dual objective (7.101), if all initial WIP and FGI values are set to $I_{i0}=W_{i0}^{k}=0$, the $\Gamma_{it}^{k}$ variables have no impact on the optimal solution value except via the artificial workcenter $K+1$ representing the finished inventory. The dual variables $\Gamma_{it}^{K+1}$ represent the maximum amount the firm should be willing to pay for an additional unit of finished inventory of product *i* in period *t*, or, equivalently, the minimum price it should charge an additional unit of demand.

The right hand side of the primal constraints (7.97) that limit output of each product by the CF computes the total output, in units of time, of product *i* for a given workload level. Thus the dual variables $\sigma_{itq}^{k}$ indicate the amount the firm should be willing to pay for an additional time unit of output of product *i* from resource *k* in period *t*. Examining the dual constraint (7.102), recall that $a_i^k$ is the time required to process one additional unit of product *i* on resource *k*, and let $q \in Q$ denote the linear segment of resource *k*'s CF with slope $\alpha_q^k$ that is binding at optimality. Rearranging (7.102) as follows provides some insight:

$$\begin{aligned}
&\sum_{\tau=t}^{T}\left(\Gamma_{i\tau}^{k+1}-\Gamma_{i\tau}^{k}\right)\\
&\underbrace{+\sum_{q\in Q}\alpha_q^k a_i^k \sigma_{itq}^k-\sum_{q\in Q}a_i^k\sigma_{itq}^k}_{\text{net impact in period t}}\\
&\underbrace{+\sum_{\tau=t}^{T}\sum_{q\in Q}\left(\alpha_q^{k+1}a_i^{k+1}\sigma_{i\tau q}^{k+1}-\alpha_q^k a_i^k\sigma_{i\tau q}^k\right)}_{\text{net impact for remainder of planning horizon}}\le p_{it}^k+\sum_{\tau=t}^{T}\left(w_{i\tau}^{k+1}-w_{i\tau}^k\right)
\end{aligned} \tag{7.107}$$

The right hand side indicates that an additional unit of output of product *i* at resource *k* in period *t* will reduce the WIP at this resource by one unit while increasing the WIP at the next resource *k* + 1 along product *i*'s routing; it will also save the incremental production cost $p_{it}^k$. The left hand side represents the total reduction in the objective function value due to this allocation over the remainder of the planning horizon. The impact in the current period *t* is the value of the additional output that can be generated from resource *k* + 1, net of the value of the output from resource *k*, and the impact in the remainder of the planning horizon in a similar fashion. Thus the price paid by the firm for the additional output allocation should not exceed the cost savings from the purchase of the additional allocation.

In an optimal solution to the formulation (7.94)–(7.100), in any period *t* the workcenters can be classified into three groups: congested workcenters where $\sum_i W_{it}^k > 0$, non-congested workcenters where $\sum_i W_{it}^k = 0$ and $\sum_i X_{it}^k > 0$, and idle workcenters where $\sum_i X_{it}^k = \sum_i W_{it}^k = 0$. We shall define congested, non-congested, and idle periods for a workcenter analogously, depending on which of the three states defined above (congested, non-congested, or idle) the workcenter is in during the period in question. Recall we assume all products *i* are processed on all workcenters $k \in K$.

During idle periods, there is no external release of any product into the workcenter *k* and no production at the preceding operation in the product's routing, i.e., $X_{jt}^{k-1} = 0$. Hence there is no production or WIP present for that product at that resource *k*. In non-congested periods, production takes place but no WIP is carried from one period to the other. In this case, the workcenter is operating at sufficiently low utilization that all material arriving from previous operations or external releases is processed in the same period; the segment *q* = 1 with $\alpha_1^q = 1$ and $\beta_1^q = 0$ is tight at optimality. If a resource *k* is congested in some period *t*, on the other hand, the entire workload available to it in that period cannot be processed into output within the period, forcing some to be carried over to the next period as WIP. This means the system is operating at higher utilization and at least one segment of the CF with index $q > 1$ is tight. Our analysis will focus on congested resources since these are where the differences with the FLT model are most clearly visible.

We define a congested interval $\Psi(k)$ for resource $k$ to be a collection of consecutive congested periods starting with a period $s$ and ending with a period $s' > s$, i.e., $\Psi(k) = \{s, s+1,\ldots,s'\}$ such that $\sum_i W_{i,s-1}^k = 0$ and $\sum_i W_{it}^k > 0$ for all $t \in \Psi(k)$ and $\sum_i W_{i,s'+1}^k = 0$.

Before we can apply complementary slackness to (7.101)–(7.106), we need several assumptions regarding the congested interval $\Psi(k)$. The complementary slackness conditions imply that $\Gamma_{it}^k = 0$ for all $t \in \Psi(k)$ since $W_{it}^k > 0$. We also assume that $W_{it}^{k+1} > 0$, so that we have $\Gamma_{it}^{k+1} = 0$, implying that the workcenter performing the next operation in the routing is also congested.

In order to be able to apply the complementary slackness conditions directly without the need to examine a wide range of cases, we will restrict our attention to periods where the system is in regular operation, i.e., $R_{it} > 0$ and $X_{it}^{k-1} > 0$ for some product $i$. These assumptions imply that $X_{it}^k > 0 \forall t \in \Psi(k)$, i.e., if a product is present at a workcenter due to either external releases or output from preceding workcenters, there must be production of that product on the workcenter. Otherwise we can release the work in a later period and save the WIP holding cost. For brevity of exposition, we shall assume that the last period $T \notin \Psi(k)$; in this case constraints (7.103) become active and are subject to a similar analysis.

We now apply complementary slackness to (7.101)–(7.106) during a congested interval $\Psi(k)$. Under the assumptions just stated, (7.102) and hence (7.107) are tight at optimality for all $t \in \Psi(k)$, yielding

$$\sum_{\tau=t}^{T}\left(\Gamma_{i\tau}^{k+1} - \Gamma_{i\tau}^{k}\right) + \sum_{q\in Q}\alpha_q^k a_i^k \sigma_{itq}^k - \sum_{q\in Q} a_i^k \sigma_{itq}^k + \sum_{\tau=t+1}^{T}\sum_{q\in Q}\left(\alpha_q^{k+1} a_i^{k+1} \sigma_{i\tau q}^{k+1} - \alpha_q^k a_i^k \sigma_{i\tau q}^k\right)$$
$$= p_{it}^k + \sum_{\tau=t}^{T}\left(w_{i\tau}^{k+1} - w_{i\tau}^k\right), \quad t \in \Psi(k) \tag{7.108}$$

Equations (7.108) collectively define the dual behavior of the optimal $\sigma_{itq}^k$ in a congested interval. It is immediately evident that, unlike the FLT model, the dual price $\sigma_{itq}^k$ associated with output of any product $i$ at resource $k$ is related to that associated with the preceding workcenter $k - 1$ in its routing. We now rearrange (7.108) in such a fashion that their meaning is clearer by defining the quantity

$$\Phi_{it}^k = a_i^k \sum_{q\in Q}\alpha_q^k \sigma_{itq}^k \quad \forall i \in I, \quad t \in \Psi(k) \tag{7.109}$$

and rewriting (7.108) as follows:

$$\sum_{\tau=s'+1}^{T}\left(\Gamma_{i\tau}^{k+1} - \Gamma_{i\tau}^{k}\right) + \sum_{q\in Q}\alpha_q^k a_i^k \sigma_{itq}^k - \sum_{q\in Q(k)}^{T} a_i^k \sigma_{itq}^k - \sum_{\tau=t}^{T}\Phi_{i\tau}^k + \sum_{\tau=t}^{T}\Phi_{i\tau}^{k+1}$$
$$= p_{it}^k + \sum_{\tau=t}^{T}\left(w_{i\tau}^{k+1} - w_{i\tau}^k\right) \quad i \in I,\ t \in \Psi(k) \tag{7.110}$$

Writing (7.110) for periods $t$ and $t + 1$ and subtracting yields

$$a_i^k\left(\sum_{q\in Q}\sigma_{i,t+1,q}^k-\sum_{q\in Q}\sigma_{itq}^k\right)-\left(\Phi_{i,t+1}^k-\Phi_{i,t}^k\right)=\left(w_i^{k+1}-w_i^k\right)-\left(\Phi_{it}^{k+1}-\Phi_{it}^k\right) \quad (7.111)$$

illustrating the fact that the dual price associated with additional output of product $i$ at workcenter $k$ in period $t$ impacts the dual prices at the downstream workcenter $k+1$ in its routing as suggested by queueing theory (Hopp and Spearman (2008), Chap. 8). Note also that the right hand side of this expression represents the impact of moving a unit of output from resource $k$ to resource $k+1$ in period $t$, while the left hand side reflects its impact across time, from period $t$ to $t+1$.

For the first workcenter $k = 1$ in the common routing (7.110) implies that

$$\sum_{\tau=s'+1}^{T}\Gamma_{i\tau}^1+\sum_{\tau=t}^{T}\Phi_{i\tau}^1=p_{it}^0+\sum_{\tau=t}^{T}w_{it}^1 \quad i\in I,\ t\in\Psi(k) \quad (7.112)$$

Writing (7.112) for periods from $s'$ back to $s$ and solving recursively yields

$$\Phi_{it}^1=\left(p_{it}^0-p_{i,t+1}^0\right)+w_{it}^1 \quad (7.113)$$

Under time-stationary costs $\left(p_{it}^0=r_{it}=r_i, w_{it}^k=w_i^k, h_{it}=h_i=w_{K+1}, p_{it}^k=p_i^k\right)$ this expression simplifies to $\Phi_{it}^1=w_i^1, i\in I, t\in\Psi(k)\setminus\{s'\}$.

The primal constraints (7.98) represent the fact that the expected total output a workcenter $k$ can produce in a given period $t$ with a specified workload $\Lambda_t^k$ is bounded above by the value $f_k\left(\Lambda_t^k\right)$ of the CF. Therefore the dual variables $\lambda_t^k$ associated with (7.98) represent the change in objective function obtained by changing the value of this upper limit, i.e., changing the expected output of the workcenter in a period for a given workload $\Lambda_t^k$. This can be interpreted as the impact on the objective function value of having one additional time unit of output available in period $t$ for allocation among the different products $i$, thus increasing the disposable output $\beta_q^k$ (again in units of time) available for allocation by the $Z_{it}$ variables. Although the dual variables $\lambda_t^k$ are free in sign as a result of constraint (7.98) being defined as an equality, in any optimal solution these variables will only take negative values since an increase in the right hand side of (7.98) cannot yield an increase in the objective function value. Applying complementary slackness to (7.98), we get:

$$\lambda_t^k=-\sum_{c\in Q}\beta_c^k\sigma_{itc}^k,\quad i\in I,\ k\in K,\ t=1,\ldots,T \quad (7.114)$$

Thus at optimality output at each resource $k$ is allocated among products to equalize the marginal value of the capacity allocated to each, in a manner consistent with the marginal value of additional output of each product $i$ in each period $t$, given by the $\sigma_{iqt}^k$. Hence in our numerical experiments below, we shall use this quantity $\lambda_t^k$ as the analog of the dual price of capacity derived for the FLT model.

**Example 7.6** To see the difference in the behavior of the dual prices related to capacity, we compare the dual solution of the two-product single-stage problem in Example 7.5 using model (7.70)–(7.75), with (7.76) replacing (7.73) as in that example, with those from the LP model using only the third, horizontal segment of the CF as a capacity constraint. Figure 7.9 plots the total processing time required to process the demand for both products in the period in which it arises, against the maximum possible output of 20 units per period as a reference. It is apparent that for most of the planning horizon the system has considerable excess capacity, but will have to build anticipatory inventory to meet the demand peaks in periods 6 and 10.

The dual prices for capacity computed by the two models ($\lambda_t$ for the ACF model and $\hat{\sigma}_t$ for the LP model) are shown in Fig. 7.10. The qualitative difference between the dual prices from the two models is immediately apparent. The LP model, which does not consider congestion, only returns positive dual prices for capacity in periods 6, 7, and 9, and these values are an order of magnitude lower than those for the ACF model. The dual prices for the ACF model begin increasing well ahead of the demand peaks representing the congestion caused by increasing releases, and reach substantial values even when the output of the system is below the theoretical maximum of 20 time units implied by the horizontal segment of the clearing function. The dual prices from the ACF are significantly higher than those for the LP model because they consider the additional workload required to raise output in the presence of congestion, which increases rapidly at high levels of output where the slope of the CF is small.

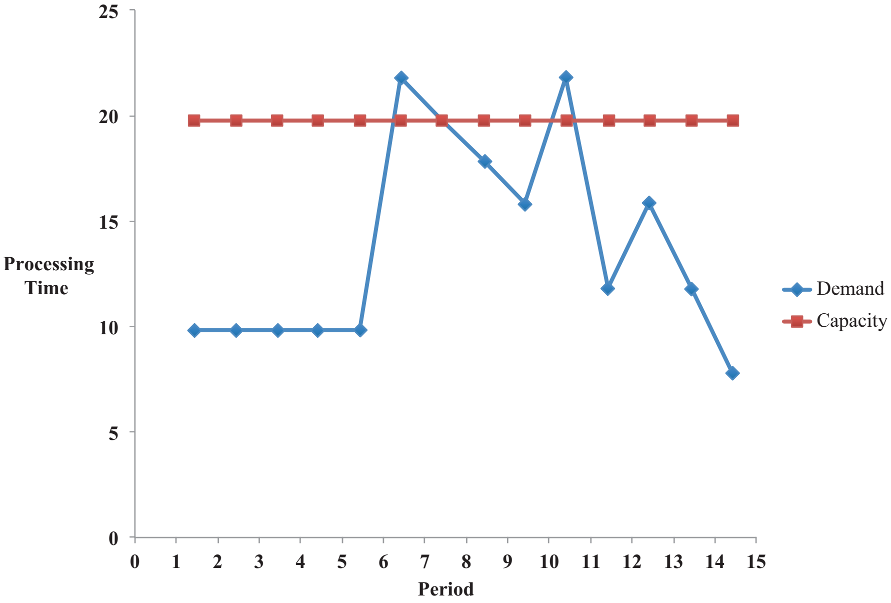


**Fig. 7.9** Demand Levels for Examples 7.5 and 7.6

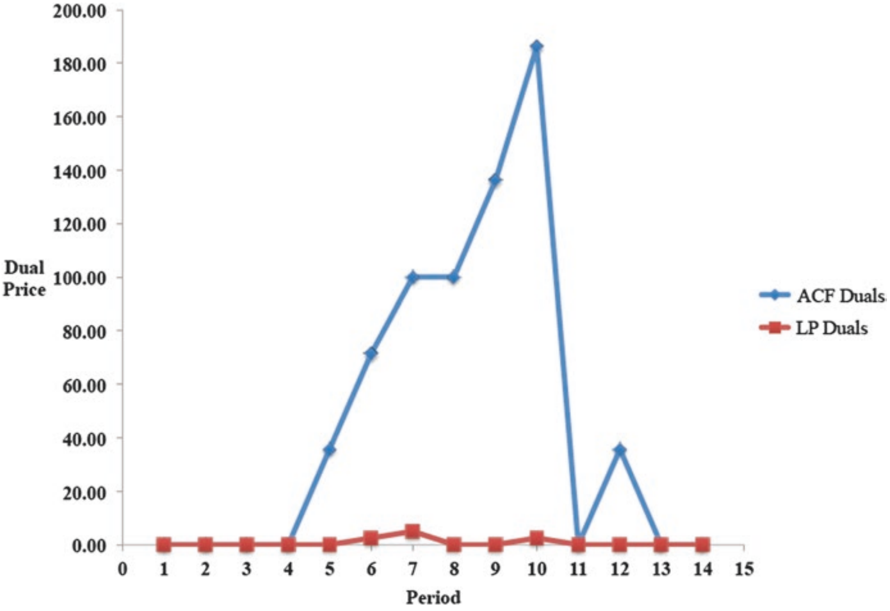


**Fig. 7.10** Comparison of Dual Prices for LP and ACF Models in Examples 7.5 and 7.6

## 7.8 Conclusions

In this chapter, we have introduced the clearing function concept that provides a systematic approach to obtain tractable optimization models for release planning that recognize the nonlinear relation between workload, output, and cycle time discussed in the queueing models of Chap. 2. This chapter has focused on univariate clearing functions that represent the expected output of a production resource in a planning period as a concave non-decreasing function of a single state variable representing the amount of work available to the resource in the period. After reviewing several different types of clearing functions that adopt different state variables, we incorporate them into a convex optimization model for the single-product case. We then extend this model to illustrate the difficulties that arise in the presence of multiple products competing for capacity at a shared resource, and present the allocated clearing function formulation that provides an effective approximate solution to these difficulties. Finally, we show that the use of clearing functions leads to more informative dual prices for capacity than do the LP models of Chap. 5; in particular, the ACF model produces meaningful dual prices when resource utilization is below 1, which the LP models of Chap. 5 cannot.

While the clearing functions and the resulting optimization models described in this chapter have several desirable properties, especially those related to dual prices for resources and the more effective modeling of congestion, they also have some accompanying disadvantages. The need to include decision variables to explicitly model WIP, and the piecewise linearization of the ACF model required to obtain an LP representation of this model, results in substantially larger formulations than those of Chap. 5. While the nonlinear form of the ACF model also yields a convex nonlinear program, due to the preservation of convexity by the perspective transformation, there is as yet little computational work exploring this area. Finally, the basic operation of the ACF model, which uses aggregate workload to estimate aggregate output and then allocates this aggregate output among competing products in a planning period, fails when this type of aggregation is no longer accurate, especially in the presence of significant setup times. In the next two chapters, we explore several more general clearing function models that seek to address these difficulties.

## References

Asmundsson J, Rardin RL, Uzsoy R (2006) Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Trans Semicond Manuf 19(1):95–111

Asmundsson J, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning with resources subject to congestion. Naval Res Logistics 56(2):142–157

Boyd S, Vandenberghe L (2009) Convex optimization. Cambridge University Press, Cambridge

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs, NJ

Carey M (1987) Optimal time-varying flows on congested networks. Oper Res 35(1):58–69

Carey M (1992) Nonconvexity of the dynamic traffic assignment problem. Transport Res B 26B(2):127–133

Carey M, Bowers M (2012) A review of properties of flow–density functions. Transport Rev 32(1):49–73

Carey M, Subrahmanian E (2000a) An approach to modelling time-varying flows on congested networks. Transport Res B 34:157–183

Carey M, Subrahmanian E (2000b) An approach to modelling time-varying flows on congested networks. Transport Res Pt B Methodological 34(6):547

Curry GL, Feldman RM (2000) Manufacturing systems modelling and analysis. Springer, Berlin

Dafermos SC, Sparrow FT (1969) The traffic assignment problem for a general network. J Res Natl Bureau Standard B Math Sci 73B(2):91–118

Franklin RE (1961) The structure of a traffic shock wave. Civil Eng Publ Works Rev 56:1186–1188

Gopalswamy K, Uzsoy R (2018) Conic programming reformulations of production planning problems research report. Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University. Raleigh, NC

Gopalswamy K, Uzsoy R (2019) A data-driven iterative refinement approach for estimating clearing functions from simulation models of production systems. Int J Prod Res 57(19), 6013–6030.

Gopalswamy K, Fathi Y, Uzsoy R (2019) Valid inequalities for concave piecewise linear regression. Oper Res Lett 47:52–58

Graves SC (1986) A tactical planning model for a job shop. Oper Res 34(4):522–533

Graves SC (1988) Safety stocks in manufacturing systems. J Manuf Oper Manag 1:67–101

Graves SC, Kletter DB, Hetzel WB (1998) Dynamic model for requirements planning with application to supply chain optimization. Oper Res 46(3):35–49

Hackman ST, Leachman RC (1989) A general framework for modeling production. Manag Sci 35(4):478–495

Hannah LA, Dunson LA (2013) Multivariate convex regression with adaptive partitioning. J Mach Learn Res 14(1):3261–3294

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Imamoto A, Tang B (2008) Optimal piecewise linear approximation of convex functions. World Congress on Engineering and Computer Science, San Francisco, CA

Johnson LA, Montgomery DC (1974) Operations research in production planning, scheduling and inventory control. Wiley, New York

Kacar NB, Moench L, Uzsoy R (2016) Modelling cycle times in production planning models for wafer fabrication. IEEE Trans Semicond Manuf 29(2):153–167

Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123

Kefeli A, Uzsoy R (2016) Identifying potential bottlenecks in production systems using dual prices from a mathematical programming model. Int J Prod Res 54(7):2000–2018

Leachman RC (2001) Semiconductor production planning. In: Pardalos PM, Resende MGC (eds) Handbook of applied optimization. Oxford University Press, New York, pp 746–762

Merchant DK, Nemhauser GL (1978) A model and an algorithm for the dynamic traffic assignment problems. Transport Sci 12(3):183–199

Missbauer H (1998) Bestandsregelung als Basis für eine Neugestaltung von PPS-Systemen. Physica, Heidelberg

Missbauer H (2002) Aggregate order release planning for time-varying demand. Int J Prod Res 40:688–718

Newell G (1961) A theory of traffic flow in tunnels. In: Herman R (ed) Theory of traffic flow. Elsevier, Amsterdam, pp 193–206

Nyhuis P, Wiendahl HP (2009) Fundamentals of production logistics: theory, tools and applications. Springer, Berlin

Parrish SH (1987) Extensions to a model for tactical planning in a job shop environment. Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA

Peeta S, Ziliaskopoulos AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. Netw Spat Econ 1(3-4):233–265

Pochet Y, Wolsey LA (2006) Production planning by mixed integer programming. Springer Science and Business Media, New York

Srinivasan A, Carey M, Morton TE (1988) Resource pricing and aggregate scheduling in manufacturing systems. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA

Teo C, Bhatnagar R, Graves SC (2011) Setting planned lead times for a make-to-order production system with master schedule smoothing. IIE Trans 43:399–414

Teo C, Bhatnagar R, Graves SC (2012) An application of master schedule smoothing and planned lead time control. Prod Oper Manag 21(2):211–223

Toriello A, Vielma JP (2012) Fitting piecewise linear continuous functions. Eur J Oper Res 219:86–95

Turkseven CH (2005) Computational evaluation of production planning formulations using clearing functions. School of Industrial Engineering, Purdue University, West Lafayette, IN

Van Aerde M, Rakha H (1995) Multivariate calibration of single regime speed–flow–density relationships. 6th Vehicle Navigation and Information Systems (VNIS) Conference, Seattle, WA

Van Ooijen HPG, Bertrand JWM (2003) The effects of a simple arrival rate control policy on throughput and work-in-process in production systems with workload dependent processing rates. Int J Prod Econ 85(1):61–68

# Chapter 8
# Multivariate Clearing Functions

The clearing functions examined in Chap. 7 all assume that the expected output of a production resource in a planning period is a function of a single, aggregate state variable characterizing the amount of work available to the resource during the planning period; hence they were termed univariate clearing functions. As discussed in Chap. 7, several alternative definitions of this aggregate workload have been proposed, including the average WIP level during the planning period, the sum of entering WIP and new releases, or solely the beginning WIP. The use of such aggregate clearing functions in production environments with multiple products created anomalous behavior in the resulting optimization models as seen in Example 7.4. The allocated clearing function formulation develops an approximate formulation that provides effective solutions to this issue and has been validated in extensive computational experiments (Asmundsson et al. 2006, 2009; Kacar et al. 2012, 2013, 2016).

However, the allocated clearing function formulation is based on the assumption that the workload on the production resource resulting from all products in the system competing for its capacity can be aggregated into a single measure of workload without major loss of accuracy. An alternative statement of this assumption is that for a given total workload, however it is measured, the total amount of output, measured in the same units, that the resource can produce in a planning period is independent of the mix of products making up that total.

Univariate clearing functions also assume that the workload information for the current state, however defined, is sufficient to characterize the output of the resource in the current period. While this assumption may be valid for planning periods that are sufficiently long that the queues representing resource behavior can reach steady state and the periods of transient behavior at the beginning of the period due to new release decisions can be neglected, it is clearly questionable in many planning situations. Planning periods are often too short for steady state to be reached, and the release decisions introduced by the planning models at the start of each period are continually creating new workload situations by design. Queueing models suggest that the output of the system in any period can potentially depend on the entire

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_8

history of the arrival and service processes previous to the period, as well as their evolution during the period itself.

In this chapter, we shall examine more complex clearing functions that attempt to address these issues. The obvious first step is to disaggregate the single state variable for each period that forms the basis of the clearing functions in Chap. 7 in different ways. This approach begins by separating the two components of the period workload $\Lambda_t$ into its two components, $R_t$ and $W_{t-1}$, and treating each as a separate state variable. The presence of multiple products makes disaggregation of both WIP and releases by products a natural step. When cycle times exceed the length of the planning period, there may also be benefit to considering the workload in previous periods. For each set of state variables selected, a specific functional form for the clearing function must also be chosen. Many of these functional forms result in nonconvex optimization models, but there is considerable computational evidence that in many cases a standard convex solver yields high-quality solutions.

We shall begin our discussion by using transient queueing models to provide an initial intuition for why additional state variables are needed. We then discuss clearing functions that explicitly attempt to represent the transient behavior of the system without assuming steady state, and then proceed to consider additional state variables related to individual products and previous periods. The discussion of lot-sizing models based on multi-dimensional clearing functions that consider WIP levels, planned output levels, and planned lot sizes as state variables is treated separately in Chap. 9 since lot sizing raises some additional issues.

## 8.1 Limitations of Single-Dimensional Clearing Functions

The functional forms of the single-dimensional clearing functions described in Chap. 7 are almost all derived from steady-state queuing models. Hence they relate the average WIP or workload of the production system in steady state over a planning period to the expected output in this period. Similarly, a clearing function estimated from simulation data reflects the environmental conditions represented in the data set used to fit the clearing function. Any order release planning model using the clearing function thus implicitly assumes that these relationships continue to hold for each period of the planning horizon. However, since both demand and release quantities will vary over time, this assumption is often problematic. The order releases obtained from the clearing function model can exhibit characteristics that systematically deviate from steady state or from the characteristics of the simulation data used for setting the clearing function parameters, invalidating the shape of the clearing function assumed by the order release model.

This issue can be demonstrated by the following simple example. Consider a single production resource that can be modeled as an *M/M/1* queuing system in steady state. Recall from Chap. 2 that the clearing function for this system is given by (2.6). A clearing function based release planning model assumes that this function is valid for each period of the planning horizon. However, only the production orders available to the resource at the start of period 1 are known with certainty since they can be observed directly. If the processing times are known with certainty, the initial WIP level $W_0$, measured in hours of work is thus known. If no releases of new work are expected during period 1, its workload will be $\Lambda_1 = W_0$, and the deterministic clearing function for period 1 will be

$$X_1 = \max\left\{\Lambda_1, C_1\right\} = \max\left\{W_0, C_1\right\} \tag{8.1}$$

unless machine breakdowns occur or work is delayed deliberately, which we shall assume is not the case. This is essentially the best-case clearing function of Hopp and Spearman (2008). Figure 8.1 compares the steady-state clearing function (7.24) derived by Missbauer for an *M/M/1* queue and (8.1). They clearly differ substantially, but a release planning model using a clearing function assumes that they are identical. In this case, the steady-state clearing function substantially and consistently underestimates the expected output of the resource in period 1 for a given workload, for the reasons discussed in (8.8) below.

This example describes an extreme case. We now generalize the underlying reasoning using the queueing-theoretical analysis presented below.

## 8.2 Transient Queueing Analysis of Clearing Functions

We arbitrarily select a particular planning period of an order release model, and consider a production resource modeled as an *M/M/1* queue that can be in transient regime. The number of the period is 1 without loss of generality, that is, the first planning period can have a negative period index. At the start of the period (time $t = 0$) the amount $W_0$ of WIP available to the resource, again measured in units of time, can be observed and hence is known with certainty. The resource is available for production for $\Delta$ time units during the planning period. We shall derive the

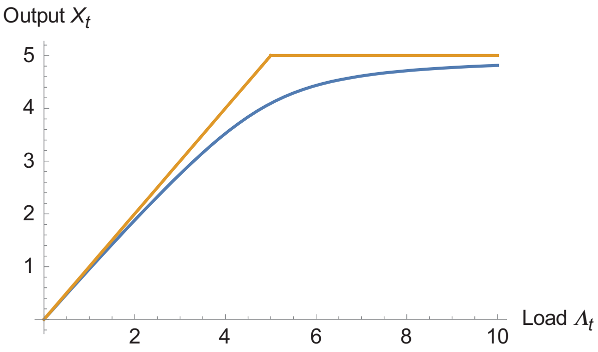


**Fig. 8.1** Clearing function for idealized situation vs. for steady-state *M/M/1* system (Eq. (7.24) with $t_e = 0.2$, $\sigma = t_e$)

functional relationship between expected load $E[\Lambda_1]$ and expected output $E[X_1]$ of the resource for this period, where both quantities are measured in units of time. The analysis follows the approach of Missbauer (2011) where WIP is measured in number of orders; the following analysis for deterministic initial WIP is due to Missbauer (2014).

It is clear that if we have $W_0 \geq \Delta$, the output $X_1 = \Delta$. We now analyze the nontrivial case where $W_0 < \Delta$. In this case, the resource operates continuously from time $t = 0$ until time $t = W_0$. Within this time interval of length $W_0$ time units, work arrives according to a Poisson arrival process with arrival rate $\lambda$, but no arriving work is processed due to the FIFO assumption (highlighting, incidentally, the dependence of the clearing function on the specific dispatching policy used in the production unit). In contrast to the initial WIP $W_0$, which is known with certainty at time $t = 0$, we assume that no information is available about orders that arrive after the start of the period. Defining $p_n(t)$ as the probability of having $n$ orders in the system at time $t$, the probability distribution of the number of orders in the system at time $W_0$, given by the number of orders arriving during the interval $[0, W_0]$, is

$$p_n\left(W_0\right) = \frac{\left(\lambda W_0\right)^n e^{-\lambda W_0}}{n!}, n = 0, 1, \cdots \tag{8.2}$$

Given this probability distribution, the output of the system during the interval $[W_0, \Delta]$ can be derived by calculating the probability of idleness for all $t$ in the interval $[W_0, \Delta]$ (Missbauer 2011). Denoting the output of the system in the time interval $[t_1, t_2]$ within period 1 by the random variable $X_1(t_1, t_2)$, the expected output in period 1 for mean arrival rate $\lambda$ and initial WIP $W_0$ can be written as:

$$E\left[X_1\left(0, \Delta\right) | W_0\right] = \begin{cases} W_0 + E\left[X_1\left(W_0, \Delta\right)\right] & \text{for } W_0 < \Delta \\ \Delta & \text{for } W_0 \geq \Delta \end{cases} \tag{8.3}$$

We must now calculate $E[X_1(W_0, \Delta)]$, the expected output in the interval $[W_0, \Delta]$. After time $t = W_0$, the arrival process continues with rate $\lambda$ until the end of the period. The state probabilities of having $n$ orders in the system at time $t$, $W_0 < t \leq \Delta$, can be calculated from the state probabilities at time $W_0$ given by (8.2) and the conditional state probabilities $p_{rn}(t)$ of having $n$ customers in the system at time $t$ given $r$ customers in the system at time 0. The latter is well-known in queueing theory (Cohen 1969: 82 ff. and 178) and is given by

$$p_{rn}\left(t\right) = \left(1-u\right)\rho^n U\left(1-u\right) + u^{\frac{1}{2}(n-r)} e^{-(1+u)t/t_{\mathrm{e}}} I_{n-r}\left(2\frac{t}{t_{\mathrm{e}}}\sqrt{u}\right) - u^{\frac{1}{2}(n-r)} \int_t^\infty e^{-(1+u)\tau/t_{\mathrm{e}}} \left\{ \begin{array}{l} I_{r+n}\left(2\frac{\tau}{t_{\mathrm{e}}}\sqrt{u}\right) - 2u^{\frac{1}{2}} I_{r+n+1}\left(2\frac{\tau}{t_{\mathrm{e}}}\sqrt{u}\right) + \\ u I_{r+n+2}\left(2\frac{\tau}{t_{\mathrm{e}}}\sqrt{u}\right) \end{array} \right\} \frac{d\tau}{t_{\mathrm{e}}} \tag{8.4}$$

for $t \geq 0$, where $I_j(x)$ denotes the modified Bessel function of the first kind, $t_e$ the mean service time, $u = \lambda t_e$ the utilization and

$$U(t) = \begin{cases} 0, & t < 0 \\ 1/2, & t = 0 \\ 1 & t > 0 \end{cases}$$

The state probabilities at time $t > W_0$ are:

$$p_r(t) = \sum_{n=0}^{\infty} p_n(W_0)\, p_{nr}(t - W_0) \quad \text{for} \quad W_0 < t \leq \Delta \tag{8.5}$$

with $p_n(W_0)$ defined by (8.2). The expected output during the interval $[W_0, \Delta]$, measured in time units, is the expected total time during this interval the server is not idle:

$$E\left[X_1(W_0,\Delta)\right] = \Delta - W_0 - \int_{t=W_0}^{\Delta} p_0(t)\, dt \tag{8.6}$$

where $p_0(t)$ is obtained from (8.5) by setting $r = 0$. Substituting into (8.3) to calculate the output per period for deterministic initial WIP $W_0$, we obtain

$$E\left[X_1(0,\Delta)|W_0\right] = \begin{cases} \Delta - \int\limits_{t=W_0}^{\Delta} \sum\limits_{n=0}^{\infty} p_n(W_0)\, p_{n0}(t - W_0)\, dt & \text{for}\, W_0 < \Delta \\ \Delta & \text{for}\, W_0 \geq \Delta \end{cases} \tag{8.7}$$

Figure 8.2 illustrates the expected output (8.7) as a function of the expected workload in the period for different values of the initial WIP $W_0$. Missbauer (2011) presents the same analysis with initial WIP measured in number of orders. In that case, the differences in the expected output for different initial WIP levels are smaller because for a finite number of orders at the server at $t = 0$ there is always a positive probability of idleness within the period due to the exponentially distributed service times. Figure 8.2 clearly demonstrates that the entire shape of the clearing function changes based on the value of $W_0$, even when the latter is deterministic and not a random variable.

The assumption of deterministic initial WIP is reasonable for the first period in the planning horizon of an order release model. However, the initial WIP $W_{t-1}$ available at the start of all subsequent planning periods $t$ is a random variable. If we interpret the planned value of this random variable calculated in the release planning model as its expectation $E[W_{t-1}]$, the concavity of the clearing function and Jensen's inequality (Billingsley 1995: 80) yield

$$E\left[X_t(W_{t-1},\Delta)\right] \leq E\left[X_t\left(E\left[W_{t-1}\right],\Delta\right)\right] \tag{8.8}$$

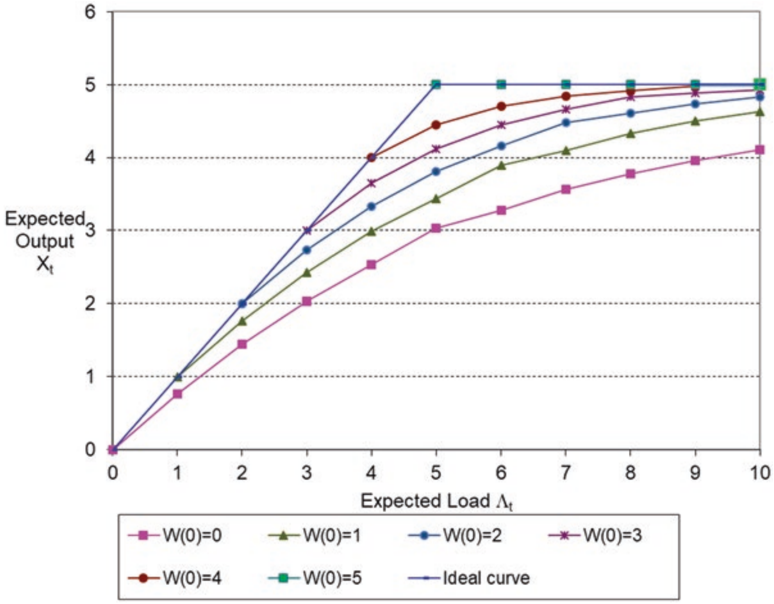


**Fig. 8.2** Clearing functions for different deterministic initial WIP levels measured in hours of work. Period length $\Delta = 5$ time units, expected service time $t_e = 1$

implying that a clearing function treating the planned value of $W_{t-1}$ as a deterministic parameter is likely to overestimate the expected output.

Continuing the analysis for period 1 with WIP measured in units of time, we define $f_{W_0}(w)$ as the probability density function of the initial WIP $W_0$. The expected output for given initial WIP $W_0$ is given by (8.7), and the expected output for stochastic initial WIP can then be obtained by conditioning as:

$$E[X_1] = \int_0^\infty E[X_1(0,\Delta)|w] f_{W_0}(w)\,dw \tag{8.9}$$

where $E[X_1(0, \Delta)|w]$ is given by (8.3).

**Example 8.1** We consider the steady-state distribution of the initial WIP for the *M/M/1* system which, by the PASTA property that Poisson arrivals see time averages (Buzacott and Shanthikumar 1993: 54), is equal to the distribution of the (actual) waiting time of the arriving customers. This distribution is given by

$$f_{W_0}(w) = (1-u)\delta_0(w) + \lambda(1-u)e^{-1/t_e(1-u)\mathrm{w}}, \quad \text{for} \quad w \ge 0 \tag{8.10}$$

where $\delta_0(w)$ denotes the Dirac Delta (unit impulse) function occurring at time $w = 0$ (Papadopoulos et al. 1993: 363).

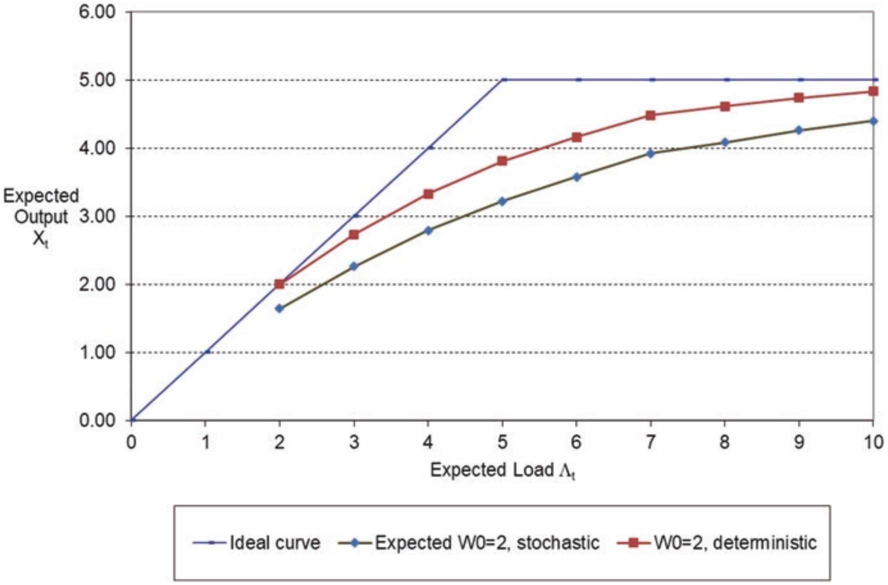


**Fig. 8.3** Clearing functions for expected $W_0 = 2$, deterministic vs. steady-state distribution. $\Delta = 5$ time units, $t_e = 1$

The clearing functions (8.10) for different values of the *expected* initial WIP are plotted in Fig. 8.3 for the same data as in Fig. 8.2. Each point of $E[X_1]$ corresponds to a specified value of the arrival rate $\lambda$ that, added to the expected initial WIP, leads to the expected workload given on the horizontal axis. For computational purposes, the numerical integral in (8.7) is discretized using 10 segments with a finite upper integration limit.

This analysis demonstrates that the expected output for a given *expected* load depends on the composition of the load (initial WIP vs. work released during the period), on the distribution of the initial WIP and also, implicitly, the probability distribution of the arriving work determined by the manner in which the new work is released over the duration of the period.

Armbruster et al. (2012) perform a similar analysis to that presented above for both a constant and time-varying arrival rate (influx, in their terminology) to the resource, analyzing the latter case using discrete-event simulation. They show that the functions depicted in Fig. 8.3 depend on the functional form of the influx over the period, concluding that “the clearing function cannot be just a parametric relationship between input and output” (p. 135).

Missbauer (2009) uses metamodels of the transient behavior of single-stage queueing systems developed from queueing models and simulation, specifically of the transient evolution of WIP over time, to estimate the output of a production resource per period. He shows that this leads to an integer, nonlinear formulation and that modeling errors occur that can lead to counterintuitive behavior. Hence at present the applicability of this approach is unclear.

The results of the analysis so far suggest that we may face a fundamental tradeoff in addressing the problem of formulating clearing functions. If one regards the clearing function as a metamodel of the behavior of the production resource of interest, the primary concern is to develop a model that best predicts the behavior of the resource for a given state at a given point of time. This suggests the use of sophisticated, high-dimensional statistical modeling methods such as Gaussian processes and time series analysis. Such techniques have been used by simulation researchers to develop the operating curves that describe expected cycle time as a function of resource utilization (Yang et al. 2006; Ankenman et al. 2010). Li et al. (2016) use similar techniques to develop a metamodel predicting the output of a production system over time based on a number of state variables, which they then use in place of a discrete-event simulation model in a simulation optimization approach.

While properly formulated and calibrated models of this kind are capable of predicting the output of a production resource or production unit in a planning period quite accurately, they are generally unsuitable for use in a mathematical programming model due to their complex functional forms. As we shall see later in this chapter, even relatively simple multivariate clearing functions lead to non-convex order release models. Hence there appears to be a basic tradeoff between computational tractability of the resulting order release model and the accuracy of the output estimates produced by a clearing function. This issue will surface frequently in the discussion of different functional forms for multivariate clearing functions in this chapter.

Selçuk et al. (2008) formulate a "short-term nonlinear" (STN) clearing function assuming that the WIP at the server is measured in number of orders. Each order that contributes to the load in a certain period is available as soon as it is needed for processing. For exponentially distributed service times, the departure process from the server is a Poisson process with mean rate equal to the service rate $\mu$ until the last order available in this period is completed, after which the server is idle. Under these assumptions, the expected output as a function of the number of available orders (i.e., the workload) can be calculated. Note that idle time at the server due to stochastic interarrival times cannot occur in this model. This simplification allows the univariate clearing function to model the transient state. The saturating shape of this CF is due to the uncertain work content of the orders, which is assumed to be unobservable even for the initial WIP at the time of planning. Asmundsson et al. (2009) use a similar but somewhat more general formulation to prove the concavity of the clearing function in a transient regime.

An approximate model of transient queuing systems that can be integrated into order release models is the stationary backlog carryover (SBC) approach introduced by Stolletz (2008) for *M(t)/M(t)/c(t)* systems and extended to *G(t)/G/1/K* systems by Stolletz and Lagershausen (2013). We shall describe the technique for an *M(t)/M/c* system, characterized by a time-varying Poisson arrival process, exponential service times, and $c$ servers. In the SBC approach, time is divided into short intervals, usually equal in length to the mean service time $t_e$, with arrival rate $\lambda_t$ during each interval $t$. We shall refer to these short intervals as micro-periods to distinguish them from the longer planning periods discussed throughout the volume. The average utilization in period 1 is assumed to be equal to the steady-state utilization of an *M/M/c/c* queue with arrival rate $\lambda_t$, which is given by

$$E\left[u_1\right] = \lambda_1 g\left(\lambda_1\right) t_{\mathrm{e}} \tag{8.11}$$

where $g(\lambda)$ denotes the steady-state fraction of served customers in an Erlang loss (*M/M/c/c*) system with $c$ servers and a mean service time $t_{\mathrm{e}}$ as a function of the arrival rate $\lambda$. Recall that a finite capacity queue or loss system with capacity $c$ can accommodate at most $c$ customers; an arriving customer encountering $c$ customers already in the system will depart without being served. Hence, in this model a fraction $P_1 = 1 - g(\lambda_1)$ of the arriving orders will be blocked from entering the system, giving the expected number of blocked orders in period 1 as

$$b_1 = \lambda_1 P_1 \tag{8.12}$$

For all subsequent micro-periods $t = 2, 3, \ldots$, an artificial arrival rate $\tilde{\lambda}_t$ that accounts for both the (artificial) backlog and new external arrivals is calculated as

$$\tilde{\lambda}_t = b_{t-1} + \lambda_t, \quad t = 2,3,\ldots \tag{8.13}$$

The expected utilization is then calculated from this artificial arrival rate as

$$E\left[u_t\right] = \tilde{\lambda}_t \, g\left(\tilde{\lambda}_t\right) t_{\mathrm{e}}, \quad t = 2,3,\cdots \tag{8.14}$$

Note that if the output estimate is correct the expected artificial backlog $b_{t-1}$ represents the expected WIP, measured in number of orders, in the real system at the end of micro-period $t - 1$ and hence the start of micro-period $t$. Hence (8.14) calculates the expected utilization as a concave saturating function of the workload $\lambda_t$, making SBC a special case of a clearing function model (Missbauer 2007), but with an equality constraint on the output. Missbauer and Stolletz (2016) formulate and test an order release model based on SBC, finding it to be mathematically consistent and solvable by standard NLP solvers. Closely related approximate queueing models for transient systems are suggested by Askin and Hanumantha (2018).

## 8.3 Transient Clearing Functions with Multiple Variables

The problems with the usual one-dimensional clearing functions are obvious: they express the relationship between load and output under long-run average (steady-state) conditions although the actual relation is conditional on the history prior to the planning period under consideration and can be very different from any steady-state condition, especially for the first period of the planning model where initial WIP is largely deterministic. This suggests that extending the one-dimensional CF with additional explanatory variables that reflect the history of the process should improve its ability to estimate output. The queueing-theoretical results derived above indicate that disaggregating the period workload into initial WIP and releases during the period is the most obvious extension. This leads to a two-dimensional CF of the form

$$X_t = f\left(W_{t-1}, R_t\right) \tag{8.15}$$

where $R_t$ denotes the work input in period $t$.

Andersson et al. (1981) propose a linear clearing function of this form without explicit reference to the queueing argument above. In our context, a saturating clearing function of the form (8.15) must be used in order to reflect the congestion phenomena arising from the stochastic nature of arrivals and service times and the finite capacity of the resource. Deriving a piecewise linear approximation of such a function based on empirical or simulated data is difficult without postulating an underlying nonlinear functional form. In contrast to steady-state queueing models, the expressions describing the behavior of transient queueing systems can only be evaluated numerically, rendering the derivation of a tractable expression for a saturating, two-dimensional clearing function difficult. Häussler and Missbauer (2014) propose what appears to be a reasonable functional form with the following properties:

- A fraction $\beta$ of the initial WIP, measured in time units (hours of work), is converted into output during the period, up to the maximum available capacity $C_t$. Simulation models generally assume $\beta = 1$. In general, $\beta$ is a parameter whose value must be estimated from the data.
- For positive releases $R_t > 0$ the output $X_t \leq Min\ \{W_{t-1} + R_t, C_t\}$.
- For given initial WIP $W_{t-1}$ the clearing function is concave and monotonically non-decreasing (saturating) in $R_t$, with

$$\lim_{R_t \to \infty} f\left(W_{t-1}, R_t\right) = C_t \tag{8.16}$$

- For $R_t > 0$ we assume that for fixed $W_{t-1}$ the increase of the clearing function with $R_t$ follows the same functional form as the one-dimensional clearing function derived from a steady-state *M/G/1* model in (7.24). The two-dimensional clearing function is then

$$X_t = \begin{cases} \beta W_{t-1} + \dfrac{\left(C_t - \beta W_{t-1}\right)}{C_t} \cdot \dfrac{1}{2}\left[C_t + k + R_t - \sqrt{C_t^{\,2} + 2C_t k + k^2 - 2C_t R_t + 2kR_t + R_t^2}\right] \\ \qquad\qquad\qquad\qquad \text{if } W_{t-1} < {C_t}/{\beta} \\ C_t \qquad \text{if } W_{t-1} \geq {C_t}/{\beta} \end{cases} \tag{8.17}$$

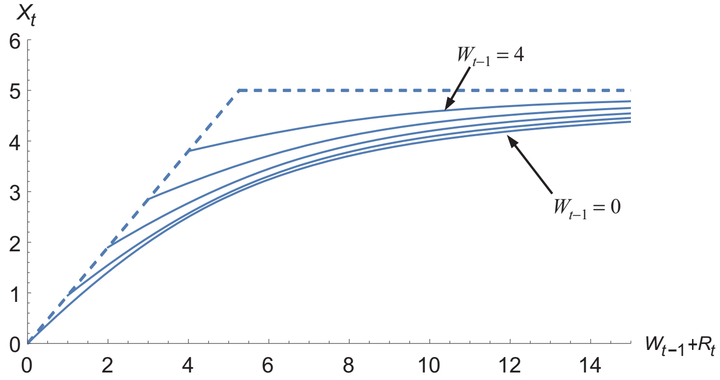


**Fig. 8.4** Two-dimensional CF (Equation (8.17)) for initial WIP levels from 0 (lower function) to 4. Parameters: $C_t = 5$, $k = 1.5$, $\beta = 0.95$. The dashed line is the ideal curve $Min\{W_{t-1} + R_t, C_t\}$

Figure 8.4 depicts (8.17) and shows that, when parameterized appropriately, this CF exhibits a shape very similar to that of the transient CF in Fig. 8.2.

This logic suggests that the fit of the CF can be improved by switching from a 1-dimensional to a 2-dimensional CF. This hypothesis is tested in Häussler and Missbauer (2014) for both simulated and empirical data. The quality of the fit is measured by the adjusted coefficient of determination (adj. $R^2$). Overall, the hypothesis is confirmed for bottleneck resources although the improvements in fit are often smaller than one might expect. To the best of our knowledge no significance test for changes in $R^2$ exists, but the sample size is large (1690 periods for simulation, 350 periods for the empirical data).

Table 8.1 shows the $R^2$ values for three machines operating at the manufacturer of optical storage media described in Chap. 1; the simulation represents a scaled-down version of this manufacturing system. Note the substantial difference between the results for simulated and empirical data caused by the noise in the empirical data, as also observed by Fine and Graves (1989). As expected, the fit for simulation data depends on the period length. In Table 8.1, a period length of five times the average processing time $t_e$ is used. For the period length of the empirical data, which is about 15 times the average processing time, the adj. $R^2$ for simulation is very close to 1.

A saturating, 2-dimensional clearing function based on $W_{t-1}$ and $R_t$ such as (8.17) leads to a convex, nonlinear order release model. Although there is little experience with this structure, it appears to be computationally tractable. Approximating (8.17) by a set of linear functions, in the manner used for 1-dimensional clearing functions, leads to a high number of constraints in the resulting LP model. Successive linear approximation in the optimal region (Hadley 1964) is an alternative as well as using NLP solvers. Determining the best way to solve the resulting models remains a topic for future research.

**Table 8.1** Adjusted $R^2$ for representative bottleneck machines in the manufacturing (Man), printing (Pri), and packing (Pack) department of an optical storage media manufacturer

| Machine | Utilization | $R^2$ 1-dim. CF | $R^2$ 2-dim. CF |
|---|---|---|---|
| *Simulation data* | | | |
| ManBNS | Gateway workcenter | | |
| PriBNS | 95.34% | 0.743 | 0.939 |
| PackBNS | 71.95% | 0.937 | 0.977 |
| *Empirical data* | | | |
| ManBN | 88.71% | 0.664 | 0.687 |
| PriBN | 82.77% | 0.578 | 0.600 |
| PackBN | 80.03% | 0.656 | 0.702 |

1-dim. CF Equation (7.24), 2-dim. CF Equation (8.17) (Häussler and Missbauer 2014)

Decomposing $W_{t-1}$ into its components $W_{t-2}$ and $R_{t-1}$, which leads to a three- or more dimensional clearing function with explanatory variables that reflect the evolution of work input and output over time, has not been considered in the research so far. Kacar and Uzsoy (2014) explore this issue using a product-based clearing function that makes no distinction between the different operations $l$ of each product at each workcenter, but fits a clearing function for each product at each workcenter. Thus the release, WIP, and output variables are defined as

$$R_{ikt} = \sum_{\{l:k=k(il)\}} R_{itl}, W_{ikt} = \sum_{\{l:k=k(il)\}} W_{itl}, X_{ikt} = \sum_{\{l:k=k(il)\}} X_{itl} \tag{8.18}$$

and clearing functions $f_{ik}(.)$ are formulated for each product $i \in I$ and workcenter $k$. Plotting the total output for all products against the total initial WIP $\sum_{i \in I} W_{ikt}$ and total releases $\sum_{i \in I} R_{ikt}$ to a workcenter $k$ subject to machine failures, as illustrated in Fig. 8.5, suggests that there is benefit in disaggregating the workload into its components, releases $R_t$ and entering WIP $W_{t-1}$, as suggested in (8.15). Hence they propose three different product-based clearing functions. Model 1 uses only state information for the current period, given by

$$X_{ikt} = \mu_{ik} + \sum_{i \in I} \beta_{ik} R_{ikt} + \sum_{i \in I} \theta_{ik} W_{ik,t-1} \tag{8.19}$$

where $\mu_{ik}$ denotes the intercept and $\beta_{ik}$ and $\theta_{ik}$ the regression coefficients to be estimated. Model 2 extends Model 1 by considering the releases of product $g$ in the immediately preceding period, yielding

$$X_{ikt} = \mu_{ik} + \sum_{i \in I} \beta_{ik} R_{ikt} + \sum_{i \in I} \theta_{ik} W_{ik,t-1} + \psi_{ik} R_{ik,t-1} \tag{8.20}$$

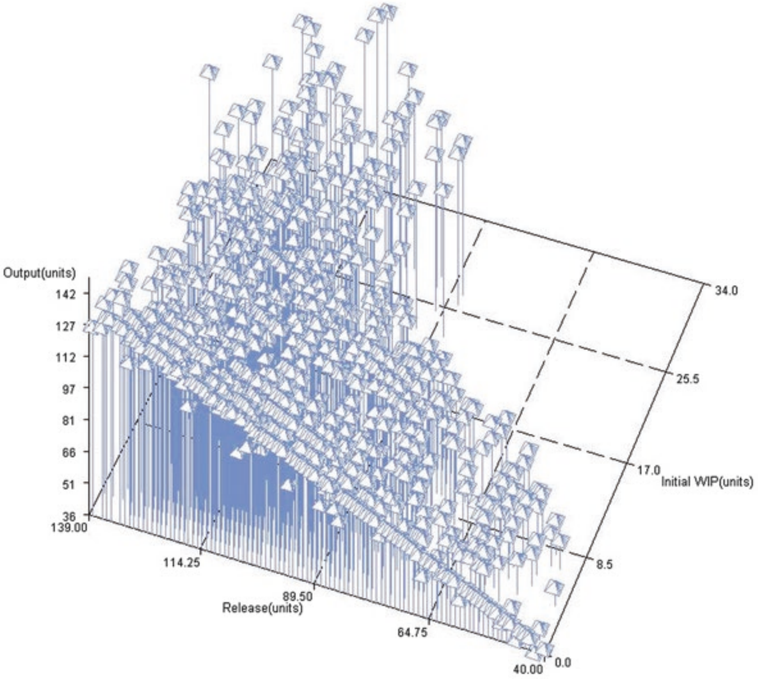


**Fig. 8.5** Total Output at a Machine Subject to Failures as a Function of Releases and Initial WIP (Kacar and Uzsoy 2014)

The final model, Model 3, augments Model 2 by adding the releases for all products in the immediately preceding period, giving

$$X_{ikt} = \mu_{ik} + \sum_{i \in I} \beta_{ik} R_{ikt} + \sum_{i \in I} \theta_{ik} W_{ik,t-1} + \sum_{i \in I} \psi_{ik} R_{ik,t-1} \tag{8.21}$$

The planning model using the product-based clearing functions differs somewhat from the ACF model used with the workload-based clearing function discussed in Sect. 7.2.3. The objective function remains the same, assuming all operations of a given product incur the same WIP holding cost. The material balance equations for finished goods inventory of each product and WIP of each operation of each product are also the same as in the ACF model. However, the constraints governing the output of each product in each period are given by

$$X_{ikt} \le f_{kt}\left(.\right), i \in I, k \in K, t = 1, \ldots, T \tag{8.22}$$

$$\sum_{i \in I} X_{ikt} \le C_k, k \in K, t = 1, \ldots, T \tag{8.23}$$

where $f_{kt}(.)$ is defined by one of (8.19), (8.20), or (8.21). Constraints (8.23) were included because, occasionally, the fitting procedure will return a fit whose intercept exceeds the theoretical capacity of the workcenter.

The comparison of the different product-based clearing functions sheds some light on the issue of whether or not to include state variables related to previous history in the clearing function. Under low utilization the clearing functions (8.19) that consider only variables for the current period are among the best performers, although the difference in expected profit between the models is sometimes small (though statistically significant). At high utilization the model (8.20) that includes the releases of the individual product from the previous period is consistently among the best performers. These results are in general intuitive: at lower utilization levels the production resources will be able to convert the majority of the available workload in a period into output, leaving little WIP at the workcenter at the end of the period. When utilization increases, cycle times will also increase, causing the releases in the previous period to affect output in the current period.

An interesting finding of this work is the analysis of the residuals from the regression models fitted. Figure 8.6 shows the residuals (difference between predicted and realized output) of one of the product-based clearing functions as a function of the observed output of one of the products. Ordinary least-squares regression assumes that the residuals should be independent and normally distributed with homogenous variance and mean zero. It is apparent from Fig. 8.6 that the model illustrated did

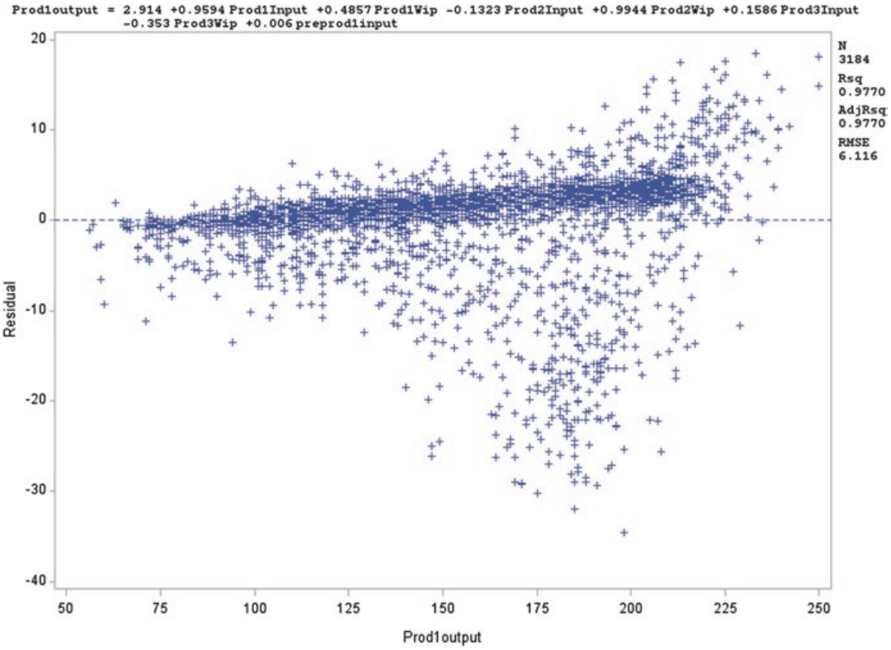


**Fig. 8.6** Residuals for Product-based Clearing Function (8.20) of Unreliable Machine

not satisfy these conditions. While the mean residual is close to zero at low output levels, as output levels increase an upward trend appears. In addition, the variance of the residuals for a given output level, shown in the figure by the vertical dispersion of the points around the horizontal axis, is also increasing, and far from symmetric, suggesting frequent underestimation in the output range 120–220 units. At very high output levels, the problem seems to be one of systematic overestimation. Clearly, the interactions between the state variables are complex, and there is much room for improvement.

In hindsight, the failure to distinguish between the workloads of different operations, i.e., workload of the same product at different stages of processing, confounds the results of these experiments considerably. Comparison of the product-based clearing function and the workload-based clearing functions used in the ACF model shows, unsurprisingly, that in five of the eight experimental conditions the workload-based clearing function outperforms the various product-based clearing functions. The product-based clearing functions perform better for both low utilization short failure cases and low utilization long failures with high demand CV. The reason for this lies in the more granular representation of the production resources in the workload-based clearing function. Recall that in the product-based clearing functions there is no information capturing the flow of material through the different operations of each product routing; the product-based clearing function considers only the total number of lots of each product processed in that period. This creates the opportunity for incorrect behavior such as that illustrated in Chap. 7 for single-variable clearing functions. The product-based clearing function for a given product must produce the different operations in the right combination, but there is nothing in the model to ensure this apart from the finished goods inventory balance equation, which meets demand for each product using output from the last operation on its routing. In contrast, the workload-based clearing function creates a single clearing function for the workcenter whose capacity is shared among the operations, and uses the allocated clearing function formulation to allocate the estimated total output of the workcenter among all operations of all products processed there. The observation that the workload-based clearing function outperforms the best product-based clearing functions in five of the eight experimental conditions, particularly those at high utilization, suggests that the product-based clearing functions as implemented in this study are deficient in multiple aspects. The results of Albey et al. (2014, 2017) discussed below, which examine different aggregations of state variables in single- and multistage production systems, also contribute to this discussion.

Häussler and Missbauer (2014) examine the fit of various 3-dimensional clearing functions to the empirical and simulated data for the manufacturer of optical media described in Chap. 1 and for simulation data specifically designed for this experiment. Since no functional form for saturating 3-dimensional clearing functions is known, they use a linear and a specific cubic function. Although minor improvements in fit were observed in some cases, the results are largely inconclusive. This suggests rapidly diminishing returns to increasing the dimensionality of the clearing functions, but this must be examined in further studies.

The findings presented so far demonstrate that the expected output in a given planning period depends, in principle, on the entire history of the process up to the current period. Neglecting this dependence leads to an inaccurate estimate of the expected output in the planning period, which can be termed an *estimation error*. The inclusion of this inaccurate clearing function in the order release model leads to suboptimal releases over time, which we shall term *optimization error*. In particular, since the clearing function represents the expected output of the system for a given state and time as opposed to its maximum possible output, the effects of temporary periods of high workload (workload peaks) are unlikely to be predicted accurately. A number of experiments have shown that CF-based order release models can lead to fluctuations in releases over time that exceed those in external demand (Missbauer 1998, 2009; Bischoff 2017), which might well be due to this estimation error. Orcun and Uzsoy (2011) observe oscillations of this type in a system where the planning model assumes a fixed lead time but realization follows a clearing function, creating a mismatch between the planning model and the system it is representing.

However, the relationship between the fit of the CF and the quality of the release schedules (at the discrete-event level) is complex (Kacar and Uzsoy (2015)). Preliminary numerical experiments with 2-dimensional CFs show that they can lead to high variations of the releases over time. Figure 8.7 depicts the optimization results for the single-stage, single-product CF model described in Sect. 7.2 (repeated for convenience) that seeks to minimize

$$\min \sum_t wW_t + \sum_t hI_t \tag{8.24}$$

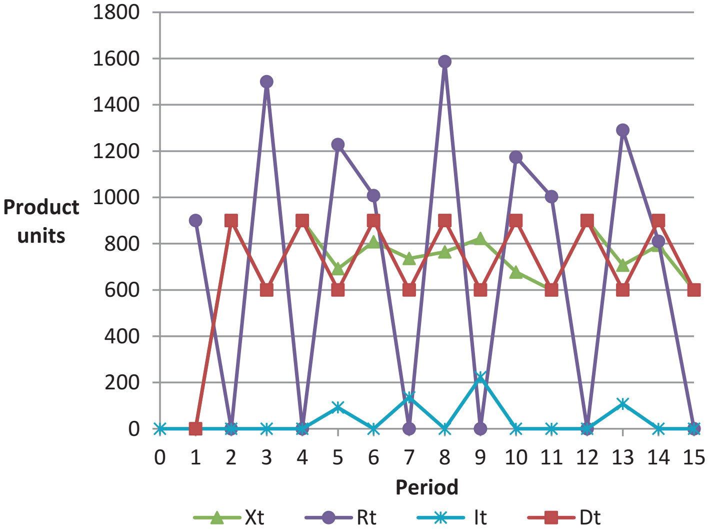


Fig. 8.7 Optimization result for oscillating demand

subject to the standard WIP and finished goods inventory balance equations, the CF (8.17) with $\beta = 1$ and nonnegativity constraints for all variables. The WIP holding cost coefficient $w = 0.5$, and the FGI holding cost $h = 1$ per unit-period. The parameters of the CF are $k = 200$, $C = 950$.

This behavior appears to arise because output in a given period can be increased by either providing initial WIP or by releasing work in the period. Providing initial WIP generates capacity more efficiently since all of it is cleared up to the available capacity. Releasing new work generates less capacity due to the nonlinearity of the CF in $R_t$. For instance, in Fig. 8.7 900 units are released in period 1, held back ($W_1 = 900$) and processed in period 2 ($X_2 = 900$) since this is cheaper than releasing more work in period 2 in order to generate a capacity of 900. This point at which it becomes more economical to release WIP rather than hold it back will change with the utilization due to the specific nonlinear shape of the CF (8.17) that is depicted as a contour plot in Fig. 8.8. This is also related to the findings of Carey (1987) that holding back behavior will arise when releasing WIP in the current period will cause congestion in later periods. This is counterintuitive and indicates that integrating the history of the process into order release models requires modification of the model structure as well. How to do this is largely a topic for future research.

The fact that the expected output in a period depends on the entire history of the process up to that period leads to another important issue: Except for initial WIP of the order release model, the values of the independent variables of the clearing function are point forecasts of a future state of the system, and hence subject to random forecast error, which influences the expected output as seen in Fig. 8.3; different realized values of the initial WIP result in a different curvature for the clearing function. Describing this forecast error for some future period as a function of the deci-

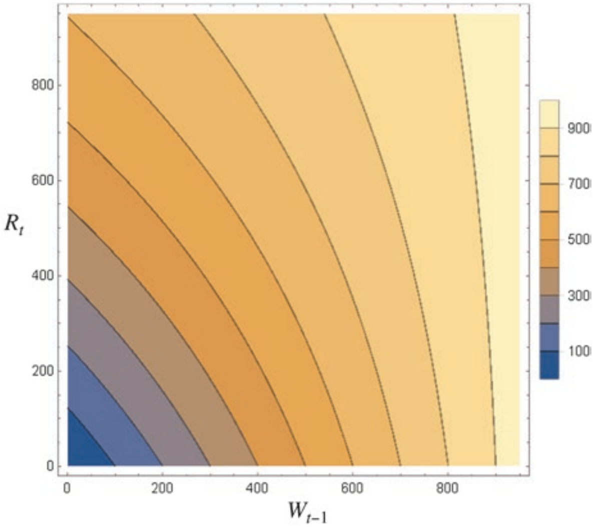


**Fig. 8.8** Contour plot of the CF (8.17) for $\beta = 1$, $C_t = 950$, $k = 200$

sion variables in a release planning model is difficult for two reasons. Firstly, it is not based on hard data that can be measured, but instead reflects the decision maker's state of knowledge at a certain time, e.g., the accuracy with which the WIP level on Thursday morning can be estimated on Monday morning, given specified work releases during the intervening period. Stochastic models of the evolution of the forecast error over time are required. The Martingale Model of Forecast Evolution (Heath and Jackson 1994) is one such approach that has been successfully applied to production planning under uncertain demand (Albey et al. 2015). Secondly, errors in the WIP estimation will increase as the future periods become more remote. Integrating these factors into the order release model results in a complex stochastic programming problem since the evolution of information over time must be considered in a rolling horizon planning framework (Missbauer 2014). While some initial efforts have been made to formulate stochastic optimization models of such problems (Aouam and Uzsoy 2012, 2015; Albey et al. 2015; Lin and Uzsoy 2016), the development of scalable, practically applicable models remains a topic for future research.

## 8.4 Multivariate Multiproduct Clearing Functions

The second principal motivation for the development of multivariate clearing functions is the need to consider the interactions of multiple products competing for capacity at the production resources of interest. This issue has already raised its ugly head in Chap. 7, where we saw that when a univariate clearing function based on a state variable aggregated over different products is used, counterintuitive behavior can result even in the absence of setup times between products. The allocated clearing function formulation addresses this issue to a degree of approximation in the absence of significant setup times. We shall show in this section that when contention between multiple products can lead to significant loss of output, as is the case in the presence of setups, the univariate clearing function fails to predict output at the level of individual products.

We shall first use a simple aggregate queueing model to explore the impact of multiple products on the output of a production resource. We then examine a number of multivariate clearing functions that explicitly address the presence of multiple products, and then consider production units with internal routing flexibility. Under these conditions it is no longer possible to describe the behavior of the production resources using a single clearing function; instead, a system of nonlinear clearing functions that describe the output of each item for fixed WIP and output levels of all other products in the system is required.

### 8.4.1 Motivation

The simple, steady-state queueing analysis used in Chap. 2 can be extended to examine the impact of product mix on system output. In that chapter, we had shown that the average utilization $u$ of a *G/G/1* queue in steady state as a function of the average WIP level $W$ can be approximated as

$$u = \frac{-(W+1)+\sqrt{(W+1)^2+4(\Psi-1)W}}{2(\Psi-1)}, \quad \text{for} \quad \Psi \neq 1 \tag{8.25}$$

where $\Psi = \left(c_\mathrm{a}^2 + c_\mathrm{e}^2\right)/2$, $c_\mathrm{a}^2$ denotes the squared coefficient of variation of the interarrival times and $c_\mathrm{e}^2$ that of the effective service time. Recall that the effective service time is a random variable representing the amount of time a job will spend in service, taking into account both the natural processing time and disruptions such as setups, quality issues, and machine failures (Hopp and Spearman (2008), Chap. 8).

If significant setup times must be incurred when switching between different products, the impact of product mix on the distribution of the effective processing time can be characterized as in Hopp and Spearman (2008). Suppose the natural processing time, the time required to process a job without any detractors such as setups and machine failures, has mean $t_0$ and variance $\sigma_0^2$. Assuming a setup is equally likely to occur after each part being processed, with the average number of parts processed between setups being $N_s$, and denoting the mean and variance of the setup time by $t_s$ and $\sigma_s^2$, respectively, the mean and variance of the effective processing time are given by Hopp and Spearman (2008), Chap. 8:

$$t_\mathrm{e} = t_0 + \frac{t_\mathrm{s}}{N_\mathrm{s}} \tag{8.26}$$

$$\sigma_\mathrm{e}^2 = \sigma_0^2 + \frac{\sigma_\mathrm{s}^2}{N_\mathrm{s}} + \left(\frac{N_\mathrm{s}-1}{N_\mathrm{s}^2}\right) t_\mathrm{s}^2 \tag{8.27}$$

The mix of products processed by the system can potentially affect all terms in the expressions above. The more frequently setups need to be performed, the smaller $N_\mathrm{s}$ will be; in addition, both the mean and variance of the setup time distribution may increase as a more diverse portfolio of products requiring different equipment configurations are processed. In practice, lot sizing policies will affect $N_s$, and continuous improvement programs such as single minute exchange of die (SMED) (Shingo 1986) seek to reduce both $t_\mathrm{s}$ and $\sigma_\mathrm{s}^2$. However, the impact of product mix on utilization, and hence output, through its impact on $c_\mathrm{e}^2$ is evident.

A simple simulation experiment reported by Albey et al. (2014) makes this point quite graphically (no pun intended!). They consider a single-stage production system capable of producing two different parts, whose processing times are lognormally distributed with a mean of $t_0$ = 100 s and a coefficient of variation of 0.13.

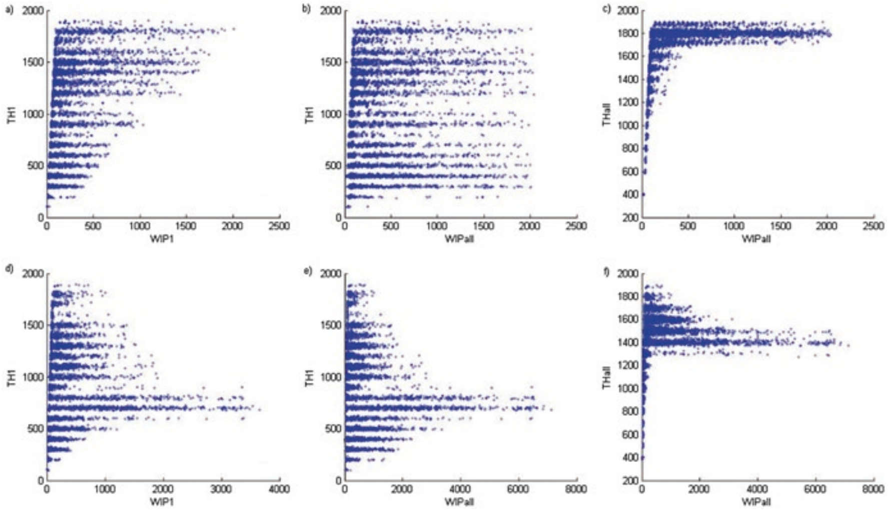


**Fig. 8.9** Impact of Product Mix on System Output

Parts are released into the system one by one following a cyclic pattern based on the heuristic of Askin and Standridge (1993). They consider two different situations: one in which there are no setup times between part types, and one where the setup time follows a triangular distribution with mean $t_s = 0.1t_0$. The demand in each period follows a Poisson distribution, leading to a mean total workload of 1600 s in each period. The total workload for the period is then disaggregated into individual products over 10 different product mixes, where the ratio of the second product to the first ranges from 0 to 5 (i.e., 0, 0.2, 0.25, 0.33, 0.5, 1, 2, 3, 4, and 5). Each product mix is simulated for 1000 different workload realizations, resulting in a total of 10,000 observations of workload and output. The resulting plot of the output of the system in a planning period of 1800 s is shown in Fig. 8.9.

The upper row of graphs represent the performance of the system without setup times. The first two graphs plot the output of Product 1 in the planning period as a function of the WIP of Product 1 and the total WIP in the system, in units of time; the rightmost graph shows total output of both products as a function of the total WIP of both products. The banded appearance of the two leftmost charts is due to the discrete product mix combinations used in the experiment. A specified output of Product 1 can be obtained for various WIP levels of that product (leftmost graph), or of all products (middle graph), depending on the amount of Product 2 in the system. Hence the output of Product 1 is not well described by either its own WIP or the total WIP of both products. The rightmost chart, however, shows that the total system output of both products is well represented by a function of the total WIP.

The lower panel of graphs tells a similar story—representing the output of Product 1 in terms of a WIP measure is inaccurate. However, in the presence of setup times, the output of Product 1 can decrease as its WIP increases, if the amount of Product 2 in the system is also increasing. For a given level of either WIP measure (Product 1's WIP or the total WIP), different output levels of Product 1 can be achieved depending on the amount of Product 2 in the system. The rightmost graph in the lower row differs qualitatively from that above it, showing that in the presence of setup times the aggregate output of the system does not present a monotonically increasing, concave shape.

Motivated by these observations, Albey et al. (2014) examine a number of different multi-dimensional clearing functions (MDCFs) for a single production resource. Their point of departure is the univariate clearing function of Karmarkar (1989), given by

$$f\left(\Lambda_t\right)=\frac{K_1\Lambda_t}{K_2+\Lambda_t},\quad \Lambda_t\geq 0 \tag{8.28}$$

where $\Lambda_t$ denotes the workload available to the resource throughout period $t$ as discussed in Chap. 7. They note that in a multiproduct environment, the output of a given product in a planning period must depend on both the amount of that particular product available to the resource during the period, and the amount of capacity allocated to other products. The allocated clearing function formulation of Chap. 7 addresses this issue by estimating the aggregate output of the resource in units of time as a function of the total workload of all products available to it, and then disaggregating this into estimates of output for individual products. Albey et al. (2014) take a different approach by formulating a MDCF for each product $i$, representing the capabilities of the resource by a system of nonlinear, linked clearing functions that use state variables related to all products in the system in the planning period. They consider two classes of these MDCFs: WIP-based MDCFs (W-MDCFs), where the impact of other products $j \neq i$ in the system is represented by the average WIP level of each product during the planning period; and output-based MDCFs (O-MDCFs), where the impact of the other products is estimated using their planned output. They experiment with several functional forms of each type, represented by the O-MDCF

$$X_i=\frac{\left(C-\sum_{j\neq i}a_jX_j\right)\bar{W}_i}{M_i-b_i\sum_{j\neq i}a_jX_j+\bar{W}_i} \tag{8.29}$$

where $X_i$ denotes the expected output of product $i$ in the planning period, $\bar{W}_i$ the planned time-average WIP level of product $i$ over the period, and $M_i$, $a_i$ and $b_i$ are user-defined parameters to be estimated from data. $C$ denotes the expected capacity of the resource in the period. The general form of the W-MDCFs is

$$X_i=\frac{a_i\bar{W}_i+b_i\sum_{j\neq i}\bar{W}_j}{M_i+\sum_{j}b_j\bar{W}_j} \tag{8.30}$$

Several different versions of each MDCF family, the details of which are given in Albey et al. (2014), were tested in computational experiments. The authors show that the MDCFs are non-convex functions, so that the resulting release planning models can be reduced to quadratically constrained nonlinear programs (Linderoth 2005; Bao et al. 2011), which are known to be strongly NP-hard but can be solved by enumerative methods using solvers such as BARON (Tawarmalani and Sahinidis 2005). Some specific MDCFs belong to the class of bilinearly constrained bilinear problems (Al-Khayyal 1992), whose non-convex nature appears to be less severe than that of the general quadratically constrained nonlinear problem. The nine different MDCFs are fitted using least-squares regression using an extensive set of training data generated from a simulation model of a resource processing four different products in different proportions. They consider three different experimental situations. In the first, there is no loss of capacity in switching from one product to another, and products are processed in FIFO order. In the second case switching from one product to another involves a tool change time, with FIFO dispatching. The final case assumes no tool change time and dispatching in order of Shortest Processing Time (SPT), to examine the impact of shop-floor dispatching policy on the performance of the various MDCFs. In all experiments, the products to be released in a period are sequenced in a cyclic pattern and released all together at the start of the planning period, which will result in a very large number of tool changes in the second experimental configuration. The performance of the MDCFs is measured by implementing them in a release planning model, consisting of balance equations for the WIP and finished goods inventory of each product and the MDCFs for each product in each period, and simulating the performance of the production system under the releases determined by these models. Since obtaining globally optimal solutions to the resulting non-convex optimization models requires very high CPU times, the authors use a convex nonlinear solver to obtain locally optimal solutions.

Under FIFO dispatching without tool changes, all but the most simplistic MDCFs perform comparably with the ACF model and a simpler LP model that ignores congestion. The striking feature of this experiment is the good performance of the simple LP model, which assumes that work released in a planning period will be converted to output within the same period. This may seem to suggest that congestion is not particularly important in this experiment, but this is unlikely, since the average utilization in each period is in excess of 0.90, with considerable variation over time. However, under the given demand conditions the resource must operate close to its full capacity for most of the planning horizon, resulting in similar behavior for all planning models. Interestingly, all planning models underestimate the realized cost of the releases they propose. Detailed results are given in the original paper.

The presence of tool changes between different products changes the situation dramatically, as seen in Fig. 8.10. The performance of the ACF model collapses, which is not surprising since it was not designed to consider capacity losses of this type. Not only does it yield higher costs than other MDCFs, but the planning model severely underestimates the realized cost. The LP model uses a conservative estimate of capacity, based on the worst-case number of tool changes, resulting in poor cost performance but, interestingly, a very accurate prediction of the realized costs

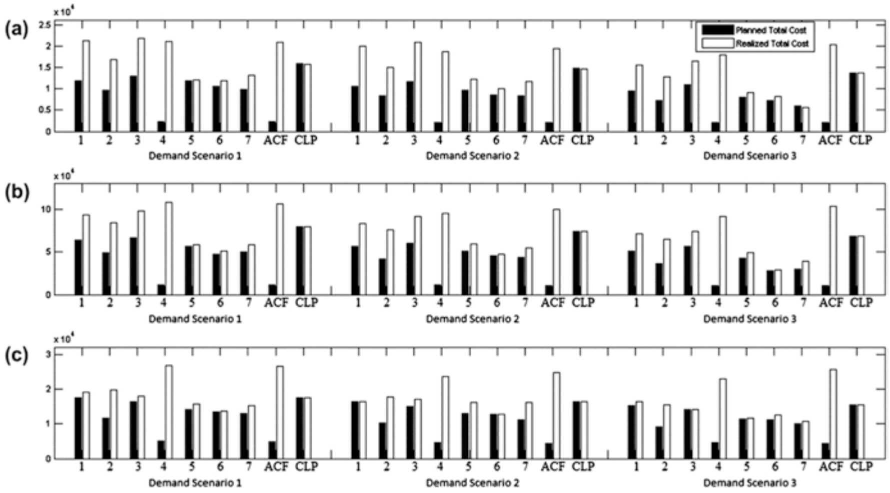


**Fig. 8.10** Performance of various MDCFs relative to ACF and conservative LP Model. MDCFs 1 through 5 are O-based, MDCFs 6 and 7 are WIP-based

of the releases it generates. The W-MDCFs are now the best performers by a considerable margin, suggesting that in the presence of mix-dependent capacity losses, detailed representation of product mix is required. The poorer performance of the O-MDCFs is likely due to the fact that the output of a particular product depends on the amount of WIP of that product available during the period. The final experiments examine the impact of shop-floor dispatching with no tool changes. Similar to the findings of Asmundsson et al. (2006), the performance of the better MDCFs and ACF are generally comparable, suggesting that the use of non-delay dispatching policies in the absence of interference between products does not adversely affect the performance of ACF, while some of the MDCFs perform quite poorly. The poor performance of certain MDCFs is likely due to the release planning model converging to a poor local optimum rather than a global one.

The primary conclusion is that while MDCFs appear to be essential for good release planning in multiproduct systems where the processing time depends on the product mix, such as in the presence of setups, the resulting optimization models are substantially more involved than the linear programs resulting from the ACF model. The non-convex nature of these optimization models ought to come as no surprise to the reader; after all, even the univariate clearing functions discussed in Chap. 7 resulted in non-convex formulations in the presence of multiple products. The good news seems to be that for many functional forms, the non-convex behavior of the MDCFs seems rather benign, allowing locally optimal solutions obtained by conventional convex solvers to provide good performance. The development of efficient solution algorithms for these models, as quadratically constrained quadratic programs or bilinear models, remains an important topic for future work. The functional form of these MDCFs is also quite similar to those derived in the next chapter for lot-sizing problems.

In a subsequent paper, Albey et al. (2017) extend the idea of clearing functions from a single production resource to a production unit consisting of multiple resources, where in addition to requiring processing on several different resources, products also have routing flexibility that allows a given operation to be performed on one of several different machines. The objective of this work is to identify a set of state variables and a functional form for a MDCF that will allow the output of the overall production unit—not individual resources—in a planning period to be estimated to an acceptable degree of accuracy.

The point of departure for this work is the MDCF form (8.30), which was initially developed for a single production resource. In a production unit consisting of multiple resources, this functional form can be implemented at several levels of aggregation. The minimal unit of work is the machine-operation pair, specifying the processing of a particular operation of a specific product on a specific machine. In the presence of routing flexibility, a given operation may be performed on several alternative, non-identical machines. Operation-machine pairs can be summed for a specified operation, a specified machine, and over products. Summation over machines combines all operations processed on a given machine, while summation over a product sums the workload from all operations performed on that product. The reader will note we have met both these aggregations already: the single-variable clearing functions developed in Chap. 7 are based on aggregate workload over all operations processed at a given machine, while the product-based clearing functions of Kacar and Uzsoy (2014) aggregate the workload from all operations of a given product at a particular resource. The authors develop MDCFs for each of these levels of aggregation, and examine their performance in the presence of different levels of utilization and processing flexibility.

The release planning models based on the MDCFs follow the basic structure of other clearing function based models, with balance equations for finished goods inventory of each product and WIP of each of the basic units of aggregation. Thus in the model based on operation-machine pairs using the MDCF form (8.30), WIP balance equations are written for each operation at each machine as in the allocated clearing function model. When using the operation-based MDCF (8.29), WIP balance equations are written for each operation in each period. The P-MDCF requires WIP balance equations for each operation, since the WIP of each operation is weighted to reflect the likelihood of its emerging as finished product in the current period. As was the case for the single-stage systems, the resulting release planning models are non-convex, but are solved to a local optimum using the KNITRO convex nonlinear solver. The performance of the MDCFs is evaluated by the performance of the production unit under the releases developed by the release planning models using them. The univariate clearing function of Srinivasan et al. (1988) is used as a benchmark for comparison, and is implemented in a release planning model using the allocated clearing function formulation, but without piecewise linearization of the clearing functions. This will be referred to as the single-dimensional clearing function (SDCF) model in our discussions. The resulting nonlinear program is convex per the discussion in Chap. 7, and is solved to a global optimum by KNITRO. The production unit considered is a job shop consisting of six machines producing four products with different routings.

The first experiment examines the performance of the MDCFs as a function of utilization with no setups required between different products and no routing flexibility; each operation can be processed on exactly one machine. The findings from this experiment are confirmatory rather than surprising: at low to medium utilization levels, whose average across all machines and periods varies from 0.6 through 0.8, the performance of all MDCFs is fairly similar, with a slight advantage to OM-MDCF based on individual operation-machine pairs. The authors compare the planned and realized costs of the different models and find very close agreement for these utilization levels, indicating that the release planning model is able to accurately predict the consequences of its decisions on the shop floor. The SDCF also exhibits close agreement between planned and realized profit, but as average utilization reaches 0.8 it yields substantially lower profit than the MDCFs, suggesting once again that it systematically underestimates the capabilities of the production unit. This latter finding once again emphasizes the need for MDCFs when multiple products are present.

At higher average utilization levels, ranging from 0.9 through 1.1—the latter representing a major overload of the system—results are qualitatively different. Major differences appear between the different MDCFs. In terms of realized profit, OM-MDCF, the least aggregated of the MDCFs, is consistently the best performer. O-MDCF is the next best, followed by P-MDCF by a wide margin. The highly aggregated P-MDCF fails dramatically at these higher utilization levels, yielding extremely poor realized performance relative to the other MDCFs.

The second experiment in this study introduces routing flexibility by incrementally adding a single alternative machine for each operation of different products: first for the operations of Product 1, then Product 2, and then for all products. However, the choice of which of the alternative machines to use for a given operation is made by the shop-floor dispatching logic and is not available to the planning models. The improvement in performance of all models with the addition of even a limited amount of flexibility for a single product is quite striking, even when it affects only one of the four products. While the single-dimensional clearing function (SDCF) is the worst performer by a wide margin when there is no flexibility, the presence of flexibility for Product 1 alone more than doubles its expected profit. The marked improvement it obtained when flexibility is allowed for Product 1 suggests that most of its problems are due to the clearing functions estimated for the machines used by that product, machines 1, 3, and 5. The realized performances of OM-MDCF and O-MDCF are now very similar, and with the higher levels of flexibility even P-MDCF provides realized profit comparable to O-MDCF. This suggests that the presence of flexibility, in the form of alternative machines for specific operations of a product, allows capacity to be pooled across machines in a manner that makes it easier for the MDCFs, and even the SDCF, to predict.

The final experiment in this study examines the impact of setups between the different products, extending the analysis in Albey et al. (2014). Results for average utilization of 1.0 are shown in Fig. 8.11. Once again, under low average utilization all MDCFs and SDCF lead to quite similar performance, but as in the earlier study of single-stage systems the situation changes markedly at high utilization. As utilization increases, the more aggregated O-MDCF and P-MDCF begin to fall behind the less

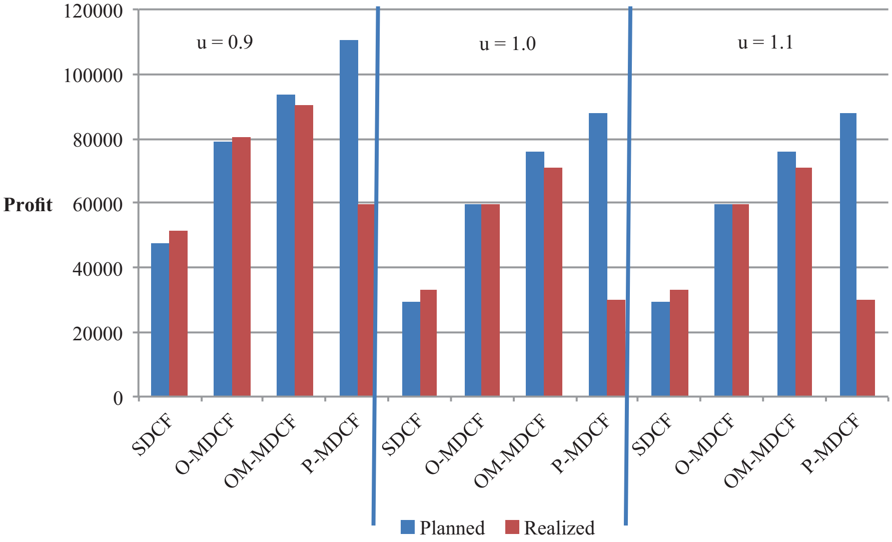


**Fig. 8.11** Performance of MDCFS and SDCF with Setups under High Utilization

aggregated OM-MDCF in relative performance. The realized profit of all models decreases with increasing utilization, due to increasing backorder costs. The close agreement between the planned and realized costs of SDCF combined with its lower profit again suggests that it is underestimating the capabilities of the system, releasing less material which makes it easier for it to realize its planned profit, which remains substantially lower than those of the MDCFs. The very close agreement between the planned and realized profit of OM-MDCF and the very poor agreement for P-MDCF are equally interesting. Since setups are incurred by the processing of specific operations on specific machines, OM-MDCF is able to predict the potential output of each operation-machine pair quite accurately. P-MDCF fails since it does not capture operation-machine level data. O-MDCF occupies an intermediate position.

Once again, the performance of the MDCFs can be explained using intuition from queueing models. The impact of setups in a multiproduct system is to increase the variability of the effective processing time distribution, making it difficult for the more aggregated clearing functions to estimate the output of the system accurately over a wide range of product mixes and operating conditions. The reader will observe the recurring theme: the more different factors contributing to the variability of the effective processing time distribution at any resource, the harder for a clearing function with a few, aggregate state variables to estimate its output accurately. Setups are incurred on the basis of specific operations at specific machines, while machine failures affect all operations at a given machine in essentially the same way. One wishes that the authors had carried their experimental design to its logical conclusion, examining the impact of flexibility on problems with setups, and introducing machine failures. One would conjecture that if machine failures are the dominant source of variability, SDCF ought to perform fairly well, while if the primary source of variability is at the level of operation-machine pairs, OM-MDCF ought to do better. The only situation in which P-MDCF might be expected to perform fairly well would be if all products required very similar processes in terms of both routing and operation processing times.

It is important to note that in this last experiment, as in the setup experiments reported for the single-stage systems, there is no attempt to perform any kind of lot sizing that would make use of setups with maximum efficiency; the cyclic sequence in which products are released approaches a worst-case situation in terms of the number of setups incurred. Clearly there is considerable scope for exploring the impact of different sequencing policies on the shape of the MDCF required to anticipate the behavior of the production unit.

## 8.5 Discussion

Our discussion of MDCFs has ranged over several different possibilities for extending the univariate clearing functions of Chap. 7, which have been the primary focus of research in this area for many years. Univariate clearing functions have been used to estimate the aggregate output of a production resource over all products, usually measured in units of time. For single product systems without sequence-dependent setup times this is, obviously, sufficient, but when multiple items compete for capacity additional logic is required. Chapter 7 discussed the difficulties encountered by univariate clearing functions in the presence of multiple products and presented the allocated clearing function formulation as an approximate, but generally effective, solution in the absence of setup times between products or under predetermined lot sizes. However, both queueing analysis and empirical observation suggest that when the amount of output the system can generate is significantly affected by the mix of the desired output, univariate clearing functions are inadequate.

The development of MDCFs requires additional state variables in the clearing functions, and the development of clearing functions estimating the output of each output item, which is usually a product but can also be defined as an operation-machine pair or multiple operation-machine pairs representing an operation that can be performed on alternative machines. The output capabilities of the system, whether a resource or a larger production unit, are captured by a system of MDCFs that jointly capture the tight interdependence of the output of different items. This approach requires the use of state variables related to each of the items produced, and use of state variables related to earlier planning periods has also been examined.

The common theme across experiments examining different MDCFs is that while univariate clearing functions are capable of estimating the aggregate output of a production resource or unit fairly accurately in the absence of setup times, their ability to estimate the mix of this output, at the level of individual items, is much more limited. Upon reflection, this should be no surprise; even in the absence of setups between products, the presence of multiple products with different service time distributions will increase the variability of the effective service time distribution, making it harder for a single-variable clearing function to produce accurate estimates of output under a wide range of operating conditions and product mixes. In the presence of significant setup times, especially when lot sizes are determined at the scheduling level, the aggregate, univariate clearing functions fail dismally, as is only to be expected.

The use of MDCFs explicitly distinguishing between individual items yields more accurate output estimates, and hence better performance by the planning models that use them, but comes at the cost of significantly larger and more complex release planning models. In particular, the use of MDCFs results in non-convex optimization models that are significantly more difficult to solve than the linear programs of Chap. 5, or the convex nonlinear models and linear programs of Chap. 7. The nature of the non-convexity should be the subject of considerable future study. Anli et al. (2007) observe that non-convex behavior arises either when operating policies at the production units are "flagrantly suboptimal" or the items produced are highly diverse in nature, leading to a highly variable effective processing time distribution. Albey et al. (2014) also find that the objective function values obtained by the BARON global solver were the same as those from the KNITRO convex nonlinear solver in all cases where BARON converged to a solution. This further suggests that the non-convexity of these models is somewhat structured, raising the possibility that more efficient solution procedures may be possible. Further exploration of this issue is clearly an interesting direction for future research and provides a useful application of global optimization methods.

The development of clearing functions that explicitly recognize the transient state of the queues describing the system, without assuming steady-state behavior within the planning periods, has also raised a number of interesting issues. Both empirical evidence and queueing arguments demonstrate quite conclusively that the shape of the clearing function is different in the transient regime from the steady-state environment that is best studied. This issue of transient behavior is compounded by the fact that release planning models treat the planned state of the system in future periods as a deterministic parameter, while in reality these are better treated as (possibly biased) forecasts of random variables. The argument from Jensen's inequality suggests that even assuming unbiased forecasts of the future state variables, treating these estimates as deterministic parameters is likely to result in systematic overestimation of the output, an observation supported by considerable experimental evidence.

There also appears to be a basic tradeoff between the accuracy of the output predictions made by a clearing function and its computational tractability. In general, adding state variables and developing MDCFs for each item produced tend to improve the accuracy of the output predictions, but greatly increase the complexity of the resulting release planning models, as well as the complexity of fitting the MDCFs themselves. Advanced, high-dimensional machine learning techniques such as metamodeling of various kinds and neural networks may be able to produce quite accurate predictions of output, but are not amenable to incorporation in mathematical programming models of the kind this volume has focused on. The use of metamodels to accelerate simulation optimization approaches, by replacing a time-consuming simulation model with a fast running metamodel as in Li et al. (2016), suggests a possible way out of this dilemma, but considerable additional work is needed in this area.

## References

Albey E, Bilge U, Uzsoy R (2014) An exploratory study of disaggregated clearing functions for multiple product single machine production environments. Int J Prod Res 52(18):5301–5322

Albey E, Norouzi A, Kempf KG, Uzsoy R (2015) Demand modeling with forecast evolution: an application to production planning. IEEE Trans Semicond Manuf 28(3):374–384

Albey E, Bilge U, Uzsoy R (2017) Multi-dimensional clearing functions for aggregate capacity modelling in multi-stage production systems. Int J Prod Res 55(14):4164–4179

Al-Khayyal FA (1992) Generalized bilinear programming: part I. Models, applications and linear programming relaxation. Eur J Oper Res 60(3):306–314

Andersson H, Axsater S, Jonsson H (1981) Hierarchical material requirements planning. Int J Prod Res 19(1):45–57

Ankenman BE, Bekki JM, Fowler J, Mackulak GT, Nelson BL, Yang F (2010) Simulation in production planning: an overview with emphasis in recent developments in cycle time estimation. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise: a state of the art handbook, vol 1. Springer, New York, pp 565–592

Anli OM, Caramanis M, Paschalidis IC (2007) Tractable supply chain production planning modeling non-linear lead time and quality of service constraints. J Manuf Syst 26(2):116–134

Aouam T, Uzsoy R (2012) Chance-constraint-based heuristics for production planning in the face of stochastic demand and workload-dependent lead times. In: Armbruster D, Kempf KG (eds) Decision policies for production networks. Springer, London, pp 173–208

Aouam T, Uzsoy R (2015) Zero-order production planning models with stochastic demand and workload-dependent lead times. Int J Prod Res 53(6):1–19

Armbruster D, Fonteijn J, Wienke M (2012) Modeling production planning and transient clearing functions. Logistics Res 5:133–139

Askin RG, Hanumantha GJ (2018) Queueing network models for analysis of nonstationary manufacturing systems. Int J Prod Res 56(1-2):22–42

Askin RG, Standridge CR (1993) Modeling and analysis of manufacturing systems. Wiley, New York

Asmundsson J, Rardin RL, Uzsoy R (2006) Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Trans Semicond Manuf 19(1):95–111

Asmundsson JM, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning models with resources subject to congestion. Naval Res Logistics 56(2):142–157

Bao X, Sahinidis NV, Tawarmalani M (2011) Semidefinite relaxations for quadratically constrained quadratic programming: a review and comparisons. Math Program Ser B 129:129–157

Billingsley B (1995) Probability and measure. Wiley, New York

Bischoff W (2017) Numerical tests of order release models with one- and two-dimensional clearing functions. Department of Information systems, production and logistics management. University of Innsbruck, Innsbruck, Austria

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs, NJ

Carey M (1987) Optimal time-varying flows on congested networks. Oper Res 35(1):58–69

Cohen JW (1969) The single server queue. North-Holland, Amsterdam

Fine CH, Graves SC (1989) A tactical planning model for manufacturing subcomponents of mainframe computers. J Manuf Oper Manag 2:4–34

Hadley G (1964) Nonlinear and dynamic programming. Addison-Wesley, Reading, MA

Häussler S, Missbauer H (2014) Empirical validation of meta-models of work centres in order release planning. Int J Prod Econ 149:102–116

Heath DC, Jackson PL (1994) Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. IIE Trans 26(3):17–30

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Kacar NB, Uzsoy R (2014) A comparison of multiple linear regression approaches for fitting clearing functions to empirical data. Int J Prod Res 52(11):3164–3184

Kacar NB, Uzsoy R (2015) Estimating clearing functions for production resources using simulation optimization. IEEE Trans Autom Sci Eng 12(2):539–552

Kacar NB, Irdem DF, Uzsoy R (2012) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. IEEE Trans Semicond Manuf 25(1):104–117

Kacar NB, Moench L, Uzsoy R (2013) Planning wafer starts using nonlinear clearing functions: a large-scale experiment. IEEE Trans Semicond Manuf 26(4):602–612

Kacar NB, Moench L, Uzsoy R (2016) Modelling cycle times in production planning models for wafer fabrication. IEEE Trans Semicond Manuf 29(2):153–167

Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123

Li M, Yang F, Uzsoy R, Xu J (2016) A metamodel-based monte carlo simulation approach for responsive production planning of manufacturing systems. J Manuf Syst 38:114–133

Lin PC, Uzsoy R (2016) Chance-constrained formulations in rolling horizon production planning: an experimental study. Int J Prod Res 54(13):3927–3942

Linderoth J (2005) A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. Math Program Ser B 103:251–282

Missbauer H (1998) Bestandsregelung als Basis für eine Neugestaltung von PPS-Systemen. Physica, Heidelberg

Missbauer H (2007) Durchlaufzeitmanagement in Dezentralen PPS-Systemen. In: Corsten H, Missbauer H (eds) Produktions- und Logistikmanagement. Festschrift für Günther Zäpfel zum 65. Geburtstag. Verlag Franz Vahlen GmbH, Munich

Missbauer H (2009) Models of the transient behaviour of production units to optimize the aggregate material flow. Int J Prod Econ 118(2):387–397

Missbauer H (2011) Order release planning with clearing functions: a queueing-theoretical analysis of the clearing function concept. Int J Prod Econ 131(1):399–406

Missbauer H (2014) From cost-oriented input-output control to stochastic programming? Some reflections on the future development of order release planning models. In: Gössinger R, Zäpfel G (eds) Management Integrativer Leistungserstellung. Festschrift Für Hans Corsten. Duncker & Humblot GmbH, Berlin, pp 525–544

Missbauer H, Stolletz R (2016) Order release optimization for time-dependent and stochastic manufacturing systems. University of Innsbruck and University of Mannheim. 26 pp

Orcun S, Uzsoy R (2011) The effects of production planning on the dynamic behavior of a simple supply chain: an experimental study. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning in the extended enterprise: a state of the art handbook. Springer, Berlin, pp 43–80

Papadopoulos HT, Heavey C, Browne J (1993) Queueing theory in manufacturing systems analysis and design. Chapman & Hall, London [u.a.]

Selçuk B, Fransoo JC, De Kok AG (2008) Work-in-process clearing in supply chain operations planning. IIE Trans 40(3):206–220

Shingo S (1986) A revolution in manufacturing: the SMED system. Productivity Press, Cambridge

Srinivasan A, Carey M, Morton TE (1988) Resource pricing and aggregate scheduling in manufacturing systems. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA

Stolletz R (2008) Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: the stationary backlog-carryover approach. Eur J Oper Res 190(2):478–493

Stolletz R, Lagershausen S (2013) Time-dependent performance evaluation for loss-waiting queues with arbitrary distributions. Int J Prod Res 51(5):1366–1378

Tawarmalani M, Sahinidis NV (2005) A polyhedral branch and cut approach to global optimization. Math Program 103(2):225–249

Yang F, Ankenman B, Nelson BL (2006) Efficient generation of cycle time-throughput curves through simulation and metamodeling. Naval Res Logistics 54(1):78–93

# Chapter 9
# Lot-Sizing Models Using Multi-dimensional Clearing Functions

The order release models described in this volume rely heavily on the functional relationship between the expected output of a production resource and its expected workload which, as discussed in Chap. 2 for the case of steady-state queues, is related to the expected cycle time by Little's Law. This relationship is significantly affected by various decision rules used within the PPC system, such as scheduling policies on the shop floor. *Lot sizing*, the decision as to how much of a product to produce each time a machine is set up for the product, is of particular importance in this respect. For a given production quantity, determined by the master production schedule, the lot sizes influence capacity utilization (via the amount of setup time required on the resource in a planning period), the mean and variance of the interarrival times (via the number and size of production lots), and the mean and variance of the service times (via the lot sizes). Lot-sizing models were among the earliest mathematical formulations of production planning problems (Harris 1915). The extensive literature on deterministic lot-sizing problems (Drexl and Kimms 1997; Brahimi et al. 2006; Pochet and Wolsey 2006; Quadt and Kuhn 2008) has generally focused on the tradeoff between fixed setup or ordering costs and inventory holding costs without considering the effects of congestion. The relationship between the lot size $Q$ and average cycle time has been explored from several angles, including simultaneous lot sizing and scheduling (Drexl and Kimms 1997) and lot streaming (Missbauer 2002; Jen Huei and Huan Neng 2005; Cheng et al. 2013). Following the discussion in Chap. 2, we begin this section with insights from simple queueing models, and then show how these can be used to develop a system of multivariate clearing functions to address a dynamic lot-sizing problem.

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_9

## 9.1 Impact of Lot Sizes on the Performance of Production Resources

As the extensive literature on deterministic economic order quantities would suggest, lot-sizing decisions have significant impact on the behavior of production systems even in completely deterministic environments. Karmarkar (1989) proposes an example by considering a synchronous production line with $N$ stations producing a single item in batches of a fixed size of $Q$ units. Batches are transferred to the next station at the completion of processing, and a setup time of $S$ time units is required for each batch. The production rate at each station is assumed to be $P$ units/time unit. Thus each batch has a cycle time of $(S + Q/P)$ time units at each station, and a total cycle time in the line of $T = N(S + Q/P)$ time units. Since under synchronous operation there will be no queueing, and one batch will complete its processing and leave the system every $T/N$ time units, the average output rate of the line will be

$$X = \frac{Q}{\left(S + \frac{Q}{P}\right)} = \frac{PQ}{PS + Q} \tag{9.1}$$

which is a saturating, non-decreasing function of the lot size. The non-decreasing, saturating nature of the function arises from the fact that as the lot size $Q$ increases, the setup time per part, which constitutes a loss of production capacity, is reduced, eventually reaching 0 as $Q \to \infty$.

Karmarkar (1987) and Zipkin (1986) were among the first to study the relationship between lot size and mean flow time in an *M*/*M*/*1* queueing system operating at a fixed output rate. Karmarkar (1987) derived this relationship for a single-server queue producing a single item. We follow this derivation for an *M*/*G*/*1* system using the following notation:

$D$: total demand per period (in product units)
$p$: processing time per unit
$S$: setup time per lot
$\lambda$: arrival rate of the lots at the server
$t_e$, $\sigma$, $c_e$: mean, standard deviation, and coefficient of variation, respectively, of the service times of the lots, given by the sum of setup and processing times
$Q$: lot size, assumed to be identical for all lots

For brevity of exposition, we will assume the coefficient of variation $c_e$ of the service times is independent of the lot size $Q$. Then the expected arrival rate of lots at the machine is given by $\lambda = D/Q$ and the expected service time of a lot $t_e = S + pQ$, yielding a utilization of $u = \lambda t_e = D(p + S/Q)$. Assuming a Poisson arrival process, the Pollaczek–Khintchine formula (Medhi 1991) gives the mean queue (waiting) time of a lot as

$$T_Q = \frac{u^2 + \lambda^2\sigma^2}{2\lambda\left(1-u\right)} = \frac{D\left(c_{\mathrm{e}}^2+1\right)\left(pQ+S\right)^2}{2\left[Q\left(1-pD\right)-DS\right]} \tag{9.2}$$

and the mean cycle time as

$$T = T_Q + S + pQ \tag{9.3}$$

Both the mean queue time (9.2) and the mean cycle time (9.3) are convex functions of the common lot size $Q$. When different lot sizes $Q_j$ for multiple products $j$ are used, the mean waiting time remains a convex function of the lot sizes, but the mean cycle time is non-convex (Karmarkar et al. 1992).

The single-product cycle time (9.2) illustrates an intuitive phenomenon termed the "Process Batching Law" by Hopp and Spearman (2008): the presence of positive setup times imposes a lower limit on the lot sizes, and as the lot size approaches this limit from above the utilization $u \to 1$, and hence the average cycle time $T \to \infty$. If the lot size becomes large, i.e., $Q \to \infty$, the impact of the setup time vanishes and $T$ increases asymptotically proportionally with the lot size. As a consequence of this structure, Karmarkar (1987) shows that $T$ is minimized for a well-defined lot size.

These insights have been refined and extended in subsequent work that is beyond the scope of this volume (Wijngaard 1989; Benjaafar 1996; Missbauer 2002; Jutz 2017). Extensions include considering multiple products, more complex material flow structures, and its integration with the inventory control system that determines the arrival process of the lots (Zipkin 1986; Vaughan 2006). The modeling approach can also be extended to multistage systems with stage-specific lot sizes (Missbauer and Jutz 2018).

In order to develop clearing function models, this model must be reformulated to express the expected output as a function of lot sizes and expected WIP. By the PASTA (Poisson Arrivals See Time Averages) property of the arrival process (Buzacott and Shanthikumar 1993, p. 54), the average *actual* waiting time of the customers (lots) $T_Q$ is identical to the average *virtual* waiting time at time $t$, defined as the waiting time that would be seen by a customer arriving at time $t$. For a single-server system, the average virtual waiting time is identical to the average WIP at the server, measured in hours of work (average remaining work). Using (9.2) the expected output $pD$, excluding time spent in setups and expressed in hours of work, can be written as:

$$X = pD = \frac{2pQT_Q}{\left(pQ+S\right)\left[2T_Q+\left(c_{\mathrm{e}}^2+1\right)\left(pQ+S\right)\right]} \tag{9.4}$$

Equation (9.4) implies that higher service time variability reduces the output for a given average WIP. The impact of the lot size $Q$ on the relationship between average WIP and output is shown in Fig. 9.1 for different lot sizes.

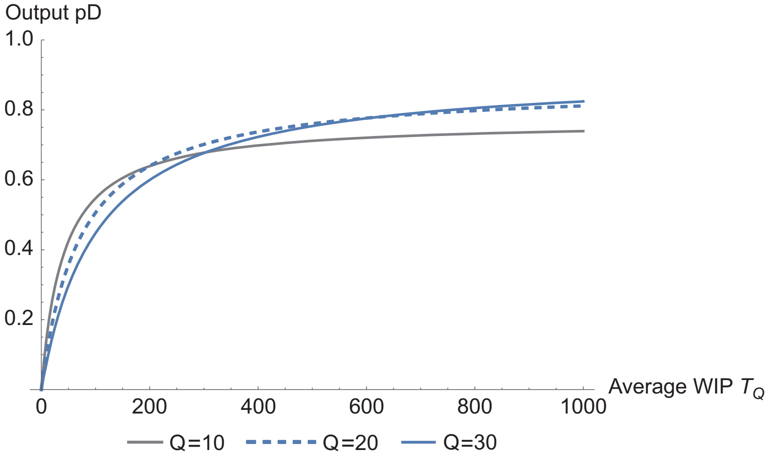


**Fig. 9.1** Expected output as a function of expected WIP for different values of the lot size $Q$ ($p = 5, S = 15, c_e = 0.5$)

This modeling approach can be applied in two ways. The relationship between lot sizes and average flow time can be used to derive standard lot sizes that yield a good compromise between the potentially conflicting goals of reducing cycle times on the one hand and minimizing setup and cycle inventory holding costs on the other (Missbauer 2002). The actual lot sizes implemented on the shop floor can be determined by modifying these standard lot sizes based on short-term demand information, leading to a hierarchical lot-sizing system (Söhner and Schneeweiss 1995). The benefit of modifying the standard lot sizes has been questioned in the literature (Wijngaard 1989). Within this decision structure the lots to release are determined outside the release model and consume the release quantities $R_{jt}$ calculated by the release model as described in Chaps. 5 through 8.

An alternative approach is to determine lot sizes and order releases simultaneously using a release model with a multi-dimensional clearing function that includes some measure of workload and the lot sizes as state variables determining the expected output in the spirit of (9.4). We discuss a model of this type in the next section.

## 9.2 A MDCF Model for Lot Sizing

In this section, we present a single-stage multi-item dynamic lot-sizing model developed by Kang et al. (2014) where the production resource is modeled as an *M*/*G*/*1* queue. The behavior of the system is modeled by a set of multi-dimensional clearing functions (MDCFs) derived by steady-state queueing analysis, instead of the empirically estimated MDCFs described in the previous chapter.

We consider a single production resource processing $N$ different products $i = 1,\ldots, N$ with deterministic processing time $p_i$ and sequence-independent setup time $s_i$ that is incurred whenever a unit of product $i$ is processed after completion of a different product. The planning horizon is divided into $T$ discrete time periods of uniform length, and all processing and setup times are expressed in units of this planning period length. Lots of product $i$ arrive at the resource following a Poisson process with rate $\lambda_i$. Due to the random arrival process, the service time is a random variable. In order to address the lot-sizing problem, the MDCFs describing the output of the resource must reflect the lot sizes. This is accomplished by assuming that the planning periods are sufficiently long that the system is in steady state, and following the analysis of Karmarkar (1987) and Karmarkar et al. (1992). Since we derive the MDCFs for a generic planning period, the period index is dropped in the following analysis.

The deterministic processing time of a lot of $Q_i$ units of product $i$ is given by:

$$P_i = s_i + p_i Q_i \tag{9.5}$$

Since lots of product $i$ arrive following a Poisson process with rate $\lambda_i$, the probability that a randomly selected batch is of product $i$ is given by $\lambda_i/\lambda$, where

$$\lambda = \sum_{i=1}^{N} \lambda_i \tag{9.6}$$

Thus the mean and variance of the random variable $P$ denoting the processing time of lots at the resource are given by:

$$E[P] = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} P_i \quad \text{and} \quad E\left[P^2\right] = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} P_i^2 \tag{9.7}$$

It is a standard result in queueing theory (Buzacott and Shanthikumar 1993, p. 62) that the expected waiting time for the *M/G/1* queue is given by:

$$T_Q = \frac{\lambda E\left[P^2\right]}{2(1-u)} = \frac{\lambda \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} P_i^2}{2(1-u)} = \frac{\sum_{i=1}^{N} \lambda_i P_i^2}{2(1-u)} \tag{9.8}$$

where $u$ denotes the average utilization as in previous chapters. The expected cycle time of product $i$ is then given by $\tau_i = P_i + T_Q$. Little's Law then yields

$$\lambda_i = \frac{\left(\dfrac{\bar{W}_i}{Q_i}\right)}{T_Q + P_i} \tag{9.9}$$

where $\overline{W}_i$ denotes the time-average WIP of product $i$ over the duration of the planning period. Since we assume the system is in steady state, the number $Y_i$ of lots of product $i$ produced during the period can be substituted for $\lambda_i$, yielding

$$Y_i = \frac{\left(\frac{\overline{W}_i}{Q_i}\right)}{\tau_i} = \frac{\left(\frac{\overline{W}_i}{Q_i}\right)}{P_i + T_Q} = \frac{\left(\frac{\overline{W}_i}{Q_i}\right)}{P_i + \frac{\sum_{i=1}^{N} Y_i P_i^2}{2(1-u)}} = \frac{\left(\frac{\overline{W}_i}{Q_i}\right)}{P_i + \frac{\sum_{i=1}^{N} Y_i P_i^2}{2\left(1-\sum_{i=1}^{N} Y_i P_i\right)}} \tag{9.10}$$

Noting that all processing times are in units of the planning period, and multiplying both sides of (9.10) by $Q_i$, we obtain the total number of units of product $i$ produced in the planning period as

$$f_i\left(Q_i, \overline{W}_i, Y_i, Q_i^{'}, W_i^{'}, Y_i^{'}\right) = Q_i Y_i = \frac{\overline{W}_i}{P_i + \frac{u_i P_i + \sum_{j \neq i} u_j P_j}{2\left(1 - u_i - \sum_{j \neq i} u_j\right)}} \tag{9.11}$$

which can be written out as

$$\begin{aligned} f_i\left(Q_i, \overline{W}_i, Y_i, Q_i^{'}, W_i^{'}, Y_i^{'}\right) &= Q_i Y_i \\ &= \frac{\overline{W}_i}{\left(s_i + p_i Q_i\right) + \frac{Y_i\left(s_i + p_i Q_i\right)^2 + \sum_{j \neq i} Y_j\left(s_j + p_j Q_j\right)^2}{2\left(1 - \left(s_i + p_i Q_i\right) - \sum_{j \neq i} Y_j\left(s_j + p_j Q_j\right)\right)}} \end{aligned} \tag{9.12}$$

where $Q_i^{'}$ denotes the vector of lot sizes $Q_j$ for all products except $i$, and $\overline{W'}_i$ and $Y_i^{'}$ are defined analogously. The MDCF (9.12) is an ugly, non-convex expression, but is actually quite intuitive: The output of a particular product $i$ in a planning period depends on its own lot size $Q_i$, the number of lots produced $Y_i$, and its time-average WIP level $\overline{W}_i$, as would be expected in a single-product model. However, it is also affected by the lot sizes, WIP levels, and number of lots of all other products. As seen in the intermediate expression (9.11), this is because these quantities determine the fraction of current machine utilization available to the product $i$ in the planning period. Thus the output mix of the machine is jointly determined by the set of $N$ MDCFs (9.12). The explicit consideration of lot sizing has resulted in the addition of state variables reflecting the lot sizes of each product during the planning period. An example of this MDCF is illustrated in Fig. 9.2; note that the level sets shown on the horizontal plane, which are the feasible combinations of WIP and lot sizes that yield the specified output, match those given by Karmarkar (1987).

Anli et al. (2007) present a MDCF with similar state variables, but take a very different approach to estimating it; they use an iterative approach between the

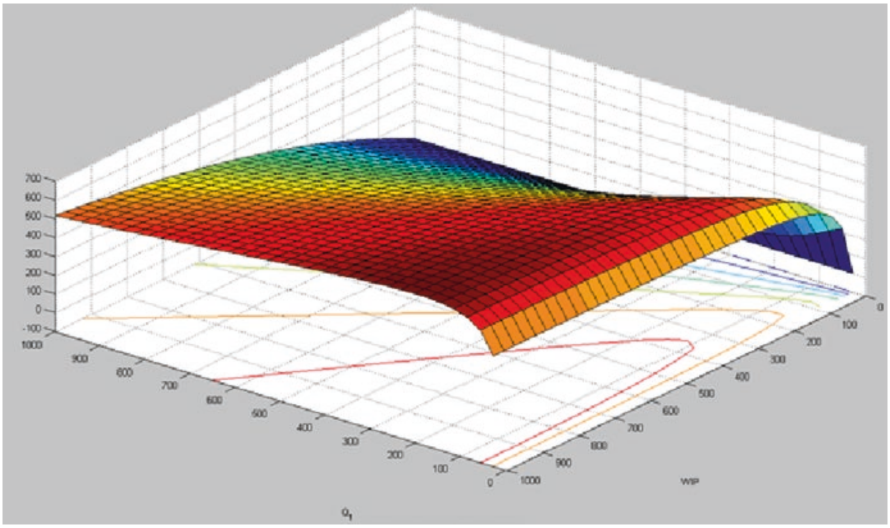


**Fig. 9.2** Illustration of MDCF with lot sizing for two-product system: output of product 1 as a function of its lot size and WIP for fixed lot size and output of product 2 (Kang et al. 2014)

individual production units for which the MDCF is being developed and the goods flow model. Tentative release plans are computed by the planning level, which are then used by the production units to estimate their realized performance. These realized performance estimates are then fed back to the planning level, which generates additional constraints derived from these estimates to refine its models of the capabilities of the production units. In the language of Schneeweiss (2003), tentative release plans are communicated to the production units and their feedback is then used to refine the planning level's anticipation functions for the production units. The approach of Anli et al. (2007) is unique in presenting an integrated, well thought out decomposition of the supply chain planning problem into multiple subproblems, including the goods flow problem, safety stock levels, and MDCFs for the individual production units, with promising computational results.

Several aspects of this MDCF are worthy of comment. Like Karmarkar (1987), it highlights the strong interdependence of products in a multiproduct queueing system: decisions made for any product, such as the level of output or the lot size, affect all other products. The use of this MDCF for a planning period of fixed finite length is clearly heuristic; the derivation assumes the queue is in steady state during the planning period, which is unlikely to be the case in general. The model also assumes that the lot sizes are decision variables associated with each planning period, and hence that these can be changed by management in each planning period. This is clearly possible for newly released orders, but it is unlikely that lots already released to production can be reconfigured without considerable disruption of ongoing operations. If the cycle time of some fraction of lots in each period exceeds the length of a planning period, it is thus likely that there will be lots of different sizes on the shop floor at least some of the time; the transient state refers not only to the number of orders at the workcenters but to the composition of the order sizes as well.

The MDCFs (9.12) can be incorporated into an integrated release planning and lot-sizing model in a straightforward manner, using the following notation:

**Decision variables:**
$Y_{it}$: number of lots of product $i$ produced in period $t$
$Q_{it}$: lot size of product $i$ in period $t$
$I_{it}$: finished goods inventory of product $i$ at the end of period $t$
$B_{it}$: amount of product $i$ backlogged at the end of period $t$
$W_{it}$: WIP of product $i$ at the end of period $t$
$\bar{W}_{it}$: time-average WIP level of product $i$ during period $t$
$R_{it}$: number of units of product $i$ released in period $t$

**Parameters:**
$h_{it}$: unit finished goods holding cost in period $t$
$w_{it}$: unit WIP holding cost in period $t$
$b_{it}$: unit backlogging cost in period $t$

The model can then be written as follows:

$$\min \sum_{i=1}^{N}\sum_{t=1}^{T}\left[h_{it}I_{it} + b_{it}B_{it} + w_{it}W_{it}\right] \tag{9.13}$$

subject to

$$I_{it} - B_{it} = I_{i,t-1} - B_{i,t-1} + Q_{it}Y_{it} - D_{it}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.14}$$

$$W_{it} = W_{i,t-1} + R_{it} - Q_{it}Y_{it}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.15}$$

$$\bar{W}_{it} = \frac{W_{it} + W_{i,t-1}}{2}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.16}$$

$$Q_{it}Y_{it} \le f_i\left(Q_t, \bar{W}_t, Y_t\right), \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.17}$$

$$\sum_{i=1}^{N} Y_{it}\left(s_i + p_i Q_{it}\right) < 1, \quad t = 1,\ldots,T \tag{9.18}$$

$$Q_{it}, Y_{it}, R_{it}, I_{it}, W_{it}, B_{it} \ge 0, \quad \text{integer} \tag{9.19}$$

Constraints (9.18) are a stability condition that is redundant when the MDCFs (9.17) are present; it is included in the model to help reduce the solution time. The model (9.13)–(9.19) is a single-stage multi-item dynamic lot-sizing model, with some interesting differences. The presence of the MDCFs leads to non-convex constraints, even when the integrality constraints are relaxed. In addition, traditional lot-sizing models focus on the tradeoff between the fixed cost of setups and inventory holding costs, while in this model setup costs are conspicuous by their absence. It can be argued that the actual cash costs of setup changes are relatively small and are usually limited to the scrap generated while adjusting the machine and tooling to the new product. In the short term, labor and machinery are all fixed costs, so the main component of a setup cost in a production environment is the opportunity cost of the lost production time. This opportunity cost, however, is difficult to estimate in practice. If the facility has sufficient excess capacity that the setup will not result in any loss of revenue, the opportunity cost of capacity associated with the setup is clearly zero; this is equally clearly not the case if the facility is highly utilized and setups result in lost sales due to reduced output.

Due to the complexity of the integer nonlinear program (9.13)–(9.19), Kang et al. (2014) relax the integrality constraints, solve the resulting non-convex model to a local optimum and then heuristically round the resulting fractional solution to an integer feasible solution. In a later paper (Kang et al. 2018), they propose a more sophisticated rounding heuristic that gives considerably improved solutions over the original approach. Due to the absence of setup costs, the performance of the model is compared to that of a model due to Erenguc and Mercan (1990), which requires some additional notation:

**Decision variables:**

$K_{it}$: binary variable equal to 1 if a setup is performed for product $i$ in period $t$, and zero otherwise

$X_{it}$: amount of product $i$ produced in period $t$

**Parameters:**

$M$: a very large number

The model can be stated as follows:

$$\min \sum_{i=1}^{N}\sum_{t=1}^{T}\left[h_{it}I_{it} + b_{it}B_{it}\right] \tag{9.20}$$

subject to:

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.21}$$

$$X_{it} \le MK_{it}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.22}$$

$$\sum_{i=1}^{N}\left[K_{it}s_{it} + p_i Q_{it}\right] \le 1, \quad t = 1,\ldots,T \tag{9.23}$$

$$K_{it} \in \{0,1\}, I_{it}, X_{it} \ge 0, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.24}$$

There are some interesting contrasts between this model and the MDCF-based model (9.13)–(9.19). The model of Mercan and Erenguc assumes that all production of a given product in a given period will be processed as a single lot, while the MDCF-based model allows multiple smaller lots. The Mercan–Erenguc model does not consider queueing effects at all, while these are central to the MDCF-based model. In fairness, the Erenguc–Mercan model was never intended to be used in a queueing environment, but rather for big-bucket lot-sizing or product cycling problems where queueing does not arise.

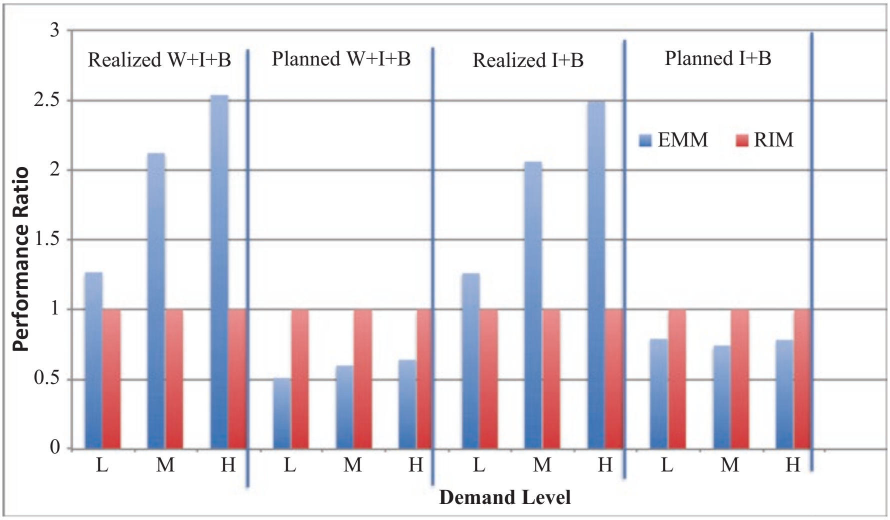


**Fig. 9.3** Performance comparison of Erenguc–Mercan model (EMM) and MDCF-based lot-sizing model (RIM)

The logical way to evaluate the performance of this integrated release planning and lot-sizing model is to simulate the behavior of the production system operating under the lot sizes and release quantities it suggests. Details of the computational experiments are given in Kang et al. (2014), but representative findings are summarized in Fig. 9.3. The planned quantities refer to the objective function values from the mathematical models, while the realized values are those observed when the decisions from the mathematical models are implemented. Since the Erenguc–Mercan model does not consider congestion, and hence ignores WIP, we report the objective functions with and without WIP costs to observe how well the mathematical models predict the consequences of their decisions. The simulation model relaxes the assumption of a constant lot size in each period; if all lots released in a given period have not exited the system by the start of the following period, lots with different sizes will coexist in the system.

It is clear from the figure that the failure of the Erenguc–Mercan model to consider WIP results in the MDCF model performing considerably better. The planned objective function of the Erenguc–Mercan model, which considers only inventory and backorder costs, is actually quite close to those components of the planned cost from the MDCF model. However, the ability of the MDCF model to produce the demand for a given period in a number of small batches results in considerable improvement in cycle times, and major differences in performance between the two models. Although it is not evident from the limited data shown, the differences between the two models are largest at low to medium demand levels. At high demand, and hence utilization, lot sizes have to be large in order for the system to meet demand. Hence all production of a product in a given period is processed in a single lot, as required by the Erenguc–Mercan model. At lower utilization levels, however, the MDCF model can take advantage of the available excess capacity by using smaller batches with more setups, resulting in lower flow times and better performance.

As discussed above, the development of this model rests on a number of heuristic assumptions: the use of steady-state queueing models to derive the MDCF and the approximate solution of the resulting nonlinear integer program by solving its continuous relaxation and rounding to an integer feasible solution. There is no doubt that each of these introduces errors, which are likely to grow as the length of the planning period decreases. However, the MDCF model is in any case unsuitable for short-term release planning due to the difficulty of adjusting lot sizes on the shop floor after release. The model is better viewed as a longer-term aggregate model that can be used to examine the impact of lot sizes in the presence of changing demand conditions. It is also likely that some of the more egregious errors introduced by these assumptions are remedied to some degree by the myopic rounding scheme implemented at the execution level in the simulation model.

## 9.3 Insights from a MDCF-Based Lot-Sizing Model

The MDCF-based lot-sizing model (9.13)–(9.19) is clearly an extension to the multi-item capacitated lot-sizing problem of the type studied by Billington et al. (1983) and Trigeiro et al. (1989) and reviewed extensively by Quadt and Kuhn (2008). These models, along with their many successors, focus on the tradeoff between the fixed costs of setups and inventory holding costs, while considering capacity constraints without congestion as reflected by constraint (9.23) in the Erenguc–Mercan model above. In this section, we present a column generation heuristic for the MDCF-based lot-sizing model developed by Kang et al. (2011), with the purpose of providing insight into the practical difficulties of estimating setup costs in production environments. Similar column generation approaches for capacitated lot-sizing problems without congestion have been developed by Lasdon and Terjung (1971) and de Graeve and Jans (2007).

The basic idea of column generation approaches for capacitated lot-sizing problems is to decompose the problem into a master problem that allocates capacity among the $N$ different products, and pricing subproblems that perform the optimal lot sizing for each product subject to the capacity allocation given by the master problem. Hence, in the Lasdon–Terjung approach (Lasdon and Terjung 1971), when the master problem allocates capacity, the pricing subproblems are single-item uncapacitated dynamic lot-sizing problems whose objective function is modified by the dual prices obtained from the master problem. Detailed presentations of Dantzig–Wolfe decomposition and column generation methods can be found in Desaulniers et al. (2005) and Lasdon (1970).

Following the usual approach to developing a column generation approach, let us denote the set of all feasible schedules for product $i$, $i = 1, \ldots, N$, by $\Upsilon_i$. Since all decision variables associated with a product $i$ in the model (9.13)–(9.19) must take integer values, the sets $\Upsilon_i$, $i = 1, \ldots, N$ will each consist of a very large number of discrete schedules. Let $\tau_i^k$ denote a column vector with $T$ entries associated with a solution $k \in \Upsilon_i$ whose $t$th entry is the capacity required by product $i$ in period $t$ for schedule $k$, given by

$$\tau_{it}^k = \left(s_i + p_i Q_{it}^k\right) Y_{it}^k \tag{9.25}$$

where $Q_{it}^k$ denotes the lot size of product $i$ in period $t$ in the schedule $k \in \Upsilon_i$ and $Y_{it}^k$ denotes the number of lots of product $i$ produced in period $t$ in this schedule. We also define the cost vector $V_i^k$ as a column vector with $T$ entries.

$$V_{it}^k = h_{it} I_{it}^k + w_{it} W_{it}^k + b_{it} B_{it}^k \tag{9.26}$$

Defining the decision variables

$$\gamma_k^i = \begin{cases} 1, \text{ if schedule } k \in \Upsilon_i \text{ is selected for product } i \\ 0, \text{ otherwise} \end{cases} \tag{9.27}$$

we can rewrite the model (9.13)–(9.19) as that of selecting exactly one schedule for each product such that the resulting schedules are capacity feasible and the objective function is minimized. The resulting master problem is given by:

(Master Problem: MP)

$$\min \sum_{i=1}^{N} \sum_{k \in \Upsilon_i} V_i^k \gamma_i^k \tag{9.28}$$

subject to

$$\sum_{k \in \Upsilon_i} \tau_{it}^k \gamma_i^k \le C_t, \quad i = 1, \ldots, N, \; t = 1, \ldots, T \tag{9.29}$$

$$\sum_{k \in \Upsilon_i} \gamma_i^k = 1, \quad i = 1, \ldots, N \tag{9.30}$$

$$\gamma_i^k \in \{0,1\}, \quad k \in \Upsilon_i, \; i = 1, \ldots, N \tag{9.31}$$

Since this is a binary set covering problem that is hard to solve, we relax the integrality constraints (9.31), replacing them with

$$0 \le \gamma_i^k \le 1, \quad i = 1, \ldots, N, \; k \in \Upsilon_i \tag{9.32}$$

to obtain the relaxed master problem (RMP).

$$\min \sum_{i=1}^{n} \sum_{k \in \Upsilon_i} V_i^k \gamma_i^k \tag{9.33}$$

subject to

$$\sum_{i=1}^{N}\sum_{k\in\Upsilon_i}\tau_{it}^{k}\gamma_i^k \leq C_t, \quad t=1,\ldots,T \tag{9.34}$$

$$\sum_{k\in\Upsilon_i}\gamma_i^k = 1, \quad i=1,\ldots,N \tag{9.35}$$

$$0 \leq \gamma_i^k \leq 1, \quad i=1,\ldots,N,\ k\in\Upsilon_i \tag{9.36}$$

Since enumerating all columns in the RMP is impractical, we use a restricted relaxed master problem (RRMP) with a limited number of columns that are generated by a column generation approach. The RRMP is initialized with an initial set of columns and solved to optimality. However, this solution is only optimal with respect to the limited set of columns considered in the RRMP; there may yet exist columns in some of the sets $\Upsilon_i$ that have not yet entered the RRMP, but which might improve the objective function if they were to enter, i.e., have negative reduced costs. A pricing subproblem is thus solved for each product $i = 1,\ldots, N$ to determine whether any columns with negative reduced costs exist.

To formulate the pricing subproblem for product $i$, we define $\alpha_{it}^k$ to be the dual variable associated with the capacity constraints (9.34) and $\propto_i^k$ those associated with constraints (9.35). Then the reduced cost for a new column to enter the basis of the RRMP will be

$$\sum_{t=1}^{T}\left[V_{it}^k + \alpha_{it}^k\left(s_i + p_i Q_{it}^k\right)Y_{it}^k\right] + \mu_i^k \tag{9.37}$$

The pricing subproblem seeks a schedule $k$ for product $i$ such that the reduced cost is negative; if no such schedule can be found for any product an optimal solution to the relaxed master problem has been obtained. We can thus state the pricing subproblem for product $i$, $i = 1,\ldots, N$, as follows:

$$\min \sum_{t=1}^{T}\left[V_{it}^k + \alpha_{it}^k\left(s_i + p_i Q_{it}^k\right)Y_{it}^k\right] + \mu_i^k \tag{9.38}$$

subject to

$$W_{it}^k = W_{i,t-1}^k + R_{it}^k - Q_{it}^k Y_{it}^k, \quad i=1,\ldots,N,\ t=1,\ldots,T \tag{9.39}$$

$$I_{it}^k = I_{i,t-1}^k + Q_{it}^k Y_{it}^k - D_{it}, \quad i=1,\ldots,N,\ t=1,\ldots,T \tag{9.40}$$

$$Q_{it}^k Y_{it}^k \leq f_i\left(Q_{it}^k, Y_{it}^k, \overline{W}_{it}^k, \widehat{\rho_{jt}^k}, \widehat{\rho_{jt}^k T_{jt}^k}\right), \quad t=1,\ldots,T \tag{9.41}$$

$$Q_{it}^k, Y_{it}^k, I_{it}^k, W_{it}^k, R_{it}^k, B_{it}^k \geq 0, \quad t=1,\ldots,T \tag{9.42}$$

where

$$\widehat{\rho_{jt}} = \sum_{j\neq i}\sum_{m\in\Upsilon_j}\tilde{\gamma}^k_{jt}Y^k_{jt}\left(s_j + p_jQ^k_{jt}\right) \tag{9.43}$$

denotes the utilization on the machine due to products other than $i$ in the optimal solution to the restricted master problem at the current iteration. This pricing subproblem is a single-item dynamic lot-sizing problem, where the amount of capacity available to the product $i$ is fixed by decisions corresponding to the other products. Dropping the constant $\propto^k_i$ we can write the objective function (9.38) as

$$\sum_{t=1}^{T}\left[h_{it}I^k_{it} + w_{it}W^k_{it} + b_{it}B^k_{it} + \alpha^k_{it}\left(s_i + p_iQ^k_{it}\right)Y^k_{it}\right] \tag{9.44}$$

and compare this to the objective function of the classical capacitated dynamic lot-sizing problem, which is given by:

$$\sum_{t=1}^{T}\left[h_{it}I_{it} + b_{it}B_{it} + S_{it}\Xi_{it}\right] \tag{9.45}$$

where $\Xi_{it}$ is a binary variable equal to 1 if product $i$ is produced in period $t$, and zero otherwise, while $S_{it}$ denotes the fixed cost of a setup. Note that classical capacitated lot-sizing models all assume a single lot of each product in a given period, which would require the additional constraint $Y^k_{it}$ ″ 1. Matching equivalent terms in (9.44) and (9.45) shows that the two objectives treat finished goods inventory and backlogs identically. Since the classical formulations do not consider congestion, and hence ignore WIP, let us assume that the cost of holding WIP is negligible. In this case, for (9.44) and (9.45) to give the same value, and hence the same solution, we must have

$$S_{it} = \alpha^k_{it}\left(s_i + p_iQ^k_{it}\right) \tag{9.46}$$

showing that even under the very restrictive assumptions imposed to achieve compatibility between the classical and MDCF-based lot-sizing models, the fixed cost of a setup must depend on the dual price of capacity at optimality—which is impossible to determine without obtaining an optimal solution for all products simultaneously.

Thus, while classical dynamic lot-sizing models can be justified in a purchasing environment, or in an environment with significant excess capacity, their use in a production environment is fraught with problems. Once the utilization of the machine reaches a certain point, it will become necessary to produce exactly one lot of each item in each period in which it is to be produced; however, the relative magnitudes of the fixed setup costs relative to inventory holding costs will determine the frequency of production. At lower utilization levels, however, (9.44) suggests that estimating the setup cost is far from trivial; at the very least, the setup cost for a product $i$ will be time dependent, driven by the evolution of its demand over time as well as that of all other products competing with it for capacity.

## 9.4 Discussion

In this chapter, we have seen that the queueing perspective of Chap. 2 leads quite naturally to a series of models describing the impact of lot-sizing decisions on the performance of production units. The MDCFs developed in the previous chapter turn out to be a suitable mechanism to describe the behavior of such systems in mathematical programming models. The resulting optimization models are generally non-convex, requiring significant additional computational effort to guarantee a global optimal solution. However, there is considerable computational evidence that the non-convexity is of a somewhat benign nature; in many cases, the use of a convex nonlinear solver leads to confirmed global optimal solutions, suggesting the existence of considerable structure in the problem that remains an objective for future research. The use of steady-state queueing models to develop the MDCFs is clearly heuristic, and open to criticism; however, the significant improvements in system performance obtained in simulation experiments suggest that these models are worth developing further.

The contrast between these models and the traditional lot-sizing models that focus on the tradeoff between setup and holding costs is also informative. The (admittedly heuristic) column generation approach outlined in Sect. 9.3 highlights the complexity of estimating setup costs accurately. The results of Sect. 9.2, on the other hand, highlight a central implication of the traditional lot-sizing models such as the multilevel capacitated lot-sizing problem and its variants, which is that all production for a planning period must be produced in a single lot. Given that setup costs are charged on a per lot basis, this is natural, but the superior shop-floor performance obtained by the MDCF model suggests that especially at medium levels of utilization the use of smaller lot sizes can lead to considerable benefits. At high utilization levels there is no capacity to spare for additional setups, and hence the results of the traditional models and the MDCF model approach each other.

The work in this chapter is clearly exploratory in nature and merely scratches the surface of a broad and complex research agenda. The extension of this type of approach to multistage systems, such as those treated by Missbauer and Jutz (2018), or multilevel systems such as those arising in the context of MRP computations is a natural direction. It is unlikely that exact solutions to such formulation can be obtained for industrial scale problem instances, especially given the non-convex nature of many MDCF planning models, but an understanding of the structure of good solutions should serve as a pathway to computationally efficient approximations.

## References

Anli OM, Caramanis M, Paschalidis IC (2007) Tractable supply chain production planning modeling non-linear lead time and quality of service constraints. J Manuf Syst 26(2):116–134

Benjaafar S (1996) On production batches, transfer batches and lead times. IIE Trans 28:357–362

Billington PJ, Mcclain JO, Thomas JL (1983) Mathematical programming approaches to capacity-constrained MRP systems: review, formulation and problem reduction. Manag Sci 29:1126–1141

Brahimi N, Dauzere-Peres S, Najid NM, Nordli A (2006) Single-item lot sizing problems. Eur J Oper Res 168:1–16

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs

Cheng M, Mukherjee NJ, Sarin SC (2013) A review of lot streaming. Int J Prod Res 51(23/24):7023–7046

DeGraeve Z, Jans R (2007) A new Dantzig-Wolfe reformulation and branch and price algorithm for the capacitated lot sizing problem with setup times. Oper Res 55(5):909–920

Desaulniers G, Desrosiers J, Solomon MM (2005) Column generation. Springer, Berlin

Drexl A, Kimms A (1997) Lot sizing and scheduling—survey and extensions. Eur J Oper Res 99:221–235

Erenguc SS, Mercan M (1990) A multifamily dynamic lot-sizing model with coordinated replenishments. Nav Res Logist 37:539–558

Harris FW (1915) Operations and cost. In: Factory management series. A. W. Shaw Co, Chicago

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Jen Huei C, Huan Neng C (2005) A comprehensive review of lot streaming. Int J Prod Res 43(8):1515–1536

Jutz S (2017) Lot sizing in a two-stage production-inventory system—a flow time oriented perspective. Department of Information systems, production and logistics management. PhD, University of Innsbruck, Innsbruck, p 194

Kang Y, Albey E, Uzsoy R (2011) A column generation approach for multiple product dynamic lot-sizing with congestion. Raleigh, NC 27695-7906, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University

Kang Y, Albey E, Hwang S, Uzsoy R (2014) The impact of lot-sizing in multiple product environments with congestion. J Manuf Syst 33(3):436–444

Kang Y, Albey E, Uzsoy R (2018) Rounding heuristics for multiple product dynamic lot-sizing in the presence of queueing behavior. Comput Oper Res 100:54–65

Karmarkar US (1987) Lot sizes, lead times and in-process inventories. Manag Sci 33(3):409–418

Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123

Karmarkar US, Kekre S, Kekre S (1992) Multi-item batching heuristics for minimization of queues. Eur J Oper Res 58:99–111

Lasdon LS (1970) Optimization theory for large systems. Macmillan, New York

Lasdon LS, Terjung RC (1971) An efficient algorithm for multi-item scheduling. Oper Res 19(4):946–969

Medhi J (1991) Stochastic models in queuing theory. Academic, Cambridge

Missbauer H (2002) Lot sizing in workload control systems. Prod Plan Control 13:649–664

Missbauer H, Jutz S (2018) A flow time oriented lot sizing model for a serial two-stage production-inventory system: analytical approximation and simulation-based optimization. In: Twentieth international working seminar on production economics, Innsbruck, Austria

Pochet Y, Wolsey LA (2006) Production planning by mixed integer programming. Springer, New York

Quadt D, Kuhn H (2008) Capacitated lot sizing with extensions: a review. 4OR Q J Oper Res 6:61–83

Schneeweiss C (2003) Distributed decision making. Springer, Berlin

Söhner V, Schneeweiss C (1995) Hierarchically integrated lot size optimization. Eur J Oper Res 86(1):73–90

Trigeiro WW, Thomas LJ, McClain JO (1989) Capacitated lot sizing with setup times. Manag Sci 35:353–366

Vaughan TS (2006) Lot size effects on process lead time, lead time demand, and safety stock. Int J Prod Econ 100:1–9

Wijngaard J (1989) Timing and lot-sizing in production control. J Manuf Oper Manag 2:35–51

Zipkin PH (1986) Models for design and control of stochastic, multi-item batch production systems. Oper Res 34(1):91–104

# Chapter 10
# Applications of Clearing Functions

The previous chapters have motivated the need for more advanced anticipation functions that can reflect, at least to a reasonable level of accuracy, the nonlinear relations between the workload of a production resource and its expected throughput. Whatever their academic interest, one would hope that the clearing function formalism could provide new insights or performance advantages over the models based on fixed exogenous lead times described in Chap. 5. We have already seen some such insights; models based on a simple upper bound on resource loading in a planning period and constraint can only provide meaningful dual prices for resources that are fully utilized, i.e., for whom the associated upper bound constraint in the period is satisfied at equality. Chapter 7 illustrates how the dual information obtained from the ACF model is considerably richer and more nuanced than that from the fixed lead time models. This chapter presents several studies where clearing functions have been applied to different problems related to production systems. In several cases, the use of clearing functions provides interesting insights that would be difficult to obtain using the conventional approach of exogenous planned lead times and maximum capacity loading.

## 10.1 Release Planning in Semiconductor Wafer Fabrication

Kacar et al. (2013, 2016) have conducted extensive computational experiments comparing the performance of clearing function models to those based on fixed lead times in the context of semiconductor wafer fabrication using the MIMAC I benchmark data sets (Fowler and Robinson 2012). The experiments proceeded in two stages: in the first, the allocated clearing function (ACF) formulation of Asmundsson et al. (2006, 2009) was compared to a fixed lead time formulation based on that given in Sect. 5.2 using integer lead times. The second paper extends this by considering fractional fixed lead times.

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_10

The MIMAC I data set, which is publicly available at https://www.sim.unihannover.de/~svs/wise0809/pds/masmlab/factory−datasets/, represents a wafer fab with more than 200 machines grouped into 84 workcenters. Processing times are deterministic, and FIFO dispatching and instantaneous material transfer between processing steps (operations) are assumed. Variability in the system arises from exponentially distributed machine failures as well as from different processing times for different operations of the two products. The facility produces two products, one with 210 unit operations and the other with 245, and the number of steppers at the photolithography workcenter is adjusted to ensure that this workcenter has the highest long-term average utilization when the two products are produced in equal proportions. Batch processing machines, where a number of lots requiring the same operation can be processed simultaneously, are represented as well as machines with sequence-dependent setup times.

Each operation in a product routing has exactly one predecessor and one successor, except for the first and last operations in the routing. Material that is processed at each operation is assumed to become available to the next operation in the product's routing as soon as it completes its processing at the current operation, and strategic inventory is not held anywhere in the fab except after the final operation. The objective is to maximize the total contribution over the planning horizon, given by the total revenue minus the sum of WIP holding, release, finished goods inventory holding and backordering costs.

The formulation with integer lead times used in the first set of experiments described in Kacar et al. (2013) is essentially that given in (5.26)–(5.29) in Chap. 5. An immediate issue is how to set the planned lead times ($L_j$ and $L_{jk}$ in the notation of Chap. 5). The authors approach this issue using the mean cycle time at each workcenter obtained from a simulation model of the fab, which was run at utilization levels corresponding to those in the experimental design (0.7 and 0.9), and observing the realized cycle times. Since, as might be expected, many of these quantities yield a fractional number of periods, the authors propose two alternative approaches to convert these into integer lead times. In the first of these, referred to as simple rounding down (SRD), all fractional lead times are truncated to the next lower integer value. The obvious disadvantage of this approach is that since all mean cycle times are rounded down, lead times will consistently underestimate the actual cycle times. An alternative approach seeks to obtain the workcenter lead times $L_{jk}$ by rounding some workcenter cycle times up and others down in a manner consistent with the observed overall mean cycle time $L_j$. This is accomplished by solving an integer program initially proposed by Turkseven (2005) that minimizes the sum of the integer lead time estimates subject to constraints ensuring they cannot be less than the observed cumulative cycle times from which they are derived. Details of the formulation are given in Turkseven (2005) and Kacar et al. (2013). The fixed lead time model using the integer lead times obtained in this manner is referred to as the IPR model.

The clearing functions for the workcenters are fitted to data from extensive simulation experiments. Demand realizations yielding seven levels of average bottleneck utilization ranging from 0.5 through 0.97 are generated, and the system simulated assuming releases of each product in each period are equal to the demand in that period. The workload and output of each workcenter are observed for each planning period, and the observations are pooled. For each workcenter, the workload axis is partitioned into two segments, each containing an equal number of data points, and simple linear regression is used to fit a linear function to the data in each segment. With the benefit of today's knowledge substantially better CF fits can be obtained, using either simulation optimization (Kacar and Uzsoy 2015) or the iterative refinement approach of Gopalswamy and Uzsoy (2019), at the expense of additional computational effort. These clearing functions are embedded in an optimization model based on (7.70)–(7.75), with the piecewise linear constraints (7.76) representing the clearing functions.

The three different models—ACF, SRD, and IPR—are tested under an extensive experimental design that compares constant versus time-varying mean demand, as well as different levels of demand variability and production uncertainty due to machine failures. The experiments consist of solving the optimization models for five independent demand realizations in each experimental configuration. The releases obtained from the optimization model are then fed into a detailed simulation model of the fab, using 20 independent replications, to observe their realized performance under random machine failures. Thus the results represent the estimated performance of the production system when the decisions from the optimization models are implemented.

Representative results from this experiment with constant mean demand are shown in Fig. 10.1. The first field of the labels denotes the machine failures (short or long); the second the average utilization (70 or 90%); and the third the demand variability (LowCV or HighCV). The ACF formulation consistently produces higher expected profit than SRD and IPR, by a considerable margin in several cases. The IPR model is dominated by SRD under short machine failures, but achieves comparable performance under the long machine failures, i.e., with higher variability in the production system. The IPR model consistently yields longer planned lead times than SRD, bringing to light a weakness in the original experimental design: the impact of a finite planning horizon on the release decisions is not well accounted for. Faced with a finite planning horizon, all models will reduce releases in the final periods of the planning horizon to avoid excessive inventories remaining on hand at the end of the horizon. Other things being equal, a fixed lead time model with longer planned lead times reduces releases earlier than one with shorter planned lead times, resulting in reduced revenue in the final periods of the horizon. This appears to be the case with IPR and SRD; the total costs of the two models are roughly comparable in magnitude, but the profit for IPR is generally lower than that for SRD. The results for time-varying mean demand are qualitatively similar and are not presented here for reasons of space.

These results suggest that the ACF model consistently yields better performing production plans compared to the use of fixed integer lead times. While SRD yields higher profit solutions than IPR in this experiment, some of the difference in performance may be due to the different behavior of the two models in the ending periods of the planning horizon, which will affect IPR more severely than SRD. Thus these

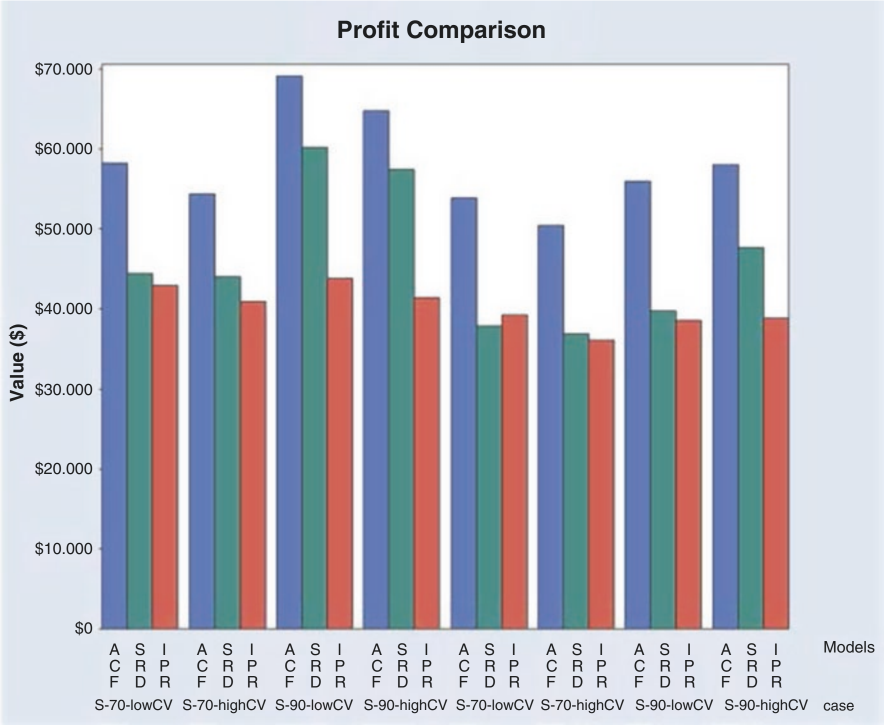


**Fig. 10.1** Comparison of planning models for constant mean demand

experiments should not be viewed as demonstrating the superiority of SRD over IPR; additional experiments carefully controlling the end of horizon effects as suggested by Hung and Leachman (1996) should be performed to carefully examine this issue. Nevertheless, these initial experiments suggest that the clearing function models have the potential to yield significantly better solutions than the fixed lead time models with integer planned lead times. It should be noted, in fairness, that the ACF formulation results in a significantly larger formulation than the fixed lead time models; each operation of each product at each workcenter requires its own decision variables, which, together with the use of multiple linear segments, significantly increases the size of the model (Kacar et al. 2016).

The obvious next step of comparing the ACF model to fixed lead time models with fractional lead times is performed in Kacar et al. (2016). The SRD and ACF formulations and the overall experimental design from the previous paper are retained and compared to three different models with fractional lead times: FLT-C, which considers fractional lead times for capacity loading constraints (5.58) only; FLT-I, which considers fractional lead times for finished goods inventory balance constraints only; and FLT-B, which considers both constraint sets. End-of-horizon effects were accounted for systematically in this experiment, providing a more accurate comparison of the different models, as well as expressions for the size of the formulations in terms of number of decision variables and constraints.

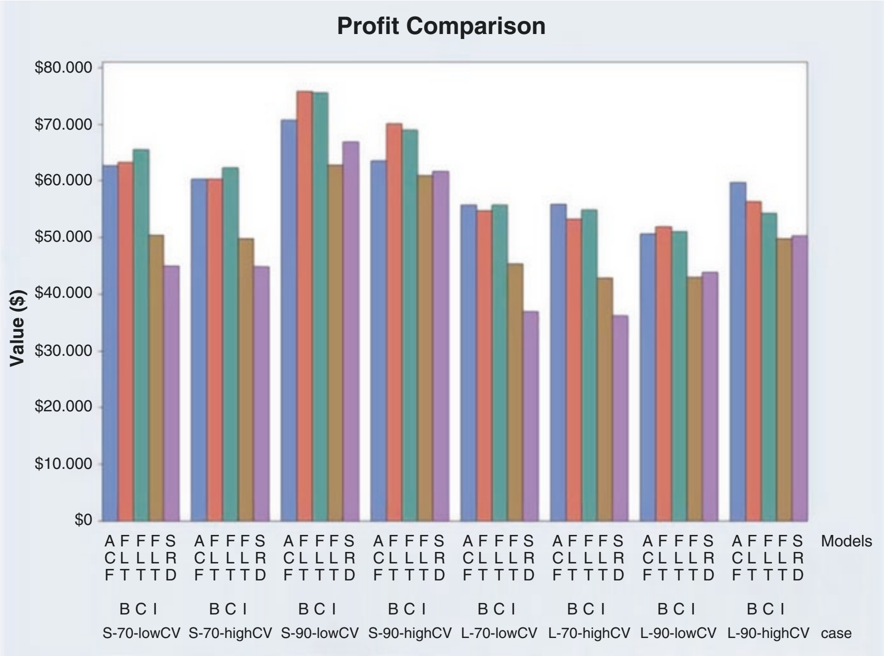


**Fig. 10.2** Profit comparison for constant average demand and fractional lead times

Comparisons of the expected profit under constant average demand are shown in Fig. 10.2. While ACF retains its superior performance over SRD, FLT-B achieves the same performance as ACF for all practical purposes, suggesting that under conditions of constant average load and product mix the advantage of the clearing function models is largely lost. Upon reflection, this is intuitive; if the average workload on the workcenters remains constant over time, the average cycle times will also remain approximately constant, and the planned lead times will accurately reflect at least the average observed cycle time. This line of thought would also suggest that when the average workload is varying over time, the clearing function models ought to perform better than a fixed lead time model with constant lead times across the entire horizon. The results of this experiment with time-varying average demands are shown in Fig. 10.3. Under short machine failures, FLT-B again outperforms ACF. However, under the long failures, ACF is now the better performer on average, although the advantage is not very large; detailed statistical analysis is given in the original paper. Additional experiments with larger demand variations suggest that the performance advantage of ACF increases as the magnitude of the changes in the average demand, and hence the average workload, increase. Albey and Uzsoy (2015) extend the experiments described above for the smaller, scaled down wafer fab data set of Kayton et al. (1997) and find that when simulation optimization is used to obtain approximately optimal planned lead times with perfect visibility of the demand, both the ACF and the FLT models with a constant fractional planned

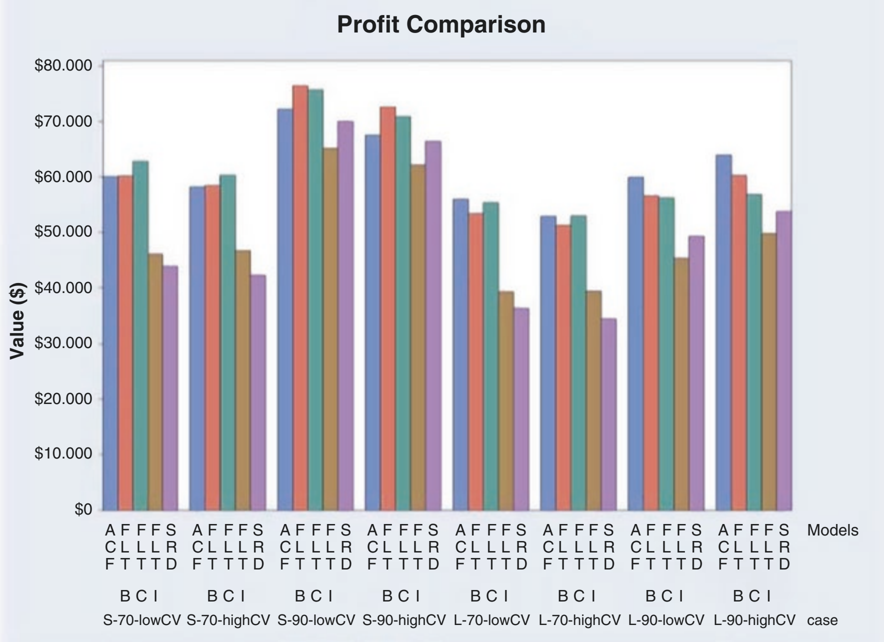


**Fig. 10.3** Expected profit comparison for time-varying average demands

lead time across the entire planning horizon achieve quite similar performance, with a slight advantage to the ACF model. The gap between the model with optimized lead times and the ACF formulation is of the order of 4%, suggesting that both ACF and the FLT model are very close to the best that can be expected without the benefit of hindsight.

The results of these experiments suggest that despite the greater sophistication and complexity of the clearing function models, their superior performance should not be taken for granted. The (in hindsight) relatively simplistic methods of fitting the clearing functions may be doing the ACF model a disservice, but the causes of this behavior require further study. Albey and Uzsoy (2015) found that while all three models represented lead times at the bottleneck workstation quite accurately, there were significant differences between planned and realized lead times at other machines; in particular, although the FLT model assumes a constant planned lead time across the entire planning horizon the observed cycle times varied quite dramatically over time. This observation, taken with the fact that the ACF and FLT models turn in essentially equal performance in terms of expected profit, suggests that fixed lead time LP models may be robust to errors in the planned lead time estimates, at least under some experimental conditions. Although the clearing function approach provides some theoretical advantages, such as the improved dual variables, there clearly remains a lot to learn about its performance in practice.

Applying the order release models in rolling horizon planning substantially changes their relationship to the actual material flow since only the first-period decisions are actually executed. Input-output control models that consider fixed lead times as time windows within which the operations can be scheduled arbitrarily (see Sect. 5.6) anticipate production smoothing more accurately and might also change the relative performance of models with fixed and variable lead times. Some of these issues are discussed in more detail in Sect. 10.2.

In summary, these comparisons of clearing function models for the problem they were originally designed to address—that of determining releases into a production system to optimally match its output to demand—are less definitive than we would like. The clearing function models appear to outperform fixed lead time models with integer planned lead times by a sometimes considerable margin in a static solution environment, where the planning model is solved for the entire planning horizon under deterministic demand and the resulting releases implemented without revision in the face of stochastic production events such as machine failures. However, this conclusion must be treated with caution given the limitations of the experimental design discussed above, and the use of a constant fractional planned lead time over the entire horizon brings the performance of the fixed lead time models to parity with the ACF model over a wide range of experimental conditions. Clearing functions appear to retain an advantage over fractional lead times when the workload on the production system varies over time, and both ACF and fractional lead time models are outperformed by a fixed lead time model using optimized lead times that differ in each period (Albey and Uzsoy 2015). This latter is only to be expected, since the use of a single clearing function to represent the system over time is clearly an approximation. However, the consistently strong performance of fixed lead time models with constant fractional lead times is less easy to explain. An obvious place to seek improvements is in the fitting of the clearing functions themselves, for which significantly better approaches now exist than were available when the Kacar et al. studies were performed. A puzzling aspect of these experiments is that the fractional fixed lead time models with a time-stationary planned lead time perform well, despite significant variations in the observed cycle times at the resources over time, while the same model with fixed integer lead times results in significantly poorer performance. The reasons for this behavior would yield interesting insights into the relation between planning and execution, and require careful experimental work.

## 10.2 Release Planning in a Rolling Horizon Environment

The order release models described in the preceding chapters, like most multi-period production planning models in the literature, assume deterministic demand. Clearly this does not accurately reflect reality, which is characterized by uncertainty in almost all model parameters, particularly the demand, which can only be forecast with limited accuracy.

Incorporating demand uncertainty into multi-period models for order release planning can be approached in various ways, depending on the formulation of the planning problem, the particular assumptions about the structure of uncertainty (e.g., uncertainty of total demand vs. uncertainty of the product mix) and the modeling approach chosen (e.g., stochastic programming vs. robust optimization). A comprehensive review of this extensive literature (Mula et al. 2006; Dolgui and Prodhon 2007; Dolgui et al. 2013) is beyond the scope of this volume. However, a common technique by which production planning systems address uncertainty and information evolution over time is the use of *rolling horizon planning*, often combined with safety stocks to absorb unplanned demand and supply variations (Sahin et al. 2013). Under this approach, which is also common in industrial practice, the production planning (or order release) problem for a $T$-period problem is approximated by solving a subproblem considering only the next $H$ periods at the start of each period $s$, $1 \leq s \leq T–H$, from which only the releases for the current period $s$ are actually executed. One period later, at the start of period $s + 1$, the system state and demand forecasts are updated based on new information and the next planning run is performed for period $s + 1,\ldots,s + H + 1$. As time advances, the planning horizon rolls through time as depicted in Fig. 10.4. An important aspect of this approach is that not only does information on the state of the system and demand forecasts evolve over time, the tentative decisions for a planning period are also revised several times before they are actually implemented. A considerable body of research dating back at least 40 years (Baker 1977; Blackburn and Millen 1980) shows that solution procedures that are optimal in a static setting fail to provide optimal solutions in a rolling horizon setting. Issues of nervousness were also rapidly discovered and addressed in a variety of ways (Blackburn et al. 1985). A recent discussion of these issues is given by Lin and Uzsoy (2016a, b). Thus, it is not at all obvious that the (admittedly limited) advantages displayed by clearing function models in the static settings of the previous section will carry over to a rolling horizon setting.

Addressing these issues, Pürgstaller and Missbauer (2012) compare a rule-based order release mechanism similar to LUMS (Sects. 4.2.3 and 4.4) and an order release model based on input/output control (Sect. 5.6) in a rolling horizon environment for the make-to-order CD/DVD manufacturer described in Chap. 1. Demand variability, product mix variability, and forecast accuracy are treated as

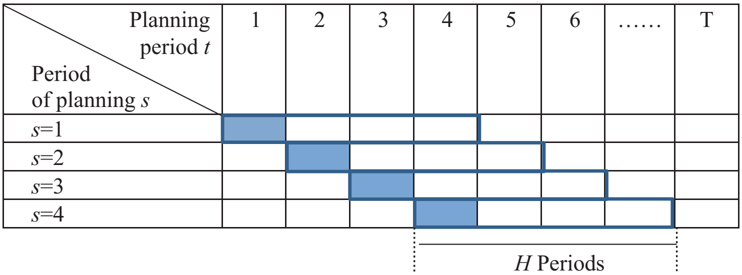


**Fig. 10.4** Release planning with rolling horizon

experimental factors; the forecast errors in consecutive periods are uncorrelated and increase as the future periods become more remote. They find that even under rather poor forecast accuracy the release planning model still outperforms the rule-based release mechanism. This suggests that exploiting the advance demand information provided by forecasts, even if it is error-prone, is beneficial for order release, which is not evident since rule-based order release mechanisms do not require explicit demand forecasts. The apparent explanation is that even substantial forecast errors have limited effect upon the first-period decisions in each period, and thus do not strongly affect actual releases. This is consistent with the analysis of linear decision rules in Holt et al. (1960), where for a quadratic objective function only the expected value of the future demand enters the optimal decision rule (p. 123) and the weights of the demand forecasts in the decision rule “become rapidly smaller as the sales period becomes more remote” (p. 118), which also holds for the costs of forecast errors (p. 173).

Häussler and Missbauer (2019) extend the Pürgstaller–Missbauer study by comparing the input/output control model with the ACF clearing function model developed in Chap. 7 in a rolling horizon setting, again for the make-to-order CD manufacturer from Chap. 1. They assume perfect demand forecasts that are “consumed” by customer orders that arrive a certain time before their due date. The distribution of this due date slack is derived from data representing the case company. They find that considering the predicted demand in the order release model outperforms an alternative approach that only considers confirmed orders already admitted to the order pool. This holds consistently for the ACF model throughout and in most cases also for the input/output control model. Given the low sensitivity of the input/output control model to forecast errors found in Pürgstaller and Missbauer (2012), it is reasonable to expect these insights to extend to scenarios with forecast errors, but this must be explored in future studies.

Albey et al. (2015) implement the chance-constrained production planning model initially developed by Norouzi and Uzsoy (2014) in a rolling horizon framework with stationary demand, using the additive Martingale Model of Forecast Evolution to represent demand. Their experiments consider a single-stage single-item capacitated production-inventory system and derive their test data from a major semiconductor manufacturer. The formulation of chance constraints explicitly considering forecast evolution exploits advance demand information with the potential to improve release decisions. Their experiments find that this is indeed the case, although the benefit of advance demand information is much reduced under high capacity utilization. This is intuitive, since when capacity is highly utilized across the entire planning horizon the optimal course of action is to simply keep the resource operating at full capacity especially when, as in this case, clearing functions are not used to capture congestion effects. A subsequent paper (Albey and Uzsoy 2016) extends this approach to a multiproduct environment using simulation optimization.

Ziarnetzky et al. (2018) significantly extend the work of Albey et al. (2015) by studying an ACF-based production planning model applied to a scaled-down simulation model of a wafer fab (Kayton et al. 1997) in a rolling horizon setting. Release quantities and safety stocks are determined simultaneously using shortfall-based chance constraints (Norouzi 2013). The paper considers two levels of demand variability, multiplicative and additive versions of the Martingale Model of Forecast Evolution, and different correlation structures for the demand forecast updates made at each planning period. They find that the model variant that considers forecast updates outperforms the variant without forecast update with respect to expected service level and profit. They conclude that "considering forecast evolution in production planning formulations for wafer fabs leads to improved performance as long as there is some excess capacity that can be exploited by the planning formulation" (p. 6130).

The impact of capacity utilization on the benefit of advance demand information is analyzed by Wijngaard (2004), who models advance demand information as the positive difference between the customer order lead time and the throughput time. He hypothesizes that "in case of highly flexible production, it is not necessary to have advance demand information," while on the other hand "in case of an inflexible production, inventory is necessary anyway and this inventory dampens the effect of precise advance demand information" (p. 96). This suggests a complex relationship between the value of advance demand information (i.e., of modeling forecast evolution in a rolling horizon setting) and the flexibility of the manufacturing system. Wijngaard and Karaesmen (2007) show that "when customer order lead times are less than a threshold value, it is allowed to aggregate the orders over time when establishing the optimal production decision" (p. 643).

Rolling horizon planning is characterized by sequentially performing optimization, implementing the first-period releases, followed by update of the demand forecast and system state and a new optimization at each planning period. This process can conceptually be approached using scenario-based stochastic programming (Birge and Louveaux 1997), robust optimization (Bertsimas and Sim 2004), or simulation optimization (Fu 2015). Aouam and Uzsoy (2012) compare stochastic programming to a linear decision rule that represents a base stock policy. At the start of the planning horizon, initial planned release and production quantities are determined by a modified clearing function model that includes chance constraints that avoid stockout with high probability. As the model rolls through time release and production quantities are modified based on updated demand and inventory information to obtain the implemented releases and observed production values. They find that when appropriately parameterized this procedure can compete effectively with multistage stochastic programming with a limited number of scenarios. However, the computational scalability of stochastic optimization methods and how to set their various parameters remains a question for further research (Aouam and Uzsoy 2015). It is interesting that by and large, the extensive literature on stochastic programming appears to have had very little impact on the domain of production planning.

Thus, while clearing function based models seem to retain at least some of their performance advantages in a rolling horizon setting, the study of production and release planning algorithms in rolling horizon environments is a complex problem with many interacting elements, making exact analytical treatments difficult. An obvious question that arises in the context of rolling horizon planning is that of how long the forecast window used at each decision epoch must be for the first-period decisions, i.e., the releases that are actually implemented, to be optimal in some sense with regard to the infinite, or at least longer, horizon problem under consideration. A forecast window length that guarantees at least optimal first-period decisions is referred to as a forecast horizon. For production planning models facing seasonal demand, it was recognized very early that "the relevant expectation and planning horizon will tend to cover a full seasonal cycle (or a shorter interval yet if storage costs are high) but is not likely to extend beyond this cycle except in the presence of a rapidly rising over-all trend" (Modigliani and Hohn 1955, p. 64f). Subsequent work analyzed this issue in more detail and established conditions for planning and forecast horizons (Kunreuther and Morton 1973; Miller 1979), most commonly in the context of dynamic lot-sizing models (Chand 1982; Chand and Morton 1986), which, incidentally, were also the context for early explorations of rolling horizon planning (Lundin and Morton 1975; Denardo and Lee 1991; Stadtler 2000; Van den Heuvel and Wagelmans 2005). The relevant planning horizon may be very small if overtime is unconstrained and relatively cheap compared to holding inventory (Kunreuther 1971).

Applying order release planning models in a rolling horizon setting raises a number of issues that are not easily solved. If in a planning run the initial WIP at a workcenter is higher than planned, this can lead to a technically infeasible solution that must be avoided by some modification of the release model. The analogous situation occurs if an order is delayed to an extent that makes timely completion impossible (Pürgstaller 2009). If backordering is allowed, the model can delay the completion of orders at a cost, but in this case the costs of delaying beyond the planning horizon cannot be modeled, that is, there is an upper limit to the delay costs considered in the model that can distort the release decisions.[1] Nonlinear, convex backordering costs can mitigate or avoid this at the expense of increased model complexity.

Although it may sound trivial, it is important to note that under rolling horizon planning the order releases over time *planned* at a given point in time (the rows in Fig. 10.4) are hardly ever executed, whereas the *actual* releases over time (the shaded elements in Fig. 10.4) are not the result of a single planning problem solved at a specific point in time. This raises several important research issues:

- What is the performance of order release models in a rolling horizon environment, as opposed to the static environment discussed in the previous section that assumes all release decisions made at the start of the planning horizon will be implemented?
- Is there a degree of demand uncertainty that degrades the performance of optimization-based multi-period order release planning to the point that rule-based release mechanisms perform just as well, or even better?
- What is the best way to handle related issues such as executing the plan updates, end-of-horizon effects, frozen horizons, etc.?

[1] The authors thank Stefan Häussler for his contribution to this idea.

Research on these issues is particularly difficult since in a rolling horizon environment planning is performed along two time dimensions: *in* a period *for* the specified planning periods. Hence a great deal depends on the state of the decision maker's information at each decision epoch, i.e., the start of each planning period where a new tentative plan is constructed. Hence a model of demand forecast evolution (Graves et al. 1986, 1998; Heath and Jackson 1994) must be an integral part of the research setting. Finally, the only practical way to evaluate the performance of a planning model in a rolling horizon environment is simulation, which requires the integration of optimization models, forecasting models and simulation in a complex research infrastructure, which is time-consuming both to build and to maintain. However, the fact remains that almost all discrete-time planning models of the type treated in this volume are implemented in practice on a rolling horizon basis, rendering better understanding of this environment crucial to progress in this field.

## 10.3 Integrated Planning of Process Improvement and Production Activities

As discussed in previous chapters, a clearing function can be viewed as a metamodel of a production resource, describing the relation between its output in a planning period and some number of state variables describing the state of the resource in the period. This suggests that the use of clearing functions may provide useful insights into the behavior of production systems subject to learning effects, where the capabilities of the system improve over time due to accumulated knowledge of product and process.

Motivated by applications in semiconductor manufacturing, Kim and Uzsoy (2008b) consider a simplified model of this type considering a single production resource that must manufacture two types of lots: production lots, denoted by $P$ in what follows, that can be used to meet demand, and engineering lots, denoted by $E$, that result in improved system performance after some time lag. The model seeks to determine the number of production and engineering lots to be released into the facility to maximize contribution (profit net of variable costs) in the face of different demand and price patterns. The demand patterns considered include demand increasing over time, as well as an increase followed by a decrease that is a more realistic description of the semiconductor product life cycle. The price scenarios represent the decline in sale price over time extensively documented in the semiconductor industry (Leachman and Ding 2007).

Formulating a model for this problem requires a mathematical model of the learning process. The first step is to specify what determines the amount of learning taking place over an interval of $t$ planning periods—is it cumulative production over the interval, cumulative engineering work (experimentation) over the interval, or some combination of the two? If some combination of the two, how much learning arises from each, and to what degree does one activity substitute for the other?

We must then determine the form of the function, specifying the maximum possible amount of improvement that can be obtained over the entire life of the product with infinite effort, and the rate at which learning will take place. It seems intuitive to postulate a learning model with diminishing returns to the underlying activity, implying that early in the life cycle improvement takes place rapidly but becomes progressively more difficult to achieve. There is an extensive literature on learning models in production systems (Yelle 1979; Anzanello and Fogliatto 2011) upon which this work can draw.

Kim and Uzsoy (2008b) postulate a simple concave, exponential learning function that increases the maximum possible output of the resource based on the total number of engineering lots processed over that period, of the form

$$\phi\left(X^{E}\left(t-d\right)\right)=V_{1}\left(1-e^{-V_{2}X^{E}\left(t-d\right)}\right) \tag{10.1}$$

where $X^E(t-d)$ denotes the total number of engineering lots processed up to production period $t$–$d$, and $d$ is the time lag between an engineering activity taking place and the resulting improvement in capacity being realized. The parameter $V_1$ denotes the maximum additional capacity that can be obtained with infinite engineering activity, while $V_2$ controls the rate at which engineering activity improves capacity, which in practice would be governed by the skill of the engineering group responsible and the complexity of the technical problems encountered. This exponential form implies decreasing returns to scale on engineering activity, where the marginal improvement in capacity from running an additional engineering lot is monotonically decreasing in the number of engineering lots processed. This learning function is then combined with the clearing function of Srinivasan et al. (1988) to obtain a time-dependent clearing function of the form

$$f\left(\sum_{\tau=0}^{t-d}X_{\tau}^{E},W_{t}\right)=\left[K_{1}+\phi\left(\sum_{\tau=0}^{t-d}X_{\tau}^{E}\right)\right]\left(1-e^{-K_{2}W_{t}}\right) \tag{10.2}$$

where $W_t$ denotes the number of production lots in WIP at the end of period $t$. This clearing function assumes that improvement can only be obtained through the processing of engineering lots. The impact of improvement is manifested in the maximum possible output of the resource in a planning period, given by $K_1+\phi(X^E(t-d))$, implying an expected effective processing time of

$$p=\frac{1}{K_{1}+\phi\left(X^{E}\left(t-d\right)\right)} \tag{10.3}$$

It can be shown that when $K_2W_t \geq 0$ and $V_2X_E(t) \geq 0$ the clearing function (10.2) is concave in both $X_E(t)$ and $W_t$. This clearing function can then be embedded in a release planning model that seeks to maximize the total contribution over the planning horizon. A concave clearing function results in a convex optimization problem that can be solved using available commercial solvers. However, analysis of the

Karush–Kuhn–Tucker optimality conditions provides some interesting insights into the marginal value of engineering activity at optimality. Figure 10.5 shows the marginal value of engineering activity over the planning horizon under different price scenarios: level (LV), slowly decreasing (SD), medium decreasing (MD), fast decreasing (FD), and increasing (INC). The INC scenario is unlikely to be encountered in practice but is included for comparison purposes. As would be expected, the value of engineering activity is high early in the planning horizon, and lower in the ending periods. Interestingly, under all price scenarios there is an extended interval in the middle of the planning horizon where the value on engineering activity is essentially constant, except in the INC scenario. As expected, in the INC scenario the marginal value of engineering activity is consistently but slightly higher than in the other price scenarios under demand that increases and then decreases over the product life cycle.

Kim and Uzsoy (2013) extend this approach to a reentrant environment where products must return to the workcenter several times for processing. They again assume that improvements are obtained based on the number of engineering lots of each operation that are processed, assuming that an engineering lot can improve only the specific operation for which it is targeted. In order to ensure the correct behavior of the model in this multiple product environment, engineering lots are assumed to contribute to resource workload. The allocated clearing function formulation of Chap. 5 is modified to incorporate the learning function in the allocation constraints (7.74), as opposed to the clearing function constraints themselves. The model assumes a single product for simplicity of exposition, but can easily be extended to multiple products with the addition of a product index. However, the model in its current form does capture the fact that engineering lots will consume

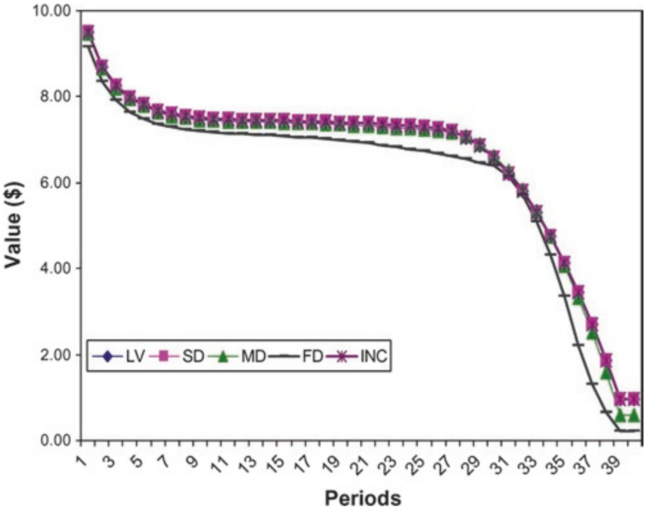


**Fig. 10.5** Marginal value of engineering activity under different price scenarios

capacity at multiple processing steps in addition to those they are specifically aimed at improving, affecting the performance of the production lots.

As in the previous model, the Karush–Kuhn–Tucker optimality conditions can be used to compute the marginal value of engineering activity, i.e., of an additional engineering lot. The authors evaluate these marginal values in a computational experiment with all cost parameters and demand held constant over time and the same engineering delay time and level of engineering skill or technical difficulty for all operations. Operations are indexed in the sequence of their appearance in the routing, so operation 1 precedes operation 2 which, in turn, precedes operation 3. As seen in Fig. 10.6, the marginal benefit of engineering activity decreases sharply for all operations over the planning horizon, which is to be expected due to the diminishing returns on engineering activity implied by the learning function. The plots in Fig. 10.6 differ from those in Fig. 10.5 because demand is constant across the planning horizon in the former, while it is first increasing and then decreasing in the latter. What is more interesting, however, is that the marginal value of engineering activity for downstream operations is consistently higher than that for upstream operations. Essentially the marginal value of engineering activity can be viewed as a surrogate for the marginal value of additional capacity; there is no benefit to increasing the output of an upstream resource if its throughput will simply accumulate as WIP at a downstream operation with less capacity.

Both models discussed above involve considerable simplification of the practical problem, the most obvious being the deterministic formulation without considering

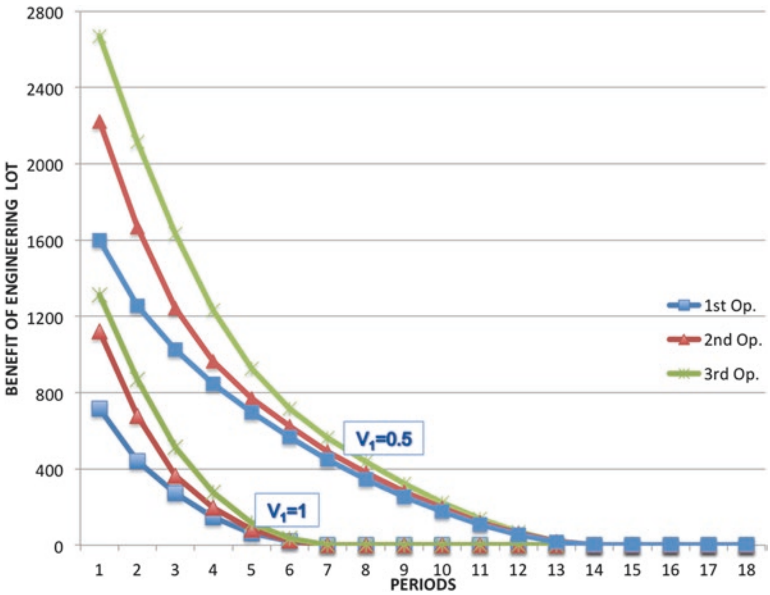


**Fig. 10.6** Marginal benefit of engineering activity for different operations for different levels of possible improvement $V_1$

any of the stochastic elements that are essential parts of the problem. However, the use of the clearing function allows explicit representation of the congestion effects by which engineering and production lots affect the performance of production system, and hence each other's cycle times and throughput. Neither of the two learning mechanisms is completely satisfactory, since in both models the learning shifts the clearing function upward across its entire range. This corresponds to reducing the average effective processing time, but does not capture the effects of learning on its variance. A substantially more complex clearing function has been suggested by Manda et al. (2016) for the case of product transitions, where a stable product is gradually replaced by a newer product whose processing is subject to higher variability, which decreases over time as learning takes place. These authors also consider a stochastic production process using simulation optimization to optimize the releases of the two products over time (Manda and Uzsoy 2018). While a nonlinear programming model similar in spirit to those of Kim and Uzsoy can be formulated, the incorporation of learning effects into the variance of the effective processing time results in non-convex models that are difficult to solve. However, it is clear that the use of clearing functions that can accurately reflect the impacts of product mix and engineering activity on the mean and variance of the effective processing time at production resources offers a useful tool to gain insight into the behavior of these production system.

## 10.4 Dynamic Pricing Under Price and Lead Time Sensitive Demand

In many capital-intensive industries demand can vary significantly over time, suggesting the use of dynamic pricing to shape demand by means of price promotions. However, aggressive use of price promotions raises the possibility of stimulating excessive demand, which will result in higher than expected resource utilization, increased cycle times, and hence missed delivery dates with the associated loss of customer goodwill and future business. This need has led to a body of research on planning models that consider both lead times and pricing (Upasani and Uzsoy 2008). Upasani and Uzsoy (2014) have approached this problem using clearing functions, providing some interesting insights and highlighting the issues that can arise when congestion is not considered. They assume that the firm behaves as a monopolist and faces a demand function

$$g\left(P,L\right)=\max\left\{0,M-aP-bL\right\} \tag{10.4}$$

where $M$ represents the maximum possible demand the market can support, $P$ the price of the product, and $L$ the lead time, respectively. In a given period $t$, the firm quotes a price $P_t$ and a delivery time $L_t$ to customers equal to the average manufacturing lead time at the start of the period. Since the manufacturing lead time (delivery time) depends on the number of orders waiting, the firm can control the maximum delivery time by limiting the number of orders accepted. In effect, the firm quotes the delivery time based on the minimum of two values: the average manufacturing lead time, and a guaranteed delivery time $L_G$ by which all orders need to be satisfied, or customers will not place orders. Hence an order received in period $t$ has to be fulfilled by period $t + L_G$. The firm needs to align its production system with this market preference by quoting an average delivery time below the value of $L_G$. This will, in turn, determine the number of orders that a firm may accept and hold in queue for processing, yielding a target production rate and a target utilization. Thus, the higher the guaranteed delivery time allowed by the market, the higher the utilization at which the firm can operate its resources. The firm may also control the quoted average delivery time by quoting a higher price, and thus accepting fewer orders. Customers may be willing to pay a higher price for lower quoted delivery times, depending on their sensitivity to lead time expressed by the value of $b$ in the demand function. The quotation of an average delivery lead time implies that some orders may be delivered earlier than promised, and the authors assume an upper limit $\nu$ on the number of orders delivered early, as well as a unit size for all orders, and they assume a clearing function of the form suggested by Karmarkar (1989). The following notation is used in the optimization model:

**Decision variables:**
$R_t$: order release quantity in period $t$
$W_t$: work-in-process (WIP) inventory at the end of period $t$
$X_t$: production quantity in period $t$
$I_t$: finished goods inventory (FGI) at end of period $t$
$P_t$: price in period $t$
$D_t$: sales quantity in period $t$
$Y_t$: quantity shipped in period $t$

**Parameters:**
$a_t$: price sensitivity of demand in period $t$
$b_t$: lead time sensitivity of demand in period t
$h_t$: finished inventory holding cost in period $t$
$w_t$: WIP holding cost in period $t$
$\phi_t$: unit production cost in period $t$
$c_t$: order release cost per unit released in period $t$
$\nu$: maximum units allowed to be shipped before due date over the horizon
$K_1$, $K_2$: parameters of the Karmarkar clearing function
$M$: intercept of the demand function
$T$: length of planning horizon
$L_G$: guaranteed delivery time (in periods)
$f(.)$: clearing function

Once again the average WIP $\hat{W}_t$ is used as the argument of the clearing function. By Little's Law, the expected lead time in period $t$ is given by $L_t = \hat{W}_t / X_t$ expressed in units of periods. Thus, the demand observed in period $t$ is given by:

$$D_t = M - a_t P_t - b_t \left( \frac{\hat{W}_t}{X_t} \right) \tag{10.5}$$

The joint price-production planning model can now be stated as follows:

$$\max \sum_{t=1}^{T} \left[ P_t \left( M - a_t P_t - b_t \left( \frac{\hat{W}_t}{X_t} \right) \right) - c_t R_t - \phi_t X_t - h_t I_t - \omega_t W_t \right] \tag{10.6}$$

subject to:

$$W_t = W_{t-1} - X_t + R_t, \quad t = 1,\ldots,T \tag{10.7}$$

$$I_t = I_{t-1} + X_t - Y_t, \quad t = 1,\ldots,T \tag{10.8}$$

$$X_t \le \frac{K_1 \hat{W}_t}{K_2 + \hat{W}}, \quad t = 1,\ldots,T \tag{10.9}$$

$$M - a_t P_t - b_t \left( \frac{\hat{W}_t}{X_t} \right) \ge 0, \quad t = 1,\ldots,T \tag{10.10}$$

$$\sum_{\tau=1}^{t} Y_\tau \ge \sum_{\tau=1}^{t-L_G} \left[ M - a_\tau p_\tau - b_\tau \left( \frac{\hat{W}_\tau}{X_\tau} \right) \right] \tag{10.11}$$

$$\sum_{\tau=1}^{t} Y_\tau \le \sum_{\tau=1}^{t-L_G} \left[ M - a_\tau p_\tau - b_\tau \left( \frac{\hat{W}_\tau}{X_\tau} \right) \right] + v \tag{10.12}$$

$$\hat{W}_t \le \frac{W_t + W_{t-1}}{2}, \quad t = 1,\ldots,T \tag{10.13}$$

$$R_t, P_t, X_t, W_t, I_t \ge 0, \quad t = 1,\ldots,T \tag{10.14}$$

The objective is to maximize total contribution, expressed as the difference between the total revenue in each period and variable operating costs, (10.7) and (10.8) are WIP and finished goods inventory balance constraints, (10.9) represents production capacity using the CF, and (10.10) defines the sales quantity. Constraint (10.11) requires that all orders be shipped within the planned delivery time, but allows orders to be shipped earlier than due, rather than being held as finished goods inventory. Since the customer may impose a limit on the number of orders shipped early over the horizon (given by the parameter $\nu$), we model this preference in constraint (10.12) and (10.13) defines the average WIP level $\hat{W}_t$ within a given period. All variables are required to be non-negative by (10.14). No cost is imposed on shipping for parsimony in the experimental design. For the same reason, there is no penalty if the average delivery time quotation exceeds the planned delivery time, but sales are reduced by the operation of the demand function.

The authors compare the performance of this model to that of a joint price-production planning model, referred to as the FLT model, that assumes a fixed delivery time $L \leq L_G$ which is specified as an exogenous parameter. The demand observed by this model in period t is thus given by:

$$D_t = M - a_t P_t - b_t L \tag{10.15}$$

The authors conduct extensive numerical experiments as well as analyze the structure of optimal solutions to both models using the Karush–Kuhn–Tucker optimality conditions. Figure 10.7 compares the planned and realized sales for the two models, assuming that the actual behavior of the production system is governed by the clearing function and there is no delivery flexibility, i.e., $\upsilon = 0$. Realized sales are computed for both models assuming that the CF represents the actual capability of the system. While the CF model is able to deliver the planned sales quantities, the FLT plans for significantly higher sales but is unable to deliver them on time because it ignores the congestion effects reflected by the CF. The CF model sets prices slightly, but not drastically, higher than the FLT model across the horizon. The FLT model, however, assumes that all demand within the maximum capacity of the system can be met within the planned lead time of one period. However, the combination of low price and low quoted lead time creates high demand that cannot be met within the planned lead time, resulting in a substantial shortfall in delivery in the later periods. Figure 10.8 shows the planned and average cycle times for the two models. The FLT model plans using a constant planned lead time, but the realized average cycle time increases throughout most of the planning horizon as the system cannot meet the high demand it has generated. The CF model, in contrast, maintains a relatively smooth sales profile by varying the planned cycle time over the horizon.

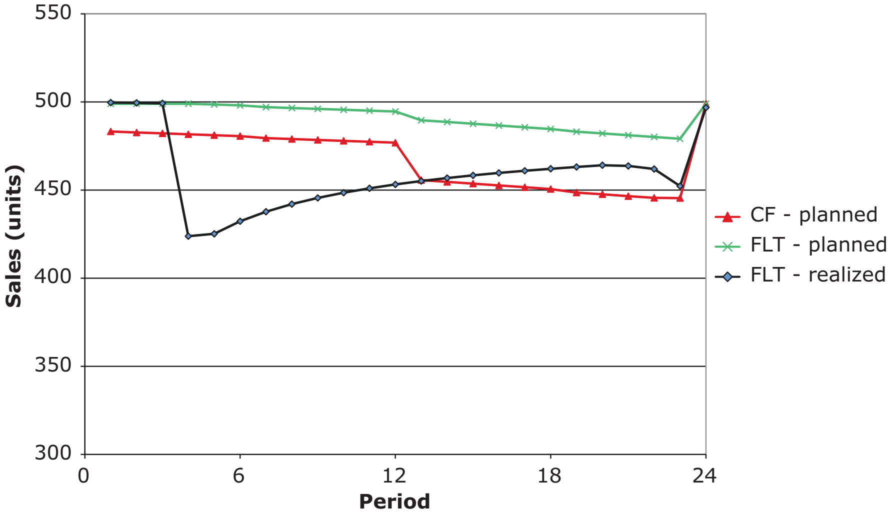


**Fig. 10.7** Comparison of sales with $u = 0.8$ and $\upsilon = 0$, $L = L_G = 1$

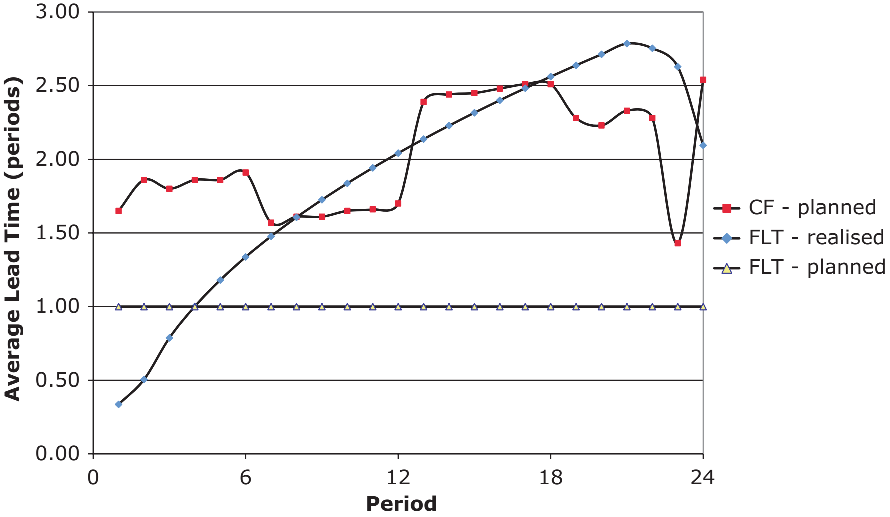


**Fig. 10.8** Lead times comparison with $u = 0.9$, $L = 2$, $L_G = 3$

The manner in which the maximum permissible loading, i.e., the planned capacity, is specified in the FLT model will clearly influence the results; in these experiments, this parameter was set to the maximum possible output obtainable from the clearing function, given by the parameter $K_1$ in (10.9). More extensive results and an in-depth discussion of the numerical experiments can be found in the original paper.

While this model is, once again, a great simplification of the practical problem, the use of the clearing function to represent the behavior of congested production resources results in qualitatively different results from those obtained using fixed planned lead times. The complexity of pricing models of this type rapidly increases; in particular, many models of this kind result in non-convex optimization models where often the best that can be done, as is the case in the models reported above, is to examine locally optimal solutions.

## 10.5 Discussion

The applications discussed in this chapter are illustrative of the types of problems that can be addressed using the clearing function construct. Sections 10.1 and 10.2 compare the performance of release planning models using clearing functions to those using fixed lead times under both static and rolling horizon conditions. The results are promising but not definitive; in both sections we find that the clearing function models perform well under some conditions and less so under others. A number of factors may have affected the results, including some aspects of the experimental design and the way in which the clearing functions were estimated. Of particular interest are the findings related to the performance of fixed lead time models; while integer lead times appear to lead to poor performance, fractional lead time models perform well even though experiments find that the observed cycle times at some resources can deviate considerably from the estimates over the planning horizon. While our understanding of how to best estimate clearing functions is highly unsatisfactory, it is interesting to observe that even though fixed lead time models have been in use for a long time, our understanding of how to specify the planned lead times they use is also far from definitive. We are not aware of any systematic experimental study of how the performance of fixed lead time release planning models is affected by the planned lead time estimates used, nor of a body of theory addressing this interesting question. A systematic exploration of this area offers some intriguing prospects for future research.

The models of engineering improvements and dynamic pricing, on the other hand, both illustrate how the use of clearing functions can lead to interesting models that provide considerable insight. In the context of engineering improvement, the use of clearing functions provides a much more complete picture of the adverse effects of engineering work on the regular production activities of the production unit. The increased uncertainty associated with the duration and occurrence of engineering activities results in increased mean and variance of the effective processing time distribution in the production unit, better capturing the externalities imposed on production by engineering work. The dynamic pricing model, on the other hand, illustrates the difficulties encountered when customer lead times affect demand, which must be met from a production unit whose resources exhibit queueing behavior. The clearing function pricing model allows the quotation of lead time and price combinations that allow demand to be met in a timely manner, as opposed to the fixed lead time case which may allow unrealistically low prices to create congestion and extend cycle times beyond what the customer can tolerate.

Other potential areas to which clearing functions can be applied include capacity expansion, explored by Kim and Uzsoy (2008a), modelling patient flow in service systems such as hospitals, and integrated planning models that address stochastic demand by planning releases and safety stocks in an integrated manner (Orcun et al. 2009). However, we need to remember that the examples in this chapter raise as many questions as they answer, suggesting an interesting and fruitful research agenda for the future.

## References

Albey E, Uzsoy R (2015) Lead time modeling in production planning. In: Winter simulation conference, IEEE, Huntington Beach

Albey E, Uzsoy R (2016) A chance constraint based multi-item production planning model using simulation optimization. In: Winter simulation conference, Arlington

Albey E, Norouzi A, Kempf KG, Uzsoy R (2015) Demand modeling with forecast evolution: an application to production planning. IEEE Trans Semicond Manuf 28(3):374–384

Anzanello MJ, Fogliatto FS (2011) Learning curve models and applications: literature review and research directions. Int J Ind Ergon 41:573–583

Aouam T, Uzsoy R (2012) An Exploratory Analysis of Production Planning in the Face of Stochastic Demand and Workload-Dependent Lead Times. Decision Policies for Production Networks. K. G. Kempf and D. Armbruster. Boston, Springer: 173–208

Aouam T, Uzsoy R (2015) Zero-order production planning models with stochastic demand and workload-dependent lead times. Int J Prod Res 53(6):1–19

Asmundsson J, Rardin RL, Uzsoy R (2006) Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Trans Semicond Manuf 19(1):95–111

Asmundsson J, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning with resources subject to congestion. Nav Res Logist 56(2):142–157

Baker KR (1977) An experimental study of the effectiveness of rolling schedules in production planning. Decis Sci 8:19–27

Bertsimas D, Sim M (2004) The price of robustness. Oper Res 52(1):35–53

Birge JR, Louveaux F (1997) Introduction to stochastic programming. Springer, New York

Blackburn JD, Millen RA (1980) Heuristic lot-sizing performance in a rolling-schedule environment. Decis Sci 11(4):691–701

Blackburn JD, Kropp DH, Millen RA (1985) MRP system nervousness: causes and cures. Eng Costs Prod Econ 9(1–3):141–146

Chand S (1982) A note on dynamic lot sizing in a rolling-horizon environment. Decis Sci 13(1):113–119

Chand S, Morton TE (1986) Minimal forecast horizon procedures for dynamic lot size models. Nav Res Logist 33:111–122

Denardo EV, Lee C-Y (1991) Error bound for the dynamic lot size model with backlogging. Ann Oper Res 28(1):213–227

Dolgui A, Prodhon C (2007) Supply planning under uncertainties in MRP environments: a state of the art. Annu Rev Control 31:269–279

Dolgui A, Ben Ammar O, Hnaien F, Louly MA (2013) A state of the art on supply planning and inventory control under lead time uncertainty. Stud Inf Control 22(3):255–268

Fowler J, Robinson J (2012). https://www.sim.uni-hannover.de/~Svs/Wise0809/Pds/Masmlab/Factory_Data-Sets/

Fu MC (ed) (2015) Handbook of simulation optimization. Springer, New York

Gopalswamy K, Uzsoy R (2019) A data-driven iterative refinement approach for estimating clearing functions from simulation models of production systems. Int J Prod Res 57(19):6013–6030

Graves SC, Meal H, Dasu S, Qui Y (1986) Two-stage production planning in a dynamic environment. In: Axsater S, Schneeweiss C, Silver E (eds) Multi-stage inventory planning and control. Springer, Berlin

Graves SC, Kletter DB, Hetzel WB (1998) Dynamic model for requirements planning with application to supply chain optimization. Oper Res 46(3):35–49

Häussler S, Missbauer H (2019) Comparison of two optimization-based order release models with fixed and variable lead times. Department of Information Systems, Production and Logistics Management, University of Innsbruck, Innsbruck

Heath DC, Jackson PL (1994) Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. IIE Trans 26(3):17–30

Holt CC, Modigliani F, Muth JF, Simon HA (1960) Planning production, inventories and work force. Prentice Hall, Englewood Cliffs

Hung YF, Leachman RC (1996) A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. IEEE Trans Semicond Manuf 9(2):257–269

Kacar NB, Uzsoy R (2015) Estimating clearing functions for production resources using simulation optimization. IEEE Trans Autom Sci Eng 12(2):539–552

Kacar NB, Moench L, Uzsoy R (2013) Planning wafer starts using nonlinear clearing functions: a large-scale experiment. IEEE Trans Semicond Manuf 26(4):602–612

Kacar NB, Moench L, Uzsoy R (2016) Modelling cycle times in production planning models for wafer fabrication. IEEE Trans Semicond Manuf 29(2):153–167

Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123

Kayton D, Teyner T, Schwartz C, Uzsoy R (1997) Focusing maintenance improvement efforts in a wafer fabrication facility operating under theory of constraints. Prod Invent Manag 38(Fourth Quarter):51–57

Kim S, Uzsoy R (2008a) Exact and approximate algorithms for capacity expansion problems with congestion. IIE Trans Schedul Logist 40(12):1185–1197

Kim S, Uzsoy RM (2008b) Integrated planning of production and engineering process improvement. IEEE Trans Semicond Manuf 21(3):390–398

Kim S, Uzsoy R (2013) Modeling and analysis of integrated planning of production and engineering process improvement. IEEE Trans Semicond Manuf 26(3):414–422

Kunreuther H (1971) Production-Planning Algorithms for the Inventory-Overtime Tradeoff. Operations Research 19(7):1717–1729

Kunreuther HC, Morton TE (1973) Planning Horizons for Production Smoothing with Deterministic Demands. Management Science 20(1):110–125

Leachman RC, Ding S (2007) Integration of speed economics into decision-making for manufacturing management. Int J Prod Econ 107:39–55

Lin PC, Uzsoy R (2016a) Chance-constrained formulations in rolling horizon production planning: an experimental study. Int J Prod Res 54(13):3927–3942

Lin PC, Uzsoy R (2016b) Estimating the costs of planned changes implied by freezing production plans. In: Rabadi G (ed) Heuristics, metaheuristics and approximate methods in planning and scheduling. Springer, New York, pp 17–44

Lundin RA, Morton TE (1975) Planning horizons for the dynamic lot size model: zabel vs. protective procedures and computational results. Oper Res 23(4):711–734

Manda, A. B. and R. Uzsoy (2018). Simulation optimization for planning product transitions in semiconductor manufacturing facilities. In: Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B. Wither simulation conference. IEEE, Gothenburg

Manda AB, Uzsoy R, Kempf KG, Kim S (2016) Modeling the impact of new product introduction on the output of semiconductor wafer fabrication facilities. In: Winter simulation conference, Arlington

Miller LW (1979) Using Linear Programming to Derive Planning Horizons for a Production Smoothing Problem. Management Science 25(12):1232–1244

Modigliani F, Hohn FE (1955) Production Planning over Time and the Nature of the Expectation and Planning Horizon. Econometrica 23(1):46–66

Mula J, Poler R, Garcia-Sabater JP, Lario FC (2006) Models for production planning under uncertainty: a review. Int J Prod Econ 103:271–285

Norouzi, A. (2013). The effect of forecast evolution on production planning with resources subject to congestion. PhD, E. P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh

Norouzi A, Uzsoy R (2014) Modeling the evolution of dependency between demands, with application to production planning. IIE Trans 46(1):55–66

Orcun S, Uzsoy R, Kempf KG (2009) An integrated production planning model with load-dependent Lead-times and safety stocks. Comput Chem Eng 33(12):2159–2163

Pürgstaller P (2009) Ein Vergleich von Regel- und Optimierungsbasierten Bestandsregelungskonzepten bei Kundenauftragsfertigung. PhD, University of Innsbruck, Innsbruck

Pürgstaller P, Missbauer H (2012) Rule-based vs. optimization-based order release in workload control: a simulation study of an MTO manufacturer. Int J Prod Econ 140:670–680

Sahin F, Narayanan A, Robinson EP (2013) Rolling horizon planning in supply chains: review, implications and directions for future research. Int J Prod Res 51(18):5413–5436

Srinivasan A, Carey M, Morton TE (1988) Resource pricing and aggregate scheduling in manufacturing systems. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh

Stadtler H (2000) Improved rolling schedules for the dynamic single-level lot sizing problem. Manag Sci 46:318–326

Turkseven CH (2005) Computational evaluation of production planning formulations using clearing functions. School of Industrial Engineering, Purdue University, West Lafayette

Upasani A, Uzsoy R (2008) Incorporating manufacturing lead times in joint production-marketing models: a review and further directions. Ann Oper Res 161:171–188

Upasani A, Uzsoy R (2014) Integrated production planning and pricing decisions in congestion-prone capacitated production systems. In: Pulat SP, Sarin SC, Uzsoy R (eds) Essays in planning, scheduling and optimization: a festschrift in honor of Prof. S. E. Elmaghraby. Springer, New York

Van den Heuvel W, Wagelmans APM (2005) A comparison of methods for lot sizing in a rolling horizon environment. Oper Res Lett 33:486–496

Wijngaard J (2004) The effect of foreknowledge of demand in case of a restricted capacity: the single-stage, single-product case. Eur J Oper Res 159:95–109

Wijngaard J, Karaesmen F (2007) Advance Demand Information and a Restricted Production Capacity: On the Optimality of Order Base-Stock Policies. OR Spectrum 29(4):643–660

Yelle LE (1979) The learning curve: historical review and comprehensive survey. Decis Sci 10:302–329

Ziarnetzky T, Moench L, Uzsoy R (2018) Rolling horizon, multiproduct production planning with chance constraints and forecast evolution for wafer fabs. Int J Prod Res 56(18):6112–6134

# Chapter 11
# Conclusions and Future Directions

Problems arising in the PPC systems that support the complex global supply chains driving the modern economy were among the earliest to be addressed with the tools of operations research (Arrow et al. 1958; Hanssmann 1959; Holt et al. 1960; Buffa and Taubert 1972; Johnson and Montgomery 1974), leading to a broad, mature body of knowledge using a variety of mathematical formalisms including mathematical programming, queueing, simulation, and stochastic optimization. By the nature of this problem domain, this volume has ranged widely over a great deal of ground, and we hope that the reader has found the journey worthwhile. This chapter concludes the book with a brief review of the principal results and their implications for future work, both related to the clearing functions that are the central concern of this volume and for the broader field of production planning models.

## 11.1 The Gordian Knot: Output, Cycle Time, and Workload

The problem at the heart of this volume is the intimate interconnection between the output, cycle time, and workload of a production unit, with, of course, the individual production resource as a special case. Whether the problem faced is that of coordinating a number of production units across a supply chain, or planning the releases of work into an individual production unit to meet demand in the best possible manner, PPC systems simply cannot operate effectively without some cognizance of the impact of their decisions on cycle times. As discussed in Chap. 2, queueing models, simulation experiments, and industrial observation all indicate that the cycle time of an order through a production unit is a random variable whose distribution depends, among potentially many other things, on the utilization of the resource, i.e., the workload available for it to process that is determined by the work release decisions made by any PPC system discussed in Chap. 1. Hence cycle times should be considered as endogenous to the planning process, rather than as an exogenous parameter, which is manifestly not the case in most of the production planning literature that

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_11

goes by that name. Most of this literature can, with more or less of a stretch, be positioned within one of the two principal frameworks for PPC systems that have emerged as a (admittedly evolving) consensus between industrial practice and academic research: Manufacturing Planning and Control (Jacobs et al. 2011) and Advanced Planning and Scheduling (APS) (Stadtler et al. 2015). The discussion of these frameworks in Chap. 3 highlights the importance of cycle times to the effective operation of both. This circularity—that planning systems need to be cognizant of cycle times, but cycle times are a consequence of the work release decisions made by the planning systems themselves—has, in our opinion, constituted a significant barrier to progress. The contents of this volume can thus be viewed as a series of attempts to address this difficulty or mitigate its negative impacts.

Chapter 4 discusses the workload control (WLC) paradigm, which constitutes a first-order response to the relation between workload, output, and cycle time. Despite their wide variety, all WLC approaches seek to identify a workload level for the production unit that will yield an acceptable compromise between the goals of maintaining low WIP and cycle times on the one hand, and sufficient output to meet demand on the other. Most such systems are rule-based, designed to operate in an environment where the demand distribution faced by the production unit remains approximately constant; they do not easily adapt to changing operating conditions, which would require recalculation of their various parameters as the environment changes. It is probably fair to say that there is as yet no unified theory governing the relations between the environmental conditions faced by such WLC systems and the values of the various parameters they require. Only a few of these approaches use an explicit model of material flow through the production unit to inform their work release decisions. The optimization models discussed in the subsequent chapters can be viewed as natural extensions of these model-based WLC approaches.

Most existing approaches to production planning, from the material requirements planning (MRP) procedure widely used in industry (Orlicky 1975; Baker 1993; Jacobs et al. 2011) to the mathematical programming models that form the central engine of many advanced planning and scheduling (APS) systems (Voss and Woodruff 2006; Hackman 2008), approach this issue using planned lead times that are treated as exogenous, workload-independent parameters. As long as lot-sizing or capacity expansion decisions are not considered, avoiding the need for integer variables, these models can generally be formulated as linear programs that can be solved with existing commercial solvers, even for very large problem instances. One of us (RU) had the opportunity a decade ago to observe the implementation of a new planning system at a major high-technology manufacturer. The complete workflow for generating a plan for a significant portion of the supply chain, involving multiple plants, multiple production lines within plants and distribution facilities, took approximately 24 h at that time, of which only 45 min was required for the solution of the optimization model. The remaining time was taken up by acquiring, formatting, and cleaning input data from the firm's ERP system and then transferring the output of the planning model back to the ERP system for execution. Chapter 5 summarizes the state of the art in these models when the planned lead time remains constant over time.

There is considerable evidence, including our own presented in Chap. 10, that despite its evident inconsistency with queueing theory the use of fixed planned lead times frequently does not lead to unacceptably bad performance. One reason for this may be that many facilities are operated within a relatively narrow range of operating conditions as defined by product mix, available resources, and demand, allowing planned lead times that provide good shop-floor performance to be arrived at over time. There is also often considerable opportunity for shop-floor decisions to mitigate the negative effects of suboptimal work release decisions by scheduling overtime, exploiting alternative resources, expediting and other such measures.

An additional advantage of planning models based on exogenous planned lead times is their intuitive nature. The idea of a delay between the release of work and its emergence as finished product is an easy one to grasp, making acceptance of the resulting planning models by their ultimate users, the managers responsible for the performance of the production units making up the supply chain, much easier than for a complex, nonlinear mathematical model. This does not mean, however, that the decisions obtained from a complex optimization model are always intuitive; anyone who has tried to explain to a manager why the optimization model chose to produce a specific amount of a specific item at a specific time on a specific resource, instead of using one of the many available alternatives (usually including the manager's favorite), will recognize the difficulty in parsing the output of a large mathematical program into a narrative explanation. The work of Greenberg (1996) on a rule-based system for explaining the results of linear programming models suggests an interesting direction for future research customizing this generic approach to specific production planning formulations.

The planning models in Chap. 5 can be viewed as optimizing work releases for a given set of planned lead times. The endogeneity of cycle times to work release decisions discussed in Chap. 2 suggests a model that can jointly optimize releases and cycle times simultaneously. Thus, if we could find the "correct" planned lead times for each planning period, the models of Chap. 5 would provide the optimal releases directly. Chapter 6 explores the difficulties that arise in identifying a consistent set of planned lead times across the planning horizon, and then focuses on approaches that decompose the planning problem into two subproblems. The first of these takes estimates of planned lead times as input and computes optimal releases based on these lead times. The second model takes a set of releases as input, and returns estimates of the resulting cycle times from which revised planned lead times can be computed. The release planning model is usually a linear program similar to those described in Chap. 5, while the lead time estimation model is usually a more or less detailed simulation model of the production unit of interest, although queueing and statistical models can also be used. A variety of such models have been proposed since the initial work of Hung and Leachman (1996), none of which have yielded conclusively positive results. Their computational burden tends to be high due to the need for multiple replications of a (often large) simulation model at each iteration. Their convergence behavior is not well understood; there appears to be no theoretical guarantee of their convergence, and experimental observations include cycling between solutions, failure to converge in any recognizable way, and dependence on the starting solution. The prime advantage of these approaches is that they combine two techniques, mathematical programming and simulation, that are each familiar to practitioners and have access to excellent commercial software. However, this approach does not build an explicit model linking output, workload, and cycle time; this model is implicit in the dynamics of the simulation or queueing model used to estimate the planned lead times given a set of releases.

Chapter 7 introduces univariate clearing functions that formulate a mathematical relation linking the expected output of a production resource in a planning period to some measure of its workload in the planning period. The basic concept was introduced, apparently independently, by several researchers in the late 1980s (Graves 1986; Srinivasan et al. 1988; Karmarkar 1989). Univariate clearing functions that are concave in their measure of workload (whatever that may be) yield convex optimization models. Since a concave function can be approximated to any degree of accuracy by a set of linear functions, it is easy to approximate these as linear programs, although the growing computational power of convex nonlinear solvers renders this less important than it once was. However, serious difficulties arise when multiple products competing for capacity on the same resource are considered, and straightforward extension of the single-product models results in clearly anomalous behavior. These difficulties are closely related to those ably explored by Carey and his coauthors in the domain of traffic modeling (Carey 1987, 1990; Carey and Subrahmanian 2000; Carey and Bowers 2012) and discussed in Chap. 6 in the context of time-varying planned lead times. After illustrating the behavior of the clearing function as a representation of a production unit, this chapter presents the allocated clearing function model of Asmundsson et al. (2006, 2009), which provides an effective although approximate solution to these difficulties and remains the state of the art at this time of writing. The chapter also illustrates one of the primary theoretical advantages of the clearing function approach over the models of Chap. 5, its ability to provide richer dual information on the marginal price of capacity at the different resources in the production unit.

The development of the allocated clearing function model exposes the limitations of the use of univariate clearing functions. The univariate clearing function estimates the aggregate output of the production unit across all products as a function of the aggregate workload of all products and then, as its name implies, allocates this aggregate output optimally among the different products. Chapter 8 departs from the observation that the allocated clearing function approach fails quite badly when the aggregate output depends heavily on the mix of products, not just on the aggregate workload. This is clearly the case when there are significant setup times between different products on the production resources; lot-sizing and sequencing decisions now have major impact on output. This chapter examines efforts to formulate multivariate clearing functions, raising the question of what additional state variables should be included. A variety of such state variables have been tried, including decomposing the workload of a product into the WIP available at the start of the period and releases during the period; inclusion of state variables related to previous periods; and using the output of each product as a state variable describing the output of all others. Many of these efforts result in non-convex optimization models, although computational evidence suggests that convex solvers can often obtain global optimal solutions in many cases, suggesting the presence of considerable structure that remains to be uncovered. Computational results, however, indicate that considerable improvement over univariate clearing functions can be obtained, at the cost of additional computational burden. It is probably safe to classify much of the work in this chapter as exploratory, leaving considerable room for future research.

Chapter 9 briefly explores the relation between the clearing function concept and lot-sizing decisions in the context of a single production resource. The seminal work of Karmarkar and his coworkers (Karmarkar 1987; Karmarkar et al. 1992) used queueing models to illustrate the relation between lot-sizing decisions and cycle times, which can then be used to derive multivariate clearing functions in which the output of a product depends on the lot sizes and output of all products in the system. The chapter then develops a non-convex optimization model using multivariate clearing functions for a single-machine dynamic lot-sizing problem and shows that this can yield significant performance improvements over prior approaches that do not consider queueing behavior. The chapter closes with an admittedly heuristic discussion using this model to illustrate the difficulty of accurately estimating the setup costs that are a crucial parameter of most lot-sizing models in the literature, which focus on the tradeoff between setup and cycle stock holding costs.

Having presented the clearing function concept in various forms in Chaps. 7 through 9, Chap. 10 examines several applications of the concept. A series of computational experiments using the allocated clearing function model for release planning for semiconductor wafer fabrication yield admittedly mixed results. While the clearing function model outperforms fixed lead time models with integer lead times, the use of fractional lead times largely eliminates the advantage of the clearing functions except under time-varying demand. Other applications include the use of clearing function models in a rolling horizon context, where they largely retain their advantage over fixed lead time models, the integrated planning of production and improvement activities, and dynamic pricing in an environment where demand is sensitive to both lead time and price. By and large, the results of the clearing function approaches are promising, especially when the richer dual information they yield can be used to gain insight into system behavior.

## 11.2 Weaknesses and Limitations of the Clearing Function Approach

Having laid out in the preceding chapters the basic motivation for the clearing function approach and the state of our knowledge to date, we would be remiss if we implied that we have a watertight case; we most certainly do not. The perceptive reader will have raised a number of criticisms themselves by this point in the volume, and there are many such both stated and implied in the previous pages. In this section, we will discuss several of the most important of these difficulties, some of which are the subject of ongoing research while others await the attention of the research community.

### *11.2.1 Why Clearing Functions?*

The first question that needs to be addressed is simply that of why the clearing function construct should be used at all. Chapter 7 pointed out that a clearing function is a metamodel describing certain aspects of the behavior of a queue. If this is indeed the case, there are a wide range of alternative approaches to choose from, such as other forms of metamodeling (Li et al. 2016), system dynamics (Sterman 2000), transient queueing models (Askin and Hanumantha 2018) and, of course, simulation (Law and Kelton 2004). Given that the reason for using clearing functions is not at all obvious, and at least one reviewer of our work has stated categorically that "… the clearing function idea is outdated," some discussion of this issue appears to be necessary.

The primary reason for using a clearing function to represent a production unit is the ability to embed it in a tractable optimization model to plan releases for the next several periods. For this purpose, what is required is a sufficiently accurate representation of the relation between workload and output; six decimal places of precision are not required in a planning model whose purpose is simply to ensure that the workload in the production unit is at the correct level to meet the desired output without unnecessarily increasing WIP and cycle time.

The other types of model described above do not lend themselves easily to the formulation of tractable mathematical programming models. It is certainly true that queueing or simulation models can be embedded in an optimization framework, using algorithms similar to those used for simulation optimization (Fu 2015). A number of models of this latter type have been presented in the literature, notably the metamodel-based simulation optimization algorithm of Li et al. (2016) and the simulation optimization approaches of Kacar and Uzsoy (2015). These models benefit from the superior ability of simulation models to incorporate detailed system dynamics that are difficult to capture in a clearing function. However, the development of the metamodel requires extensive simulation experiments to collect data and fit the metamodel, while simulation optimization is very time-consuming. The use of the clearing function construct is aimed at enabling the use of a mathematical programming model to optimize releases as well as providing the information from the dual solution that may help management better understand the behavior of their system. It is very unlikely that a clearing function can provide a highly precise prediction of output in each period, but that is not its purpose; it seeks to provide sufficiently accurate descriptions of system behavior to ensure that the planning model using it maintains the system workload in a state that will sustain the desired output level.

The advantage of a mathematical programming model, in turn, is that it allows rapid solution of a complex optimization problem, especially when it can be formulated as a linear program. The estimation of the clearing function used in the model will require considerable computational effort, but this work can be performed offline and is not part of the run generating the planning solution. In contrast, any model utilizing a detailed simulation of the production unit of interest, whether a full-blown simulation optimization approach or one of the iterative multi-model approaches discussed in Chap. 6, requires multiple replications of simulation runs for the actual solution of the release planning model, requiring considerably more time.

### *11.2.2 Choice of Functional Form for the Clearing Function*

While the idea of a concave non-decreasing functional relation between the workload and the expected output of a production resource is quite intuitive, it should be evident to the reader by this point that the state of our knowledge as to what functional forms to use and how to estimate their parameters from either industrial or simulation data is as yet highly unsatisfactory. The derivation of clearing functions from steady-state queueing models is inherently dangerous in a discrete-time planning model unless planning periods are long enough for the underlying queues to at least approximately reach steady state, which is frequently not the case in practice. The early work of Asmundsson et al. (2009) revealed that using conventional least-squares regression to fit one of the empirical functional forms discussed in Chap. 7 results in systematic overestimation of the expected output, due to the relation captured by Jensen's inequality (8.8) discussed in Sect. 8.2. Gopalswamy and Uzsoy (2019) identify a number of additional issues arising in the fitting of clearing functions to simulation data, and the design of appropriate simulation experiments to obtain such data.

Even if the issues associated with estimating clearing functions of a tractable computational form were addressed satisfactorily, the currently common approach of fitting a single clearing function that is expected to represent the behavior of the production resource in all planning periods is clearly a significant approximation, as discussed in Sect. 8.2 and illustrated in Fig. 8.2. The simulation optimization approach of Kacar and Uzsoy (2015) found that fitting a clearing function to each planning period gave superior results to using a single clearing function for all periods. However, this simulation optimization approach is computationally demanding, especially when used in a rolling horizon environment.

Yet another difficulty with the use of clearing functions arises in multistage environments. Expression (2.1) shows that the expected cycle time at a given resource depends on the mean and variance of both interarrival and service times; the variability of the interarrival times in turn depends on decisions made at upstream resources. Thus, at least in theory, the shape of the clearing function at a given resource is affected by the production decisions at upstream resources. The clearing function models we have discussed in this volume do not take this type of interrelation between production resources, or production units, into account. Instead they assume that the mean and variance of interarrival and service times at each resource are independent of other resources, an assumption referred to in queueing theory as decomposition. Clearly some error is introduced into the models by this approach. One would expect this to become especially serious in multistage systems with setup times at each resource, where lot-sizing decisions at each stage in each period may affect the shape of clearing functions at downstream resources.

The situation for MDCFs is still more complicated. The fact of the matter is that at present we have no firm theoretical foundation for deciding which state variables to include in an MDCF; it is notable, and lamentable, that most of the MDCFs proposed to date draw their functional form from steady-state queueing analyses, often of very simple models. For example, the MDCFs of Albey et al. (2014, 2017) follow the functional form suggested by Karmarkar (1989) which is motivated primarily by steady-state analysis of the *M/M/1* queue. The experimental work of Gopalswamy and Uzsoy (2019) suggests that the empirical functional forms used extensively in the past cannot provide good fits across the entire operating range of workloads a production resource will encounter.

Our current state of knowledge suggests that the best approach to fitting clearing functions available at present is the use of concave piecewise linear regression, which can be formulated as a mixed integer program (Toriello and Vielma 2012; Gopalswamy et al. 2019) although the solution of large models with many data points remains computationally challenging. The piecewise linear approach allows great modeling flexibility and yields a clearing function that when implemented in the allocated clearing function model of Chap. 7 results in a linear program. However, the establishment of a strong theoretical and methodological foundation for the fitting of clearing functions, encompassing both the choice of state variables and of a suitable functional form, remains important directions for future research. The promising performance of clearing function based production planning models presented in this volume suggests that this effort may well be worthwhile.

## 11.3 Some Directions for Future Research

The limitations of the clearing function approach discussed in the previous section suggest a broad range of interesting research questions for the future, many of which lie at the intersection of what have traditionally been viewed as quite distinct research streams. The basic idea of a clearing function lies at the intersection of queueing and mathematical programming models of production systems, research areas that have developed largely independently until today. In this section, we discuss several longer-term research efforts that can build on the clearing function concept, but which address much broader issues spanning several research streams and mathematical modeling tools.

It is clear from the discussion in Chaps. 5 and 6 of this volume that, from a technical perspective, order release models and mechanisms that assume fixed, exogenous lead times are quite mature, although their integration into the overall PPC system can raise numerous questions. Load-dependent lead times, on the other hand, are much more difficult to handle technically (and also organizationally, although this is not our primary focus), and this research stream is far from mature. This provides great opportunities for researchers to advance the frontier of our knowledge in this domain that is, as described in Chap. 1, an essential element of the PPC architecture in most discrete manufacturing companies. We now briefly describe some of the most important research questions, starting with technical issues and proceeding to more conceptual topics.

### *11.3.1 Parameter Setting for Order Release Models*

Order release models with exogenous lead times require lead time parameters that are often taken to be constant over time as in Chap. 5, but can also vary over time as discussed in Chap. 6. Clearing function models must specify the functional form and the shape parameters of the clearing functions. These parameters anticipate the behavior of the production units, but since both the realized cycle times and the realized output are random variables subject to often unknown and changing probability distributions, the parameters cannot simply be set to the "correct" values. This is especially evident for clearing functions where the conditional distribution of the output for a given planned load depends on various factors including the order release pattern itself, due to the planning circularity described in Chap. 2. Therefore, the choice of parameter values encompasses both an anticipation aspect (how accurately will the clearing function anticipate the realized output from the production unit or resource?) and an implicit decision as to the tradeoffs between WIP and FGI inventory levels and due date performance (with the importance of the latter depending on whether safety stocks are maintained, as discussed below). The performance of order release models can be quite sensitive to the parameter values as indicated, e.g., by the performance differences between fixed lead time models with integer and fractional lead times discussed in Chaps. 5 and 10. Very similar questions can be raised in terms of estimating suitable fixed lead times: assuming the distribution of the cycle times in each planning period was known, what is the optimal value of the planned lead times?

Considering the anticipation aspect of the parameter setting problem, one would assume that best performance can be achieved by setting, e.g., clearing function parameters to the values obtained from observation, such as running a least-squares regression over observed load–output data. However, the parameter setting that yields the best system performance can be substantially different (Kacar and Uzsoy 2015) and the mechanisms behind these deviations are not fully understood. We must also keep in mind that the vast majority of research on parameterization issues is performed on simulated data. Empirical data exhibit substantial noise which makes functional relationships between load and output difficult to identify (Fine and Graves 1989; Häussler and Missbauer 2014), and further research is needed to examine the validity of insights obtained from simulations to real-life situations.

### *11.3.2 A Deeper Understanding of Clearing Functions: Properties, Theoretical Basis, and Integration with Order Release Models*

In Chap. 7, clearing functions were motivated by queueing models that suggest a concave, saturating functional relationship between WIP and output, caused primarily by the variability of the arrival and departure process. However, the clearing function concept was introduced by Graves (1986) assuming that output in a period is proportional to the load in that period. The smoothing parameter that is implied by this model is assumed to be "set … so that the resulting time series for production is consistent with the work center capability" which can be obtained by assuming that "As a queue builds at a work center, a manager will direct more resources to the work center to reduce the queue to normal levels" (p. 524). Hence this proportional clearing function models the effect of a production control rule, which is quite distinct from the variability argument invoked to justify the nonlinear, saturating shape. Linear and saturating clearing functions differ not just with respect to their shape, but also with respect to the underlying phenomenon they seek to represent. It is important to keep both modeling aspects of clearing functions in mind—modeling variability versus modeling production control rules. The latter aspect opens up the possibility of modeling behavioral aspects such as load-dependent processing times, possible capacity loss due to congestion (e.g., because material must be shuffled by the production workers), etc. The modeling of these often largely informal factors is still at its beginning—an important research question within behavioral operations management—and can substantially influence the behavior of clearing function models.

Applying clearing functions in a transient regime leads to additional complications. While it is easy to prove that decomposing the workload in a period $t$ (the explanatory variable of most one-dimensional clearing functions) into its components and formulating a multi-dimensional clearing function that takes the history of the process into account can improve output estimation, incorporating this function into order release models can lead to oscillating order releases as discussed in Chap. 8. Naïvely we would assume that more accurate anticipation of the output should improve the performance of the optimization model, but apparently things are not that simple. This indicates that a comprehensive, consistent theory of order release models incorporating functions that estimate the conditional future output is not yet available. This aspect also raises the question of which characteristics of queueing systems are most critical to the model behavior and thus should be modeled most accurately. These might include the steady-state behavior, the WIP and output evolution in the transient phase, or the propagation of variability through the workcenters, including the transient phase.

Production orders that are released are eventually finished, unless they are canceled deliberately; only their finish time is uncertain. Clearing function models express this timing uncertainty as an uncertainty in output quantities across the periods. This maintains the basic structure of production planning models established in the pioneering works described in Sect. 4.6, and is a significant modeling decision since this basic structure was not originally designed for handling lead times. Integrating lead time variables into this modeling framework, that is, expressing the timing uncertainty directly, either leads to intractable model structures or discards the tight relationship between WIP and cycle time demonstrated in Chap. 2. The difficulty with cycle time oriented release models and iterative approaches is an immediate consequence. Research on alternative modeling approaches such as robust optimization is an obvious possibility.

### *11.3.3 Integration of Order Release into the Overall Supply Chain*

Both model-based and rule-based order release mechanisms mainly deal with order releases to single production units. This is adequate for MTO companies where customer orders translate directly to production orders in the order pool and are processed mainly by a single production unit, like the CD/DVD manufacturer in Sect. 1.2.2. However, if the orders must be processed sequentially by multiple production units as in semiconductor manufacturing (Sect. 1.2.1), the order releases to the production units must be coordinated according to the BOM structure. Fixed lead times greatly simplify this material coordination task (de Kok and Fransoo 2003) and can, in principle, be combined with clearing functions (Jansen et al. 2013). Extending load-dependent lead time models to incorporate material coordination along the multistage production-inventory system is much more difficult and remains a topic for future research. A central question is whether a sequential approach that derives the demand for the end items of a production unit from the planned releases to the downstream production units, or an integrated model that encompasses all production units simultaneously, extending the fixed lead time supply chain model in de Kok and Fransoo (2003), is preferable. This problem can be viewed, at an extreme, as that of incorporating the queueing behavior of resources described in Chap. 2 into the MRP logic of the MPC framework described in Chap. 3; should we modify the release schedule obtained by the MRP logic after the fact to accommodate the effects of limited capacity and queueing, or should this be done within the MRP run itself in some way? Analogously, if master planning as implemented in APS systems is applied the queueing perspective must be integrated into the capacity and lead time modeling used at this level.

Order release models integrate the tasks of production smoothing and cycle time anticipation and control. Production smoothing is also performed at the master planning/MPS level to ensure that the number of end items requested from a production unit in each period (the $D_{jt}$ parameters in the release models) is consistent with the capabilities of the production unit. Hence the smoothing capabilities required at the order release stage depend on the smoothing logic at the master planning/MPS level, and perhaps even at the lot-sizing level. Seamless integration of these levels requires consistency between the decision models at each level, in particular in how they anticipate the dynamic behavior of the production units. The anticipation models applied at the master planning/MPS level should be aggregate versions of the respective models applied for order release. Since master planning models and resource profiles for master production scheduling are generally not WIP based, this is not trivial. The research task behind this is the aggregation of transient queueing networks. Finding suitable approximations that are applicable in practice is still largely unsolved; Zäpfel and Missbauer (1993) give a first attempt to handle this aggregation problem. To what extent aggregation methods designed for the steady state, such as the effective processing time approach of Kock et al. (2011), Hopp and Spearman (2008), and Veeger et al. (2011), can be extended to the transient states implied by load-dependent lead times remains to be clarified.

Integrating order release models into the material coordination task also requires the determination of inventory levels of stock keeping units between the production units, including safety stocks. Order release models that explicitly determine WIP levels within the production units provide, at least in principle, the possibility of considering the interaction between WIP and safety stock that is evident from inventory theory—the stockout probability for a given demand distribution depends on both the safety stock and the WIP level, with WIP providing some functionality of safety stock (Graves 1988). Simultaneous optimization of safety stocks and lead times or planned WIP, respectively, using clearing function models and an aggregate representation of the production units has been demonstrated in Albey et al. (2015), Albey and Uzsoy (2016), Aouam and Uzsoy (2012, 2015), and Orcun et al. (2009). Extending this work to multistage systems and/or more detailed representations of the production units is an obvious research topic.

Optimizing inventory levels refers to handling uncertainty of demand and other planning parameters. Since demand uncertainty is usually higher for more remote planning periods, which implies that for a certain period it becomes smaller as the time of planning proceeds, simply optimizing the (safety) stock levels for given demand distributions can lead to exaggerated planned stock levels for the more distant planning periods that most likely are corrected in the course of rolling horizon planning; the order release and capacity plans are biased systematically. Stochastic demand and specific planning rules for responding to demand render future production quantities random variables (de Kok and Fransoo 2003), as considered in the original work of Holt et al. (1960). Aouam and Uzsoy (2012) find that in their simple setting linear decision rules for updating the planned production quantities perform well, which is an encouraging result and in line with adjustable robust optimization (Gorissen et al. 2015). How to extend this to realistic settings and how to assign/split up the task of setting planned inventory levels to/between the order release and the master planning/MPS level is a challenging research topic since it encompasses the hierarchical design of the entire PPC system. Again the need to consider the impact of planned WIP at the master planning level is evident.

### *11.3.4 Advanced Techniques for Flow Time and Output Modeling*

Determining lead times and order release quantities simultaneously requires an anticipation model that predicts the future values of these performance measures for a given release schedule. Univariate clearing functions consider only some measure of WIP as explanatory variable, while multi-dimensional clearing functions (MCDFs) allow more accurate representation of the causal mechanisms that lead to certain values for cycle times and output. The conceptual problems of MDCFs, especially when transient effects are modeled, are described in Chap. 8. Very accurate cycle time and output prediction can be obtained by discrete-event simulation, but this leads to the difficulties described in Sect. 6.6.

This dilemma motivates the use of metamodeling for cycle time and output prediction and metamodel-based rather than simulation-based optimization (Barton and Meckesheimer 2006). When appropriately trained or parameterized, the metamodel, which is usually a deterministic function, yields estimates of the performance measures of the production unit very close to the simulation output as a function of the input variables, which in our case are the order releases over time. The impact of the relevant parameters that describe the properties of the material flow, such as machine failure characteristics, lots sizes, and operation times of the production lots, are either coded in the metamodel or are declared as arguments of the metamodel depending on its specification. Metamodels can be represented by generic functions such as polynomials, functions that are based on certain theoretical requirements on their shape (e.g., the MDCF in Häussler and Missbauer 2014), or by artificial neural networks. Applying metamodels to anticipate output and cycle times is an extension of MDCFs. Based on Yang and Liu (2012) who propose a metamodel for the transient analysis of queueing systems, Li et al. (2016) develop a metamodel that receives the release quantities in the planning periods as input and yields the first two moments of WIP and output in the planning periods. Given this metamodel, the releases are approximately optimized using a multi-objective genetic algorithm. The metamodel considers both the departures and the queue length over the relevant past periods for the output prediction, making reasonable assumptions about the underlying time series model, and is fitted using extensive simulation data. The model performs very well compared to a simulation-based optimization with excessively long computer run times. Since there is no sharp boundary between MDCFs and metamodels, there might be a number of ways to formulate and refine metamodels for anticipating output and cycle times that can be explored in future research.

In the metamodeling approach, the release decision is decomposed into the two phases of pre-computing the metamodel and optimizing the decision variables, and the metamodel is fitted which usually requires large amounts of data that normally can only be obtained by simulation. Both of these aspects raise difficulties: the decomposition into metamodeling and optimization might result in suboptimal solutions, and the effects of informal shop-floor control rules might be difficult to capture in a simulation model. This motivates the use of machine learning techniques that strictly learn from observed data, either to learn the response of the production unit to control inputs (e.g., releases) or to learn near-optimal control inputs for a given state of the system directly (for this distinction, see Bertsimas et al. 2019). While the application of machine learning at the scheduling level has been explored extensively (Aytug et al. 2005), very few papers apply machine learning at the order release level. Lee et al. (1997) use machine learning to select the release sequencing rule in a CONWIP system. Paternina-Arboleda and Das (2001) use reinforcement learning to optimize the operation of an extended CONWIP system which also constrains the buffers at the workcenters and allows emergency authorizations of releases, similar to the force release option in LUMS (see Chap. 4). Häussler and Schneckenreither (2019) use an artificial neural network to predict the cycle time of a new order entering the production unit and, based on this estimation, determine the release times of the production orders, thus decomposing the problem of jointly determining the release times of the orders to single-order release problems that are combined by an algorithm developed in the paper. Clearly these approaches are first attempts, and further research in this area seems fruitful.

## 11.4 Conclusions

The domain of production planning is viewed by many as a mature area where all interesting problems have already been solved. We hope that the results presented in this volume have raised more questions in the mind of the reader than they have answered; this has been the effect of this work on the authors, in any event. There remain many challenging problems that, if even approximately solved, have the potential to yield significant economic benefit to many sectors of the economy. The convergence of vast computing power, data collection and storage technologies and extremely efficient optimization solvers, as well as developments in data analytics, stochastic optimization and machine learning, open new possibilities for advances in this area which has, after all, been central to the development of operations research, operations management, production economics, and industrial engineering since the inception of those disciplines. It is, we believe, a good time to be working in production planning and will only get better.

## References

Albey E, Uzsoy R (2016) A chance constraint based multi-item production planning model using simulation optimization. In: Winter simulation conference, Arlington, VA

Albey E, Bilge U, Uzsoy R (2014) An exploratory study of disaggregated clearing functions for multiple product single machine production environments. Int J Prod Res 52(18):5301–5322

Albey E, Norouzi A, Kempf KG, Uzsoy R (2015) Demand modeling with forecast evolution: an application to production planning. IEEE Trans Semicond Manuf 28(3):374–384

Albey E, Bilge U, Uzsoy R (2017) Multi-dimensional clearing functions for aggregate capacity modelling in multi-stage production systems. Int J Prod Res 55(14):4164–4179

Aouam T, Uzsoy R (2012) Chance-constraint-based heuristics for production planning in the face of stochastic demand and workload-dependent lead times. In: Armbruster D, Kempf KG (eds) Decision policies for production networks. Springer, London, pp 173–208

Aouam T, Uzsoy R (2015) Zero-order production planning models with stochastic demand and workload-dependent lead times. Int J Prod Res 53(6):1–19

Arrow KJ, Karlin S, Scarf H (1958) Studies in the mathematical theory of inventory and production. Stanford University Press, Stanford

Askin RG, Hanumantha GJ (2018) Queueing network models for analysis of nonstationary manufacturing systems. Int J Prod Res 56(1–2):22–42

Asmundsson J, Rardin RL, Uzsoy R (2006) Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Trans Semicond Manuf 19(1):95–111

Asmundsson JM, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning models with resources subject to congestion. Nav Res Logist 56(2):142–157

Aytug H, Lawley MA, McKay KN, Mohan S, Uzsoy R (2005) Executing production schedules in the face of uncertainty: a review and some future directions. Eur J Oper Res 161:86–110

Baker KR (1993) Requirements planning. In: Graves SC, AHG RK, Zipkin PH (eds) Handbooks in operations research and management science. Logistics of production and inventory, vol 3. Elsevier Science Publishers, Amsterdam, pp 571–627

Barton RR, Meckesheimer M (2006) Chapter 18: Metamodel-based simulation optimization. In: Henderson SG, Nelson BL (eds) Handbooks in operations research and management science, vol 13. Elsevier, Amsterdam, pp 535–574

Bertsimas D, Dunn J, Mundru N (2019) Optimal prescriptive trees. INFORMS Journal on Optimization 1(2):164–183

Buffa ES, Taubert WH (1972) Production-inventory systems; planning and control. R. D. Irwin, Homewood

Carey M (1987) Optimal time-varying flows on congested networks. Oper Res 35(1):58–69

Carey M (1990) Extending and solving a multiperiod congested network flow model. Comput Oper Res 17(5):495–507

Carey M, Bowers M (2012) A review of properties of flow–density functions. Transp Rev 32(1):49–73

Carey M, Subrahmanian E (2000) An approach to modelling time-varying flows on congested networks. Trans Res B Methodol 34:157–183

de Kok AG, Fransoo JC (2003) Planning supply chain operations: definition and comparison of planning concepts. In: de Kok AG, Graves SC (eds) Handbooks in operations research and management science. Supply chain management: design, coordination and operation, vol 11. Elsevier, Amsterdam, pp 597–675

Fine CH, Graves SC (1989) A tactical planning model for manufacturing subcomponents of mainframe computers. J Manuf Oper Manag 2:4–34

Fu MC (ed) (2015) Handbook of simulation optimization. Springer, New York

Gopalswamy K, Uzsoy R (2019) A data-driven iterative refinement approach for estimating clearing functions from simulation models of production systems. Int J Prod Res 57(19):6013–6030

Gopalswamy K, Fathi Y, Uzsoy R (2019) Valid inequalities for concave piecewise linear regression. Oper Res Lett 47:52–58

Gorissen BL, Yanikoglu I, den Hertog D (2015) A practical guide to robust optimization. Omega 53:124–137

Graves SC (1986) A tactical planning model for a job shop. Oper Res 34(4):522–533

Graves SC (1988) Safety stocks in manufacturing systems. J Manuf Oper Manag 1:67–101

Greenberg HJ (1996) The analyze rulebase for supporting LP analysis. Ann Oper Res 65:91–126

Hackman ST (2008) Production economics: integrating the microeconomic and engineering perspectives. Springer, Berlin

Hanssmann F (1959) Optimal inventory location and control in production and distribution networks. Oper Res 7(4):483–498

Häussler S, Missbauer H (2014) Empirical validation of meta-models of work centres in order release planning. Int J Prod Econ 149:102–116

Häussler S, Schneckenreither M (2019) Adaptive order release planning with dynamic lead times: a machine learning approach. University of Innsbruck, Innsbruck

Holt CC, Modigliani F, Muth JF, Simon HA (1960) Planning production, inventories and work force. Prentice Hall, Englewood Cliffs

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Hung YF, Leachman RC (1996) A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. IEEE Trans Semicond Manuf 9(2):257–269

Jacobs FR, Berry WL, Whybark DC, Vollmann TE (2011) Manufacturing planning and control for supply chain management. Irwin/McGraw-Hill, New York

Jansen MM, de Kok TG, Fransoo JC (2013) Lead time anticipation in supply chain operations planning. OR Spectr 35(1):251–290

Johnson LA, Montgomery DC (1974) Operations research in production planning, scheduling and inventory control. Wiley, New York

Kacar NB, Uzsoy R (2015) Estimating clearing functions for production resources using simulation optimization. IEEE Trans Autom Sci Eng 12(2):539–552

Karmarkar US (1987) Lot sizes, lead times and in-process inventories. Manag Sci 33(3):409–418

Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123

Karmarkar US, Kekre S, Kekre S (1992) Multi-item batching heuristics for minimization of queues. Eur J Oper Res 58:99–111

Kock AAA, Veeger CPL, Etman LFP, Lemmen B, Rooda JE (2011) Lumped parameter modelling of the litho cell. Prod Plan Control 22(1):41–49

Law AM, Kelton WD (2004) Simulation modeling and analysis. McGraw-Hill, New York

Lee CY, Piramuthu S, Tsai YK (1997) Job shop scheduling with a genetic algorithm and machine learning. Int J Prod Res 35(4):1171–1191

Li M, Yang F, Uzsoy R, Xu J (2016) A metamodel-based Monte Carlo simulation approach for responsive production planning of manufacturing systems. J Manuf Syst 38:114–133

Orcun S, Uzsoy R, Kempf KG (2009) An integrated production planning model with load-dependent lead-times and safety stocks. Comput Chem Eng 33(12):2159–2163

Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York

Paternina-Arboleda CD, Das TK (2001) Intelligent dynamic control policies for serial production lines. IIE Trans 33(1):65–77

Srinivasan A, Carey M, Morton TE (1988) Resource pricing and aggregate scheduling in manufacturing systems. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh

Stadtler H, Kilger C, Meyr H (2015) Supply chain management and advanced planning. Concepts, models, software, and case studies. Springer, Berlin

Sterman JD (2000) Business dynamics: systems thinking and modeling for a complex world. McGraw-Hill, New York

Toriello A, Vielma JP (2012) Fitting piecewise linear continuous functions. Eur J Oper Res 219:86–95

Veeger CPL, Etman LFP, Lefeber E, Adan IJBF, van Herk J, Rooda JE (2011) Predicting cycle time distributions for integrated processing workstations: an aggregate modeling approach. IEEE Trans Semicond Manuf 24(2):223–236

Voss S, Woodruff DL (2006) Introduction to computational optimization models for production planning in a supply chain. Springer, New York

Yang F, Liu J (2012) Simulation-based transfer function modeling for transient analysis of general queueing systems. Eur J Oper Res 223(1):150–166

Zäpfel G, Missbauer H (1993) Production planning and control (PPC) systems including load-oriented order release—problems and research perspectives. Int J Prod Econ 30/31:107–122

# Index

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3