Chapter 1
Introduction

In July 1958, Jack Kilby of Texas Instruments suggested building all of the components of a circuit completely in silicon [1]. By September 12, 1958, Kilby had built a working model of the first "solid circuit," the size of a pencil point. A couple of months later in January 1959, Robert Noyce of Fairchild Semiconductor developed a better way to connect the different components of a circuit [2, 3]. Later, in the spring of 1959, Fairchild Semiconductor demonstrated the first planar circuit—a "unitary circuit." The first monolithic integrated circuit (IC) was born, where multiple transistors coexisted with passive components on the same physical substrate [4]. Microphotographs of the first IC (Texas Instruments, 1958), the first monolithic IC (Fairchild Semiconductor, 1959), and the high performance i7-6700K Skylake Quad-Core microprocessor with up to 4.2 GHz clock frequency (Intel Corporation, 2015) are depicted in Fig. 1.1.

In 1960, Jean Hoerni invented the planar process [5]. Later, in 1960, Dawon Kahng and Martin Atalla demonstrated the first silicon based metal oxide semiconductor field effect transistor (MOSFET) [6], followed in 1967 by the first silicon gate MOSFET [7]. These seminal inventions resulted in the explosive growth of today's multi-billion dollar microelectronics industry. The fundamental cause of this growth in the microelectronics industry has been made possible by technology scaling, particularly in CMOS technology.

The goal of this chapter is to provide a brief perspective on the development of ICs, introduce power delivery and management in the context of this development, motivate the use of on-chip voltage regulators and decoupling capacitors, and provide guidance and perspective to the rest of this book. The evolution of integrated circuit technology from the first ICs to highly scaled CMOS technology is described in Sect. 1.1. As manufacturing technologies supported higher integration densities and switching speeds, the primary constraints and challenges in the design of integrated circuits have also shifted, as discussed in Sect. 1.2. The basic nature of the problem of distributing power and ground in integrated circuits is described

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_1

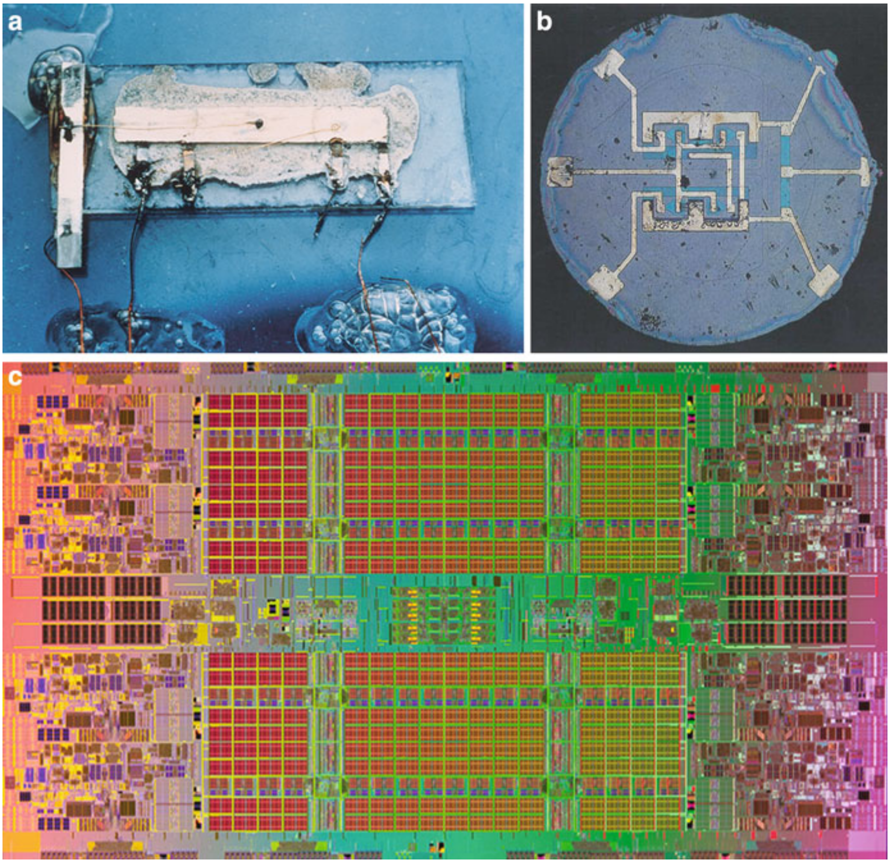

Fig. 1.1 Microphotographs of early and recent integrated circuits (IC) (the die size is not to scale); (**a**) the first IC (Texas Instruments, 1958), (**b**) the first monolithic IC (Fairchild Semiconductor, 1959), (**c**) the high performance i7-6700K Skylake Quad-Core microprocessor (Intel Corporation, 2015)

in Sect. 1.3. The adverse effects of variations in the power supply voltage on the operation of a digital integrated circuit are discussed in Sect. 1.4. Finally, the chapter is summarized in Sect. 1.5.

1.1 Evolution of Integrated Circuit Technology

The pace of IC technology over the past three decades is well characterized by Moore's law. As noted in 1965 by Gordon Moore, the integration density of the first commercial integrated circuits has doubled approximately every year [8]. A prediction was made that the economically effective integration density, i.e.,

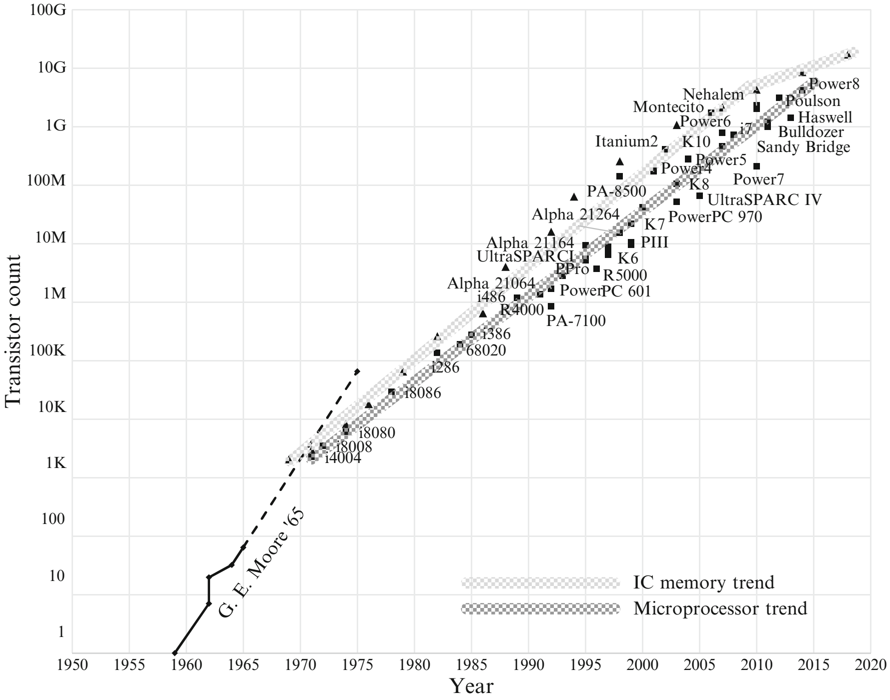

Fig. 1.2 Evolution of transistor count of CPU/microprocessor and memory ICs. In the lower left corner, the original Moore's data [8] is displayed followed by the extrapolated prediction (*dashed line*). The *wide lines* are linearized trends for both IC memory and microprocessors

the number of transistors on an integrated circuit leading to the minimum cost per integrated component, will continue to double every year for another decade. This prediction has held true through the early 1970s. In 1975, the prediction was revised to suggest a new, slower rate of growth–doubling of the IC transistor count every two years [9]. This trend of exponential growth of IC complexity is commonly referred to as "Moore's law." Since the start of commercial production of integrated circuits in the early 1960s, circuit complexity has risen from a few transistors to several billions of transistors functioning together on a single monolithic substrate. This trend is expected to continue at a comparable pace for another decade [10].

The evolution of the integration density of microprocessor and memory ICs is shown in Fig. 1.2 along with the original prediction described in [8]. As seen from the data illustrated in Fig. 1.2, DRAM IC complexity has grown at an even higher rate, quadrupling roughly every three years. The progress of microprocessor clock frequencies is shown in Fig. 1.3. Associated with increasing IC complexity and clock speed is an exponential increase in microprocessor performance (doubling every 18 to 24 month). This performance trend is also referred to as Moore's law.

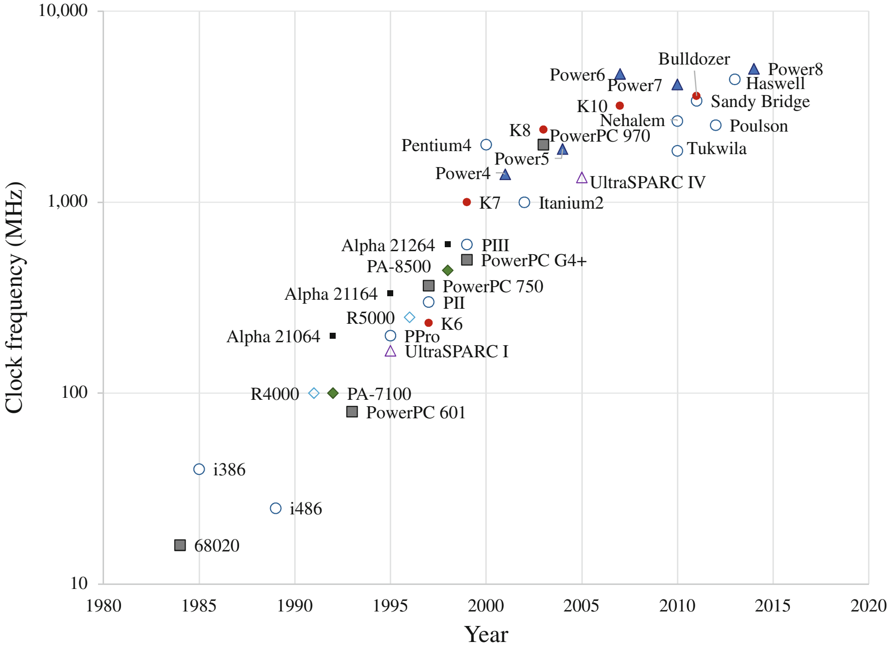

Fig. 1.3 Evolution of microprocessor clock frequency. Several lines of microprocessors are shown in *different colors and shapes*

The principal driving force behind this spectacular improvement in circuit complexity and performance has been the steady decrease in the feature size of semiconductor devices. Advances in optical lithography have allowed manufacturing of on-chip structures with increasingly higher resolution. The area, power, and speed characteristics of transistors with a planar structure, such as MOS devices, improve with the decrease (i.e., scaling) of the lateral dimensions of the devices. These technologies are therefore referred to as *scalable*. The maturing of scalable planar circuit technologies, first PMOS and later NMOS, has allowed the potential of technology scaling to be fully exploited as lithography has improved. The development of planar MOS technology culminated in CMOS circuits. The low power characteristics of CMOS technology deferred the effects of thermal limitations on integration complexity and permitted technology scaling to continue unabated through the 1980s, 1990s, 2000s, and 2010s making CMOS the digital circuit technology of choice.

From a historical perspective, the development of scalable ICs was simultaneously circuitous and serendipitous, as described by Murphy, Haggan, and Troutman [11]. Although the ideas and motivation behind scalable ICs seem straightforward from today's vantage point, the emergence of scalable commercial ICs was neither inevitable nor a result of a well guided and planned pursuit. Rather, the original motivation for the development of integrated circuits was circuit miniaturization for military and space applications. Although the active devices of the time, discrete transistors, offered smaller size (and also lower power dissipation with higher reliability) as compared to vacuum tubes, much of this advantage was lost at the circuit level, as the size and weight of electronic circuits were dominated by passive components, such as resistors, capacitors, and diodes. Thus, the original objective was to reduce the size of the passive elements through integration of these elements onto the same die as the transistors. The cost effectiveness and commercial success of high complexity ICs were highly controversial for several years after the integrated circuit was invented. Successful integration of a large number of transistors on the same die seemed infeasible, considering the yield of discrete devices at the time [11].

Many obstacles precluded early ICs from scaling. The bulk collector bipolar transistors used in these early ICs suffered from performance degradation due to high collector resistance and, more importantly, the collectors of all of the on-chip transistors were connected through the bulk substrate. The speed of a bipolar transistor does not, in general, scale with the lateral dimensions (i.e., vertical NPN and PNP doping structures typically determine the performance). In addition, early device isolation approaches were not amenable to scaling and consumed significant die area. On-chip resistors and diodes also made inefficient use of die area. Scalable schemes for device isolation and interconnection were therefore essential to truly scale ICs. It was not until these problems were solved and the structure of the bipolar transistor was improved that device miniaturization led to dramatic improvements in IC complexity.

The concept of scalable ICs received further development with the maturation of the MOS technology. Although the MOS transistor is a contemporary of the first ICs, the rapid progress in bipolar devices delayed the development of MOS ICs at the beginning of the IC era. The MOS transistor lagged in performance as compared with existing bipolar devices and suffered from reproducibility and stability problems. The low current drive capability of MOS transistors was also a serious disadvantage at low integration densities. Early use of the MOS transistor was limited to those applications that exploited the excellent switch-like characteristics of the MOS devices. Nevertheless, the circuit advantages and scaling potential of MOS technology were soon realized, permitting MOS circuits to gain increasingly wider acceptance. Gate insulation and the enhancement mode of operation made MOS technology ideal for direct-coupled logic [12]. Furthermore, MOS did not suffer from punch-through effects and could be fabricated with higher yield. The compactness of MOS circuits and the higher yield eventually resulted in a fourfold density advantage in devices per IC as compared to bipolar ICs. Ironically, it was the refinement of bipolar technology that paved the path to these larger scales of integration, permitting the efficient exploitation of MOS technology. With advances in lithographic resolution, the MOS disadvantage in switching speed as compared to bipolar devices gradually diminished. The complexity of bipolar ICs had become primarily constrained by power dissipation. As a result, MOS emerged as the dominant digital integrated circuit technology.

1.2 Evolution of Design Objectives

Advances in fabrication technology and the emergence of new applications have induced several shifts in the principal objectives in the design of integrated circuits over the past 50 years. The evolution of the IC design paradigm is illustrated in Fig. 1.4.

In the 1960s and 1970s, yield concerns served as the primary limitation to IC integration density and, as a consequence, circuit compactness and die area were the primary criteria in the IC design process. Due to limited integration densities, a typical system at the time would contain dozens to thousands of small ICs. As a result, chip-to-chip communications traversing board-level interconnect limited overall system performance. As compared to intra-chip interconnect, board level interconnect have high latency and dissipate large amounts of power, limiting the speed and power of a system. Placing as much functionality as possible into a yield limited silicon die supported the realization of electronic systems with fewer ICs. Fewer board level contacts and interconnections in systems comprised of fewer ICs improved system reliability and lowered system cost, increased system speed (due to lower communication latencies), reduced system power consumption, and decreased the size and weight of the overall system. Producing higher functionality per silicon area with the ensuing reduction in the number of individual ICs typically achieved an improved cost/performance tradeoff at the system level. A landmark example of that era is the first Intel microprocessor, the 4004, commercialized at the end of 1971 [13]. Despite the limitation to 4-bit word processing and initially operating at a mere 108 kHz, the 4004 microprocessor was a complete processor core built on a monolithic die containing approximately 2300 transistors. A microphotograph of the 4004 microprocessor is shown in Fig. 1.5.

The impact of off-chip communications on overall system speed decreased as the integration density increased with advances in fabrication technology, lowering the

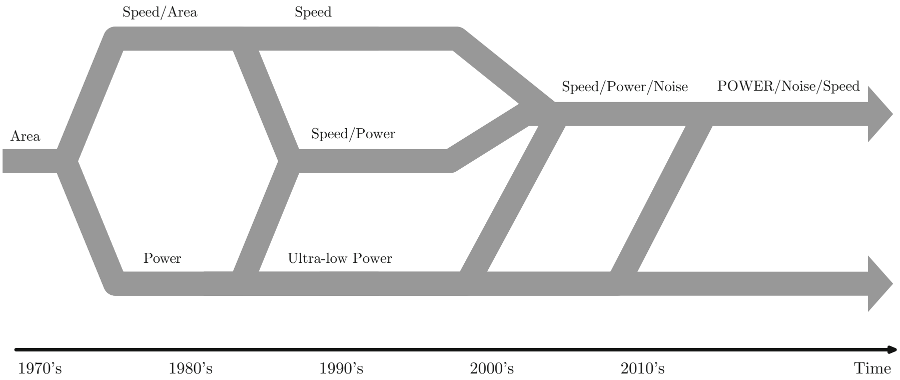

Fig. 1.4 Evolution of design criteria in CMOS integrated circuits

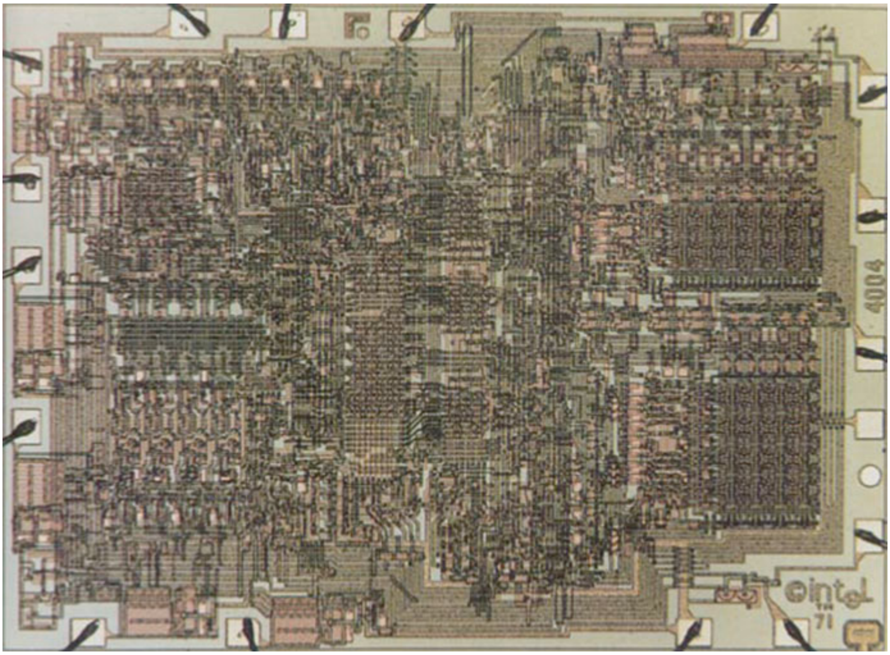

Fig. 1.5 Microphotograph of the 4004—the first microprocessor manufactured on a monolithic die

number of ICs comprising a system. System speed became increasingly dependent on the speed of the component ICs (and less dependent on the speed of the board-level communications). By the 1980s, circuit speed had become the design criterion of greatest significance. Concurrently, a new class of applications emerged, principally restricted by the amount of power consumed. These applications included digital wrist watches, handheld calculators, pacemakers, and satellite electronics. These applications established a new design concept—design for ultra-low power, i.e., power dissipation being the primary design criterion, as illustrated by the lowest path shown in Fig. 1.4.

As device scaling progressed and a greater number of components were integrated onto a single die, on-chip power dissipation began to produce significant economic and technical difficulties. While the market for high performance circuits could support the additional cost, the design process in the 1990s had focused on optimizing both speed and power, borrowing a number of design approaches previously developed for ultra-low power products. The proliferation of portable electronic devices further increased the demand for power efficient and ultra-low power ICs, as shown in Fig. 1.4.

A continuing increase in power dissipation exacerbated system price and reliability concerns, making power a primary design metric across an entire range of applications. The evolution of power consumed by several lines of commercial

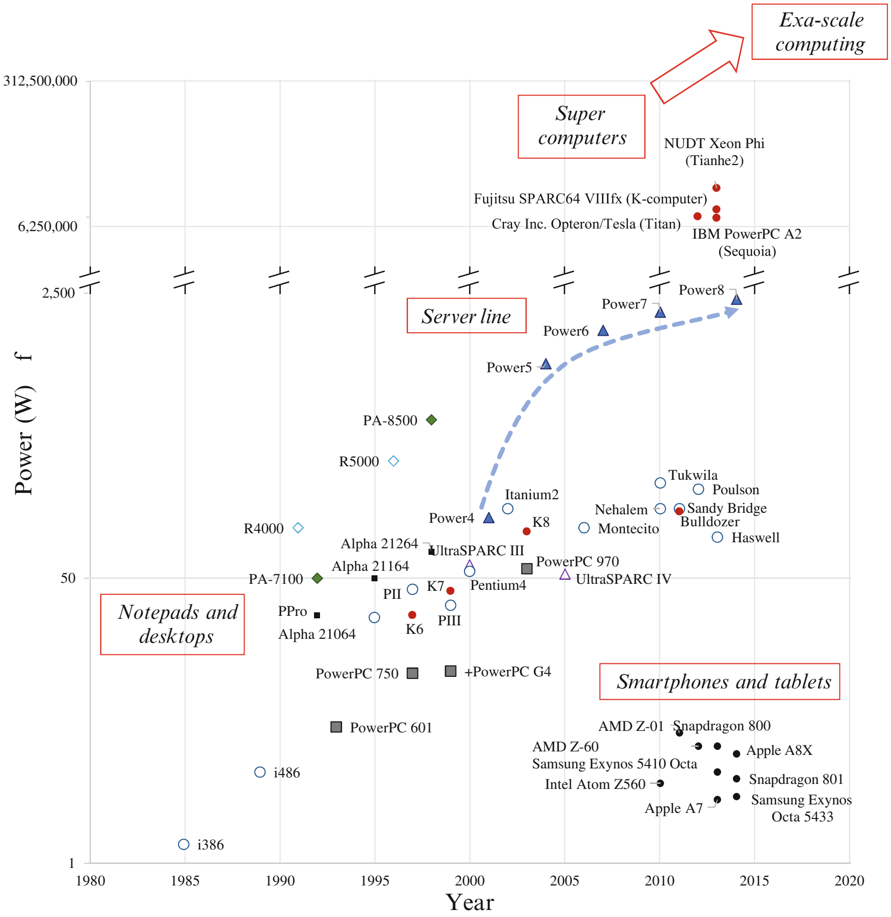

Fig. 1.6 Evolution of microprocessor power consumption. Several lines of microprocessors are shown in *different colors and shapes*

microprocessors is shown in Fig. 1.6. Furthermore, aggressive device scaling and increasing circuit complexity have caused severe noise (or signal integrity) issues in high complexity, high speed integrated circuits. Although digital circuits have traditionally been considered immune to noise due to the inherently high noise margins, circuit coupling through the parasitic impedances of the on-chip interconnect has significantly increased with technology scaling. Ignoring the effects of on-chip noise is no longer possible in the design of high speed digital ICs. These changes are reflected in the convergence of “speed” and “speed/power” design criteria to “speed/power/noise,” as depicted in Fig. 1.4.

As device scaling continued in the twenty first century, more than seven billions transistors have successfully been integrated onto a single die [14], keeping up with Moore’s law. As a result, the overall power dissipation increased accordingly, exceeding the maximum capability of conventional cooling technologies. Any further increase in on-chip power dissipation would require either expensive and challenging technology solutions, such as liquid cooling, significantly increasing the overall cost of a system, or innovations in system architecture that exploit massive integration levels or local functional characteristics. Moreover, an explosive growth of portable and handheld devices, such as cell phones and personal device assistants (PDAs), resulted in a shift of design focus towards low power. As an architectural solution for low power in high performance ICs, multi-core systems emerged [15–18], trading off silicon area with on-chip power dissipation. Since the emphasis on ultra-low power design continues in the second decade of the twenty first century, major design effort is focused on reducing system-level power dissipation.

1.3 The Issue of Power Delivery and Management

The issue of power delivery is illustrated in Fig. 1.7, where a simple power delivery system is shown. The system consists of a power supply, a power load, and interconnect lines connecting the supply to the load. The power supply is assumed to behave as an ideal voltage source providing nominal power and ground voltages, V_{dd} and V_{gnd}. The power load is modeled as a variable current source $I(t)$. The interconnect lines connecting the supply and the load are not ideal; the power and ground lines have, respectively, a finite parasitic resistance R_p and R_g, and inductance L_p and L_g. Resistive voltage drops $\Delta V_R = IR$ and inductive voltage drops $\Delta V_L = L\, dI/dt$ develop across the parasitic interconnect impedances, as the load draws current $I(t)$ from the power delivery system. The voltage levels across the load terminals, therefore, change from the nominal level provided by the supply, dropping to $V_{\text{dd}} - IR_p - L_p\, dI/dt$ at the power terminal and rising to $V_{\text{gnd}} + IR_g + L_g\, dI/dt$ at the ground terminal, as shown in Fig. 1.7.

This uncertainty in the supply voltages is referred to as power supply noise. Power supply noise adversely affects circuit operation through several mechanisms,

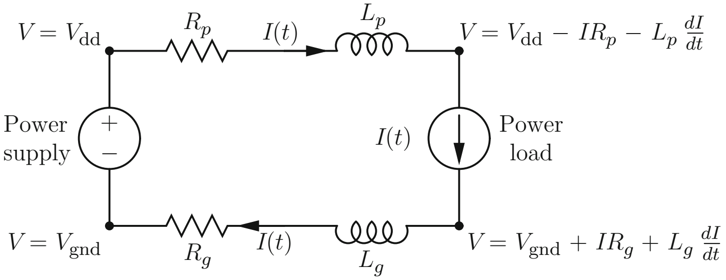

Fig. 1.7 Power delivery system consisting of the power supply, power load, and non-ideal interconnect lines

as described in Sect. 1.4. Proper design of the load circuit ensures correct operation under the assumption that the supply levels are maintained within a certain range near the nominal voltage levels. This range is called the power noise margin. The primary objective in the design of the power delivery system is to supply sufficient current to each transistor on an integrated circuit while ensuring that the power noise does not exceed target noise margins.

The on-going miniaturization of integrated circuit feature size has placed significant requirements on the on-chip power and ground distribution networks. Circuit integration densities rise with each nanometer technology generation due to smaller devices and larger dies; the current density and total current increase accordingly. Simultaneously, the higher speed switching of smaller transistors produces faster current transients within the power distribution network. Both the average current and the transient current are rising exponentially with technology scaling. The evolution of the average current of high performance microprocessors is illustrated in Fig. 1.8.

With thermal design power (TDP) of over 130 W (e.g., the TDP of the Intel Sandy Bridge, Poulson, and Tukwila microprocessors is, respectively, 130 , 170 , and 185 W [19]) and power supply voltage as low as 0.8 V [20], the current in contemporary microprocessors is approaching 200 A and will further increase with technology scaling. Forecasted demands in the power current of high performance

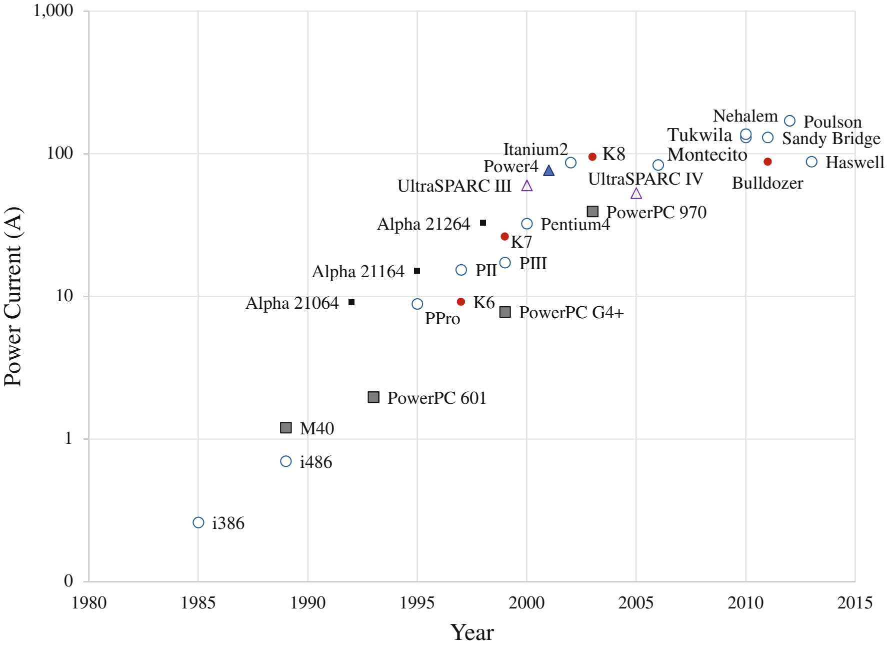

Fig. 1.8 Evolution of the average current in high performance microprocessors. Several lines of microprocessors are shown in *different colors and shapes*

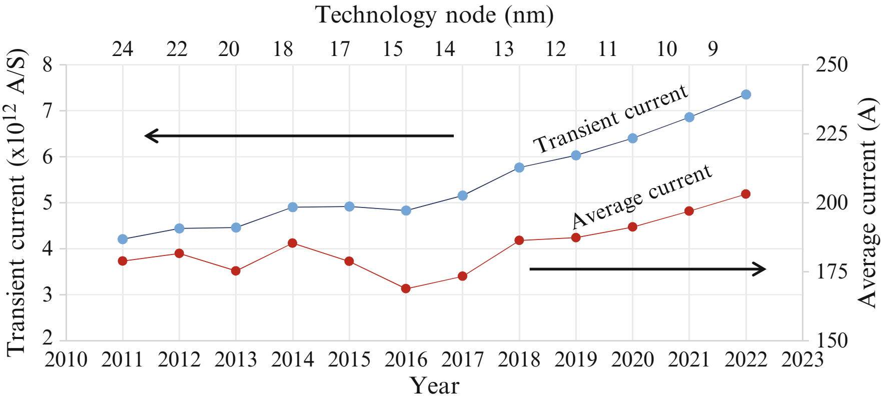

Fig. 1.9 Increasing power current requirements of high performance microprocessors with technology scaling, according to the ITRS roadmap [10]. The average current is the ratio of the circuit power to the supply voltage. The transient current is the product of the average current and the on-chip clock rate, $2\pi f_{\mathrm{clk}}$

microprocessors are illustrated in Fig. 1.9. The rate of increase in the transient current is expected to more than double the rate of increase in the average current, as indicated by the slope of the trend lines depicted in Fig. 1.9.

The faster rate of increase in the transient current as compared to the average current is due to increasing on-chip clock frequencies. The transient current in modern high performance microprocessors is approximately one teraampere per second (10^{12} A/s) and is expected to rise, exceeding seven teraamperes per second by 2022. A transient current of this high magnitude is due to switching hundreds of amperes within tens to hundreds of picoseconds. Fortunately, the rate of increase in the transient current has slowed with the introduction of lower speed multi-core microprocessors. In a multi-core microprocessor, similar performance is achieved at a lower frequency at the expense of increased circuit area.

Insuring adequate signal integrity of the power supply under these high current requirements has become a primary design issue in high performance, high complexity integrated circuits. The high average currents produce large ohmic *IR* voltage drops [21], and the fast transient currents cause large inductive $L\, dI/dt$ voltage drops [22] (ΔI noise) within power distribution networks [23]. Power distribution networks are designed to minimize these voltage drops, maintaining the local supply voltage within specified noise margins. If the power supply voltage sags too low, the performance and functionality of the circuit is severely compromised. Alternatively, excessive overshoot of the supply voltage can affect circuit reliability. Further exacerbating these issues is the reduced noise margins of the power supply as the supply voltage is reduced with each new generation of nanometer process technology, as shown in Fig. 1.10.

To maintain the local supply voltage within specified design margins, the output impedance of a power delivery system should be low as seen at the power

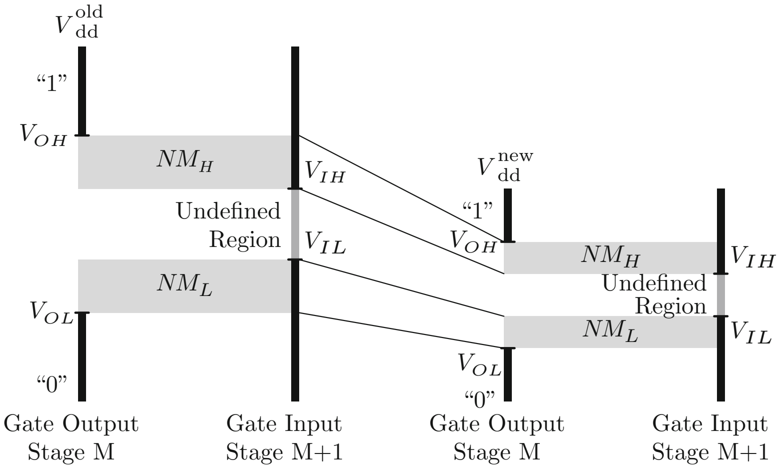

Fig. 1.10 Reduction in noise margins of CMOS circuits with technology scaling. NM_H and NM_L are the noise margins, respectively, in the high and low logic state

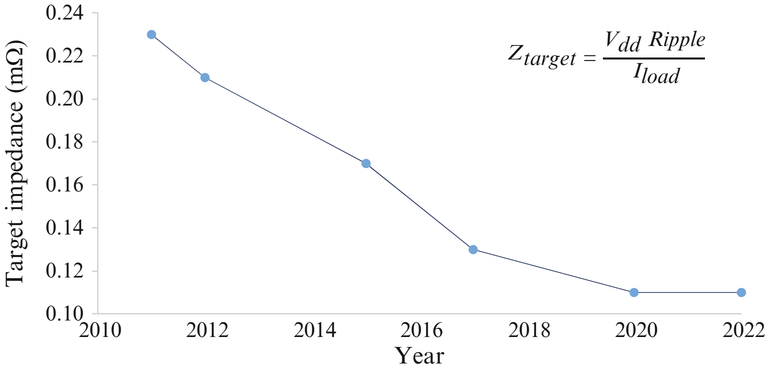

Fig. 1.11 Projections of the target impedance of a power delivery system. The target impedance will continue to drop for future technology generations at an aggressive rate of 1.25 X per technology node [24]

terminals of the circuit elements. IC technologies are expected to scale for another decade [10]. As a result, the average and transient currents drawn from the power delivery network will continue to rise. Simultaneously scaling the power supply voltage, however, has become limited due to threshold variations. The target output impedance of a power delivery system in high speed, high complexity ICs such as microprocessors will therefore continue to drop, reaching an inconceivable level of 150 μΩ by the year 2022 [24], as depicted in Fig. 1.11.

With transistor switching times as short as a few picoseconds, on-chip signals typically contain harmonic frequencies as high as ∼100 GHz. For on-chip wires, the inductive reactance ωL dominates the overall wire impedance beyond ∼10 GHz.

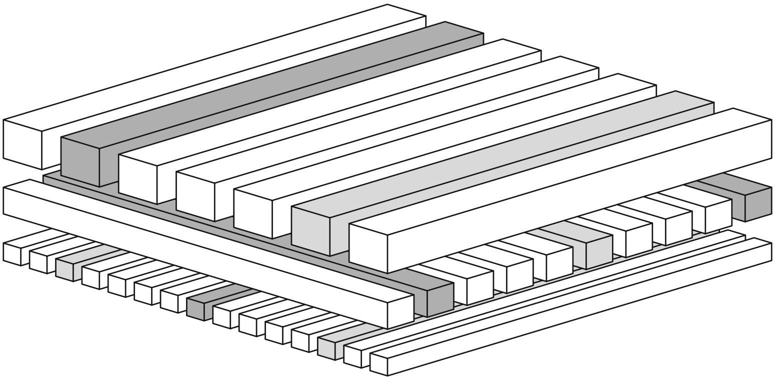

Fig. 1.12 A grid structured power distribution network. The ground lines are *light gray*, the power lines are *dark gray*, and the signal lines are *white*

The on-chip inductance affects the integrity of the power supply through two phenomena. First, the magnitude of the ΔI noise is directly proportional to the power network inductance as seen at the current sink. Second, the network resistance, inductance, and decoupling capacitance form an *RLC* tank circuit with multiple resonances. The peak impedance of this *RLC* circuit must be lowered to guarantee that target power supply noise margins are satisfied. Thus, information characterizing the inductance is needed to correctly design and analyze power delivery systems.

Power distribution networks in high performance digital ICs are commonly structured as a multilayer grid. In such a grid, straight power/ground (P/G) lines in each metalization layer span the entire die (or a large functional unit) and are orthogonal to the lines in the adjacent layers. The power and ground lines typically alternate in each layer. Vias connect a power (ground) line to another power (ground) line at the overlap sites. This power grid organization is illustrated in Fig. 1.12, where three layers of interconnect are depicted with the power lines shown in *dark gray* and the ground lines shown in *light gray*. The power/ground lines are surrounded by signal lines.

A significant fraction of the on-chip resources is committed to insure the integrity of the power supply voltage levels. The global on-chip power delivery system is typically determined at early stages of the design process, when little is known about the local current demands at specific locations on an IC. Additional metal resources for the global power delivery system are typically allocated at later stages of the design process to improve the local electrical characteristics of the power network. A complete redesign of the surrounding global signals can be prohibitively expensive and time consuming. For these reasons, power delivery systems tend to be conservatively designed [25], sometimes using more than a third of the on-chip metal resources [26, 27]. Overengineering the power delivery system is, therefore, costly in modern interconnect limited, high complexity digital integrated circuits.

Performance objectives in power delivery systems, such as low impedance (low inductance and resistance) to satisfy noise specifications under high current loads, small physical area, and low current densities (for improved reliability) are typically in conflict. Widening the lines to increase the conductance and improve the electromigration reliability also increases the grid area. Replacing wide metal lines with narrow interdigitated P/G lines increases the line resistance if the grid area is maintained constant or increases the physical area if the net cross section of the lines is maintained constant. It is therefore important to make a balanced choice under these conditions. A quantitative model of the inductance/area/resistance tradeoff in high performance power distribution networks is therefore needed to achieve an efficient power delivery system. Another important objective is to provide quantitative tradeoff guidelines and intuition in the design of high performance power delivery systems.

Decoupling capacitors are often used to reduce the impedance of a power distribution system and provide the required charge to the switching circuits, lowering the power supply noise [28]. At high frequencies, however, the on-chip decoupling capacitors can be effective due to the high parasitic impedance of the power network connecting a decoupling capacitor to the current load [29]. On-chip decoupling capacitors, however, reduce the self-resonant frequency of a power delivery system, resulting in high amplitude power supply voltage fluctuations at the resonant frequencies. A hierarchical system of on-chip decoupling capacitors should therefore be carefully designed to provide a low impedance, resonant-free power delivery system over the entire range of operating frequencies, while delivering sufficient charge to the switching circuits to maintain the local power supply voltages within target noise margins [30].

In earlier technology generations, high quality DC voltages and currents were delivered from off-chip voltage converters to on-chip load circuitry within carefully designed electrical power grids, producing a power system which was passive in nature. To maintain sufficient quality of power under increasing current densities and parasitic impedances, the power needs to be locally regulated with distributed on-chip voltage converters close to the load. This concept of distributed power delivery poses new power design challenges in modern ICs, requiring circuit level techniques to convert and regulate power at points-of-load (POL), methodological solutions for distributing on-chip power supplies, and automated design techniques to co-design distributed power supplies and decoupling capacitors.

While the quality of power can be addressed with a POL approach, the emerging trends of heterogeneity, on-chip integration, and dynamic control require fundamental changes in traditional power delivery approaches—power delivery systems should not be viewed as a passive power distribution network but rather as systems that need to be efficiently and proactively managed. The regulation of DC voltages close to the load, distributed on-chip current delivery, and local intelligence are all required to efficiently manage power resources in high performance ICs. To address these novel challenges, traditional power delivery and management systems need to be conceptually reorganized. Specialized power delivery circuits, locally intelligent power routers, microcontrollers, and power managing policies have become basic building blocks for delivering and managing power in modern heterogeneous systems.

1.4 Deleterious Effects of Power Distribution Noise

Power noise adversely affects the operation of an integrated circuit through several mechanisms. These mechanisms are discussed in this section. Power supply noise produces uncertainty in the delay of the clock and data signals, as described in Sect. 1.4.1. Power supply noise also increases the uncertainty of the timing reference signals generated on-chip (clock jitter), lowering the clock frequency of the circuit, as discussed in Sect. 1.4.2. The reduction of noise margins is discussed in Sect. 1.4.3. Power supply variations diminish the maximum supply voltage, degrading the speed of operation, as described in Sect. 1.4.4.

1.4.1 Signal Delay Uncertainty

The propagation delay of on-chip signals depends on the power supply voltage during a signal transition. The source of the PMOS transistors in pull-up networks within logic gates is connected to the highest supply voltage directly or through other PMOS transistors. Similarly, the source of the NMOS transistors within a pull-down networks is connected to the lowest supply voltage (directly or through other NMOS transistors). The drain current of an MOS transistor increases with the voltage difference between the transistor gate and source. When the rail-to-rail power voltage is reduced due to power supply variations, the gate-to-source voltage of the NMOS and PMOS transistors is less, lowering the output current of the transistors. The signal delay increases accordingly as compared to the delay under a nominal power supply voltage. Conversely, a higher power voltage and a lower ground voltage shortens the propagation delay. The effect of the power noise on the propagation delay of the clock and data signals is, therefore, an increase in both delay uncertainty and the delay of the data paths [31, 32]. Consequently, power supply noise limits the maximum operating frequency of an integrated circuit [33–35].

1.4.2 On-Chip Clock Jitter

A phase-locked loop (PLL) is often used to generate the on-chip clock signal. An on-chip PLL generates an on-chip clock signal by multiplying the system clock signal. Certain changes in the electrical environment of a PLL, power supply voltage

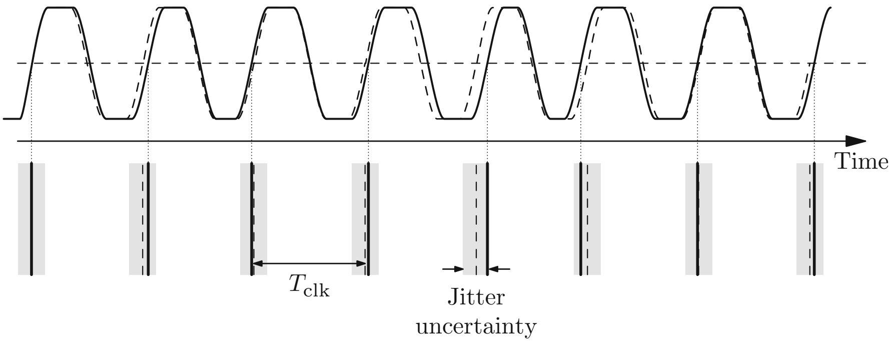

Fig. 1.13 Cycle-to-cycle jitter of a clock signal. The phase of the clock signal (*solid line*) randomly deviates from the phase of an ideal clock signal (*dashed line*)

variations in particular, affect the phase of the on-chip clock signal. A feedback loop within the PLL controls the phase of the PLL output and aligns the output signal phase with the phase of the system clock. Ideally, the edges of the on-chip clock signal are at precisely equidistant time intervals determined by the system clock signal. The closed loop response time of modern PLL is typically hundreds of nanoseconds (e.g., 300 ns in [36]). Disturbances of shorter duration than the PLL response time result in deviations of the on-chip clock phase from ideal timing objectives. These deviations are referred to as clock jitter [37, 38]. The clock jitter is classified into two types: cycle-to-cycle jitter and peak-to-peak jitter.

Cycle-to-cycle jitter refers to *random* deviations of the clock phase from the ideal timing, as illustrated in Fig. 1.13 [39]. Deviation from the ideal phase at one edge of a clock signal is independent of the deviations at other edges. That is, the cycle-to-cycle jitter characterizes the variation of the time interval between two adjacent clock edges. The average cycle-to-cycle jitter asymptotically approaches zero with an increasing number of samples. This type of jitter is therefore characterized by a mean square deviation. This type of phase variation is produced by disturbances of duration shorter or comparable to the clock period. Active device noise and high frequency power supply noise (i.e., of a frequency higher or comparable to the clock frequency) contribute to the cycle-to-cycle jitter. Due to the stochastic nature of phase variations, the cycle-to-cycle jitter directly contributes to the uncertainty of the time reference signals across an integrated circuit. Increased uncertainty of an on-chip timing reference results in a reduced operating frequency [39].

The second type of jitter, peak-to-peak jitter, refers to *systematic* variations of on-chip clock phase *as compared to the system clock*. Consider a situation where several consecutive edges of an on-chip clock signal have a positive cycle-to-cycle variation, i.e., several consecutive clock cycles are longer than the ideal clock period, as illustrated in Fig. 1.14 (due to, for example, prolonged power supply variations from the nominal voltage). The timing requirements of the on-chip circuits are not violated provided that the cycle-to-cycle jitter is sufficiently small. The phase difference between the system clock and the on-chip clock,

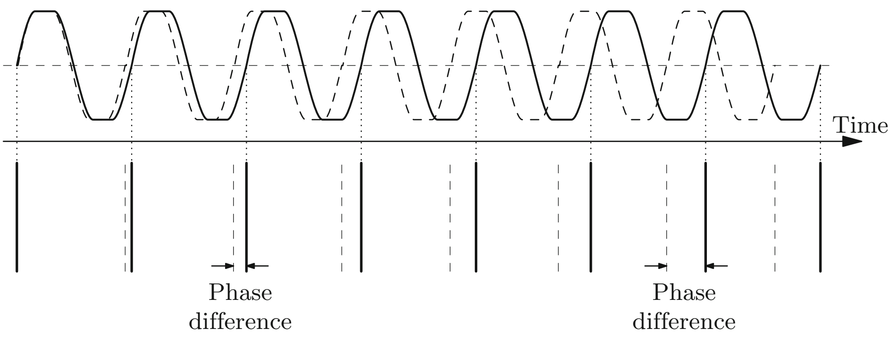

Fig. 1.14 Peak-to-peak jitter of a clock signal. The period of the clock signal (the *solid line*) systematically deviates from the period of the reference clock (the *dashed line*), leading to accumulation of the phase difference

however, accumulates with time. Provided the disturbance persists, the phase difference between the system and the on-chip clocks can accumulate for tens or hundreds of clock cycles, until the PLL feedback adjustment becomes effective. This phase difference degrades the synchronization among different clock domains (i.e., between one portion of an integrated circuit and other system components controlled by different clock signals). Synchronizing the clock domains is critical for reliable communication across these domains. The maximum phase difference between two clock domains is characterized by the peak-to-peak jitter.

The feedback response time is highly sensitive to the power supply voltage [40]. For example, the PLL designed for the 400 MHz IBM S/390 microprocessor exhibits a response time of approximately 50 clock cycles when operating at a 2.5 V power supply and disturbed by a 100 mV drop in supply voltage. The recovery time from the same disturbance increases manyfold when the supply voltage is reduced to 2.3 V and below [40].

1.4.3 Noise Margin Degradation

In digital logic styles with single-ended signaling, the power and ground delivery system also serves as a voltage reference for the on-chip signals. If a transmitter communicates a low voltage state, the output of the transmitter is connected to the ground distribution network. Alternatively, the output is connected to the power distribution network to communicate the high voltage state. At the receiver end of the communication line, the output voltage of the transmitter is compared to the power or ground voltage *local to the receiver*. Spatial variations in the power supply voltage create a discrepancy between the power and ground voltage levels at the transmitter and receiver ends of the communication line. The power noise induced uncertainty in these reference voltages degrades the noise margins of the on-chip signals. As the operating speed of integrated circuits has risen, crosstalk noise among on-chip signals has also increased. Providing sufficient noise margins of the on-chip signals is therefore a design issue of paramount importance.

1.4.4 Degradation of Gate Oxide Reliability

The performance characteristics of an MOS transistor depend on the thickness of the gate oxide. The current drive of the transistor increases as the gate oxide thickness is reduced, improving the speed and power characteristics. Reduction of the gate oxide thickness in process scaling has therefore been instrumental in improving transistor performance. A thin oxide layer, however, poses the problems of electron tunneling and oxide layer reliability [41]. As the thickness of the gate silicon oxide has reached several molecular layers (tens of angstroms) in contemporary digital CMOS processes, the power supply voltage is limited by the maximum electric field across the gate oxide layer [35]. Variations in the power supply voltage can increase the voltage applied across the ultra-thin gate oxide layer above the nominal power supply, degrading the long term reliability of the semiconducting devices [42]. Overshoots of the power and ground voltages should be limited to avoid significant degradation in the transistor reliability characteristics.

1.5 Summary

A historical background, general motivation, and relevant aspects related to integrated circuits in general and on-chip power networks in particular are presented in this introductory chapter. This chapter is summarized as follows.

- The development of integrated circuits has rapidly progressed after the first planar circuit—a “unitary circuit”
- Current microprocessors integrate many billions of transistors on a single monolithic substrate
- The clock frequency of modern microprocessors is in the range of several gigahertz
- The power consumption of mobile, notepad/desktop, and supercomputing microprocessor-based server farms, respectively, are in the range of a few watts, several hundreds of watts, and millions of watts.
- Different design criteria for integrated circuits have evolved over the past several decades with changing technology and application characteristics
- The issue of effective power delivery is fundamental to the successful operation of high complexity ICs. As current demand requirements have increased, voltage margins have been reduced, constraining the impedance of the power delivery system

- Voltage fluctuations within the power delivery system are causing a variety of problems, such as signal delay uncertainty, clock jitter, smaller noise margins, and reliability concerns due to degradation of the gate oxide
- Point-of-load power delivery is fundamental to maintain high quality of power as current densities and parasitic impedances have increased
- To support heterogeneous dynamically on-chip controlled systems, power resources should be intelligently managed in real-time