14

Multiple On-Chip Power Supply Systems

With recent developments in nanometer CMOS technologies, excessive power dissipation has become a limiting factor in integrating a greater number of transistors onto a single monolithic substrate. With the introduction of systems-on-chip and systems-in-package (SiP) technologies, the problem of heat removal has further worsened. Unless power consumption is dramatically reduced, packaging and performance of ultra large scale integration (ULSI) circuits will become fundamentally limited by heat dissipation.

Another driving factor behind the push for low power circuits is the growing market for portable electronic devices, such as PDAs, wireless communications, and imaging systems that demand high speed computation and complex functionality while dissipating as little power as possible [293]. Design techniques and methodologies for reducing the power consumed by an IC while providing high speed and high complexity systems are therefore required. These design technologies will support the continued scaling of the minimum feature size, permitting the integration of a greater number of transistors onto a single monolithic substrate.

The most effective way to reduce power consumption is to lower the supply voltage. Dynamic power currently dominates the total power dissipation, quadratically decreasing with supply voltage [294]. Reducing the supply voltage, however, increases the circuit delay. Chandrakasan *et al.* demonstrated in [295] that the increased delay can be compensated by shortening the critical paths using behavioral transformations such as parallelization and pipelining. The resulting circuit consumes less average power while satisfying global throughput constraints; albeit, at the cost of increased circuit area [296].

Power consumption can also be reduced by scaling the threshold voltage while simultaneously reducing the power supply [297]. This approach, however, results in significantly increased standby leakage current. To limit the leakage current during sleep mode, several techniques have been proposed, such as multi-threshold voltage CMOS [298], [299], variable threshold voltage schemes [300], [301], and circuits with an additional transistor behaving as a sleep switch [302]. These techniques, however, require additional process steps and/or additional circuitry to control the substrate bias or switch off portions of the circuit [301].

The total power dissipation can also be reduced by utilizing multiple power supply voltages [299], [303], [304]. In this scheme, a reduced voltage V_{dd}^L is applied to the non-critical paths, while a higher voltage V_{dd}^H is provided to the critical paths so as to achieve the specified delay constraints [299]. Multi-voltage schemes result in reduced total power without degrading the overall circuit performance. Multiple on-chip power supply systems are the subject of this chapter. Various circuit techniques exploiting multiple power supply voltages are presented in Section 14.1. Challenges to ICs with multiple supply voltages are discussed in Section 14.2. Choosing the optimum number and magnitude of the multi-voltage power supplies is discussed in Section 14.3. Some conclusions are offered in Section 14.4.

14.1 ICs with multiple power supply voltages

The strategy of exploiting multiple power supply voltages consists of two steps. Those logic gates with excessive slack (the difference between the required time and the arrival time of a signal) is first determined. A reduced supply voltage V_{dd}^L is provided to those gates to reduce power. Note that in most practical applications, the number of critical paths is only a small portion of the total number of paths in a circuit. Excess slack therefore exists in the majority of paths within a circuit. Determining those gates with excessive time slack is therefore an important and complex task [299]. A variety of computer-aided design (CAD) algorithms and tools have been developed to evaluate the delay characteristics of high complexity ICs such as microprocessors [305], [306]. Multi-voltage low power techniques are reviewed in this section. A low power technique with multiple power supply voltages is presented in Section 14.1.1. Clustered voltage scaling (CVS) is presented in

Section 14.1.2. Extended clustered voltage scaling (ECVS) is discussed in Section 14.1.3.

14.1.1 Multiple power supply voltage techniques

A critical delay path between flip flops FF_1 and FF_2 in a single supply voltage, synchronous circuit is shown in Fig. 14.1. Since the excessive slack remains in those paths located off the critical path, timing constraints are satisfied if the gates in the non-critical paths use a reduced supply voltage V_{dd}^L. A dual supply voltage circuit in which the original power supply voltage V_{dd}^H of each of the gates along the non-critical delay paths is replaced by a lower supply voltage V_{dd}^L is illustrated in Fig. 14.2. If a low voltage supply is available, the gates with V_{dd}^L can be selected to reduce the overall power using conventional algorithms such as gate resizing [307].
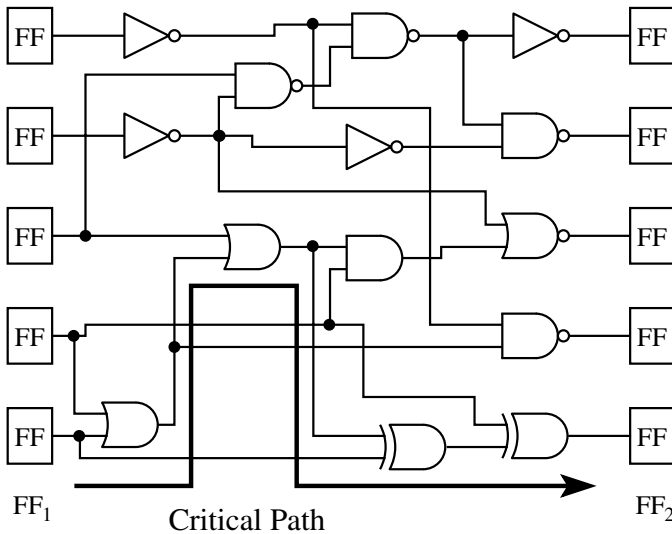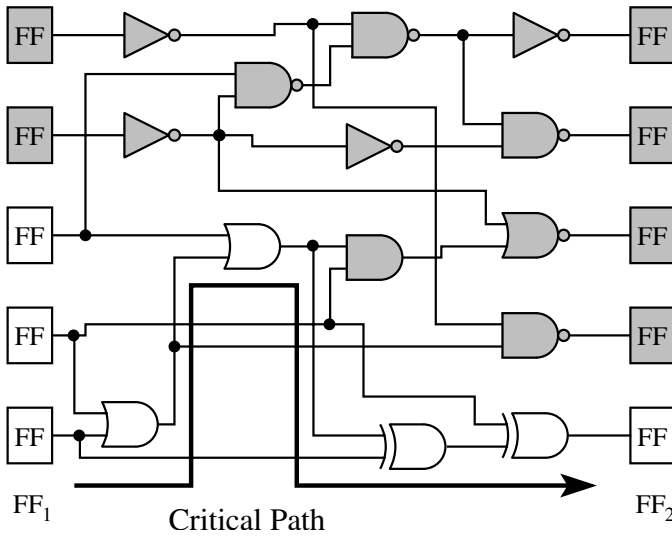

Fig. 14.1. An example single supply voltage circuit.

A circuit with multiple power supply voltages, however, can result in DC current flowing in a high voltage gate due to the direct connection between a low voltage gate and a high voltage gate. If a gate with a

reduced supply voltage is directly connected to a gate with the original supply voltage, the "high" level voltage at node A is not sufficiently high to turn off the PMOS device in a CMOS circuit, as shown in Fig. 14.3. The PMOS device in the high voltage gate is therefore weakly "ON," conducting static current from the power supply to ground. These static currents significantly increase the overall power consumed by an IC, wasting the savings in power achieved by utilizing a multi-voltage power distribution system.

Level converters are typically inserted at node A to remove the static current path [308]. A simple level converter circuit is illustrated in Fig. 14.4. The level converter restores the full voltage swing from V_{dd}^{L} to V_{dd}^{H}. Note that a great number of level converters is typically required, increasing the area and power overhead. The problem of utilizing a dual power supply voltage scheme is formulated as follows.

Problem formulation: For a given circuit, determine the gates and registers to which a reduced power supply voltage V_{dd}^{L} should be applied such that the overall power and number of level converters are minimized while satisfying system-level timing constraints [309].

Fig. 14.2. An example dual supply voltage circuit. The gates operating at a lower power supply voltage V_{dd}^{L} (located off the critical delay path) are shaded.
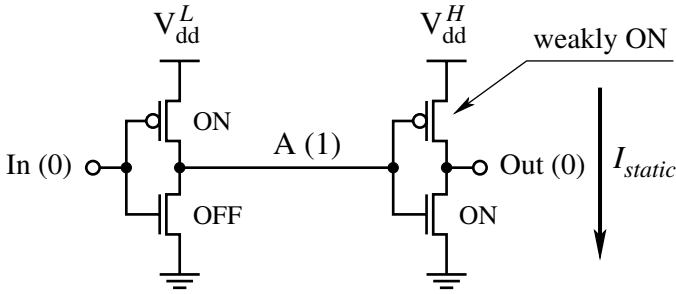
Fig. 14.3. Static current as a result of a direct connection between the V_{dd}^L gate and the V_{dd}^H gate.
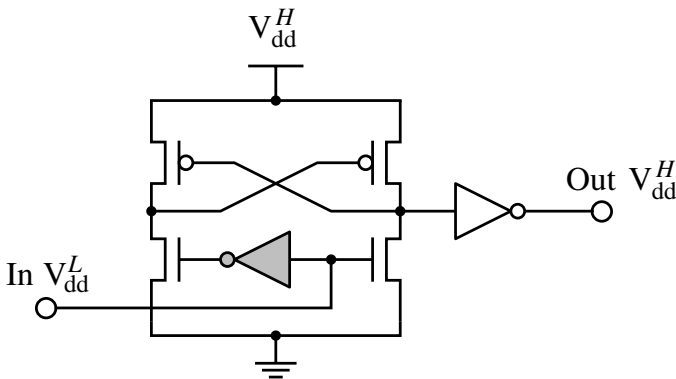

Fig. 14.4. Level converter circuit. The inverter operating at the reduced power supply voltage V_{dd}^L is shown in grey.

14.1.2 Clustered voltage scaling (CVS)

The number of level converters can be reduced by minimizing the connections between the V_{dd}^L gates and the V_{dd}^H gates. The CVS technique, proposed in [310], results in a circuit structure with a greatly reduced number of level converters, as shown in Fig. 14.5.

To avoid inserting level converters, the CVS technique exploits the specific connectivity patterns among the gates, such as a connection between V_{dd}^H gates, between V_{dd}^L gates, and between a V_{dd}^H gate and a V_{dd}^L gate. These connections do not require level converters to remove any static current paths. Level converters are only required at the interface between the output of a V_{dd}^L gate and the input of a V_{dd}^H gate. The number of required level converters in the CVS structure shown in

Fig. 14.5 is almost the same as the number of V_{dd}^L flip flops. The CVS technique therefore results in fewer level converters, reducing the overall power consumed by an integrated circuit.
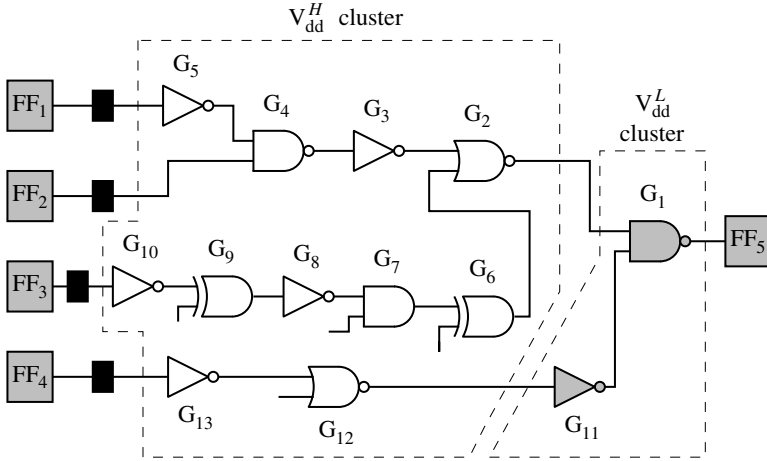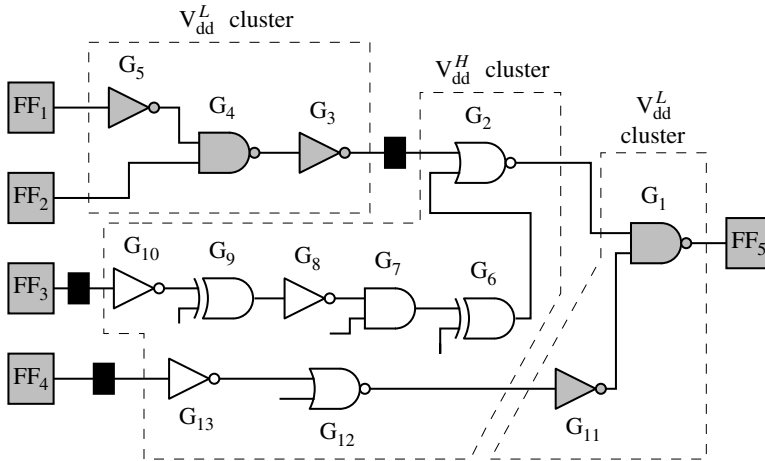

Fig. 14.5. A dual power supply voltage circuit with the clustered voltage scaling (CVS) technique [310]. The gates operating at a lower supply voltage are shaded. The level converters are shown as black rectangles.

14.1.3 Extended clustered voltage scaling (ECVS)

The number of gates with a lower power supply voltage can be increased by optimally choosing the insertion points of the level converters, further reducing overall power. As an example, in the CVS structure shown in Fig. 14.5, the path delay from flip flop FF_3 to gate G_2 is longer than the delay from FF_1 to G_2. Moreover, applying a lower power supply to gate G_2 can produce a timing violation. A high power supply should therefore be provided to G_2. From CVS connectivity patterns described in Section 14.1.2, note that G_3 also has to be supplied with V_{dd}^H. Alternatively, in a CVS structure, G_3 cannot be supplied with V_{dd}^L although excessive slack remains in the path from FF_1 to G_2. Similarly, G_4 and G_5 should be connected to V_{dd}^H to satisfy existing timing constraints. If the insertion point of the level converter adjacent to FF_1 is moved to the interface between G_3 and G_2, gates G_3, G_4, and G_5 can be

connected to V_{dd}^L, as illustrated in Fig. 14.6. Note that the structure shown in Fig. 14.6 is obtained from the CVS network by relaxing existing limitations on the insertion positions of the level converters. Such a technique is often referred to as the extended clustered voltage scaling technique [309], [311].

Fig. 14.6. A dual power supply voltage circuit with the extended clustered voltage scaling (ECVS) technique [309]. The gates operating at a lower supply voltage are shaded. The level converters are shown as black rectangles.

14.2 Challenges in ICs with multiple power supply voltages

The application of power reduction techniques with multiple supply voltages in modern high performance ICs is a challenging task. Circuit scheduling algorithms require complex computations, limiting the application of CVS and ECVS techniques to specific paths within an IC. Primary challenges of multi-voltage power reduction schemes are discussed in this section. The issues of area overhead and related trade-offs are introduced in Section 14.2.1. Power penalties are presented in Section 14.2.2. The additional design complexity associated with level converters and integrated DC–DC voltage converters is discussed in

Section 14.2.3. Several placement and routing strategies are described in Section 14.2.4.

14.2.1 Die area

As described in Section 14.1, level converter circuits are inserted at the interface between specific gates in power reduction schemes with multiple power supply voltages to reduce static current. Multi-voltage circuits require additional power connections, significantly increasing routing complexity and die area. Additional area results in greater parasitic capacitance of the signal lines, increasing the dynamic power consumed by an IC. As a result of the increased area, the time slack in the critical paths is often significantly smaller, reducing the power savings of a multi-voltage scheme. A tradeoff therefore exists between the power savings and area overhead in ICs with multiple power supply voltages. The critical paths should therefore be carefully determined in order to reduce the overall circuit power.

14.2.2 Power dissipation

Multi-voltage low power techniques require the insertion of level converters to reduce static current. The number of level converters depends upon the connectivity patterns at the interface between each critical and non-critical path. Improper scheduling of the critical paths can lead to an excessive number of level converters, increasing the power. The ECVS technique with relaxed constraints for level converters should therefore be used, resulting in a smaller number of level converters.

   Note that the magnitude of the overall reduction in power is determined by the number and voltage of the available power supply voltages, as discussed in Section 14.3. It is therefore important to determine the optimum number and magnitude of the power supply voltages to maximize any savings in power. Also note that lower power supply voltages are often generated on-chip from a high voltage power supply using DC−DC voltage converters [312], [313]. The power and area penalties of the on-chip DC−DC voltage converters should therefore be considered to accurately estimate any savings in power.

   Several primary factors, such as physical area, the number and magnitude of the power supply voltages, and the number of level converters contribute to the overall power overhead of any multi-voltage low power

technique. Complex multi-variable optimization is thus required to determine the proper system parameters in order to achieve the greatest reduction in overall power [314].

14.2.3 Design complexity

Note that while significantly reducing power, a multiple power supply voltage scheme results in significantly increased design complexity. The complexity overhead of a multi-voltage low power technique is due to two aspects. The level converters not only dissipate power, but also dramatically increase the complexity of the overall design process. A level converter typically consists of both low voltage and high voltage gates, increasing the area and routing resources. Multiple level converters also increase the delay of the critical paths. High speed, low power level converters are therefore required to achieve a significant reduction in overall power while satisfying existing timing constraints [308], [315]. Standard logic gates with embedded level conversion as reported in [315] support the design of circuits without the addition of level converters, substantially reducing power, area, and complexity.

Monolithic DC – DC voltage converters are often integrated on-chip to enhance overall energy efficiency, improve the quality of the voltage regulation, decrease the number of I/O pads dedicated to power delivery, and reduce fabrication costs [316]. To lower the energy dissipated by the parasitic impedance of the circuit board interconnect, the passive components of a low frequency filter (*e.g.*, the filter inductor and filter capacitor) are also placed on-chip, significantly increasing both the required area and design complexity. A great amount of on-chip decoupling capacitance is also often required to improve the quality of the on-chip power supply voltages [317]. The area and power penalty as well as the increased design complexity of the additional on-chip voltage converters should therefore be considered when determining the optimal number and magnitude of the multiple power supply voltages.

14.2.4 Placement and routing

To achieve the full benefit offered by multiple power supply voltage techniques, various design issues at both the high level and physical level should be simultaneously considered. Existing electronic design automation (EDA) placement and routing tools for conventional circuits with single power supply voltages, however, cannot be directly

applied to low power techniques with multiple power supply voltages. Specific CAD tools, capable of placing and routing physical circuits with multiple power supplies based on high level gate assignment information, are therefore required. The placement and routing of ICs with multiple power supply voltages is a complex problem. Three widely utilized layout schemes are described in this section.
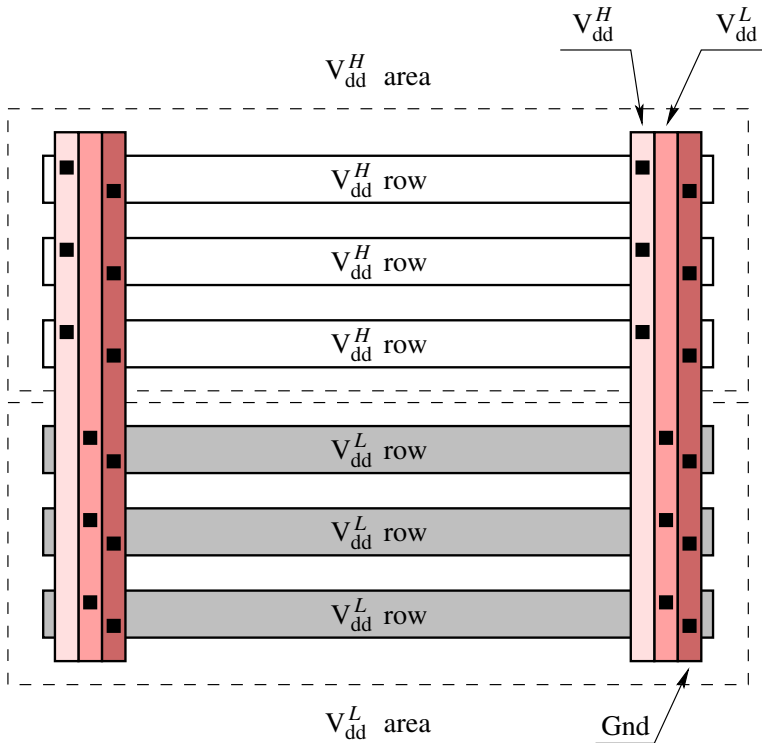
Area-by-area architecture

The simplest architecture for a circuit with dual power supply voltages is an area-by-area architecture [309], as shown in Fig. 14.7. In this architecture, the V_{dd}^L cells are placed in one area, while the V_{dd}^H cells are placed in a different area. The area-by-area technique iteratively generates a layout with existing placement and routing tools using one of the available power supply voltages. This architecture, however, results in a degradation in performance due to the substantially increased interconnect length between the V_{dd}^L and V_{dd}^H cells.

Row-by-row architecture

The layout architecture proposed in [318] is illustrated in Fig. 14.8. In this architecture, the V_{dd}^L cells and V_{dd}^H cells are placed in different rows. Each row only consists of V_{dd}^L cells *or* V_{dd}^H cells. This layout technique is therefore a row-by-row architecture. Note that in this architecture, a V_{dd}^L row is placed next to a V_{dd}^H row, reducing the interconnect length between the V_{dd}^L cells and the V_{dd}^H cells. The performance of a row-by-row layout architecture is therefore higher as compared to the performance of an area-by-area architecture. The row-by-row technique also results in smaller area, further improving system performance. Another advantage of this technique is that an original V_{dd}^H cell library can be used for the V_{dd}^L cells. Since the layout of the V_{dd}^L cells are the same as those of the V_{dd}^H cells, the original layout of the V_{dd}^H cells can be treated as V_{dd}^L cells. A lower power supply voltage can be provided to the V_{dd}^L cells.
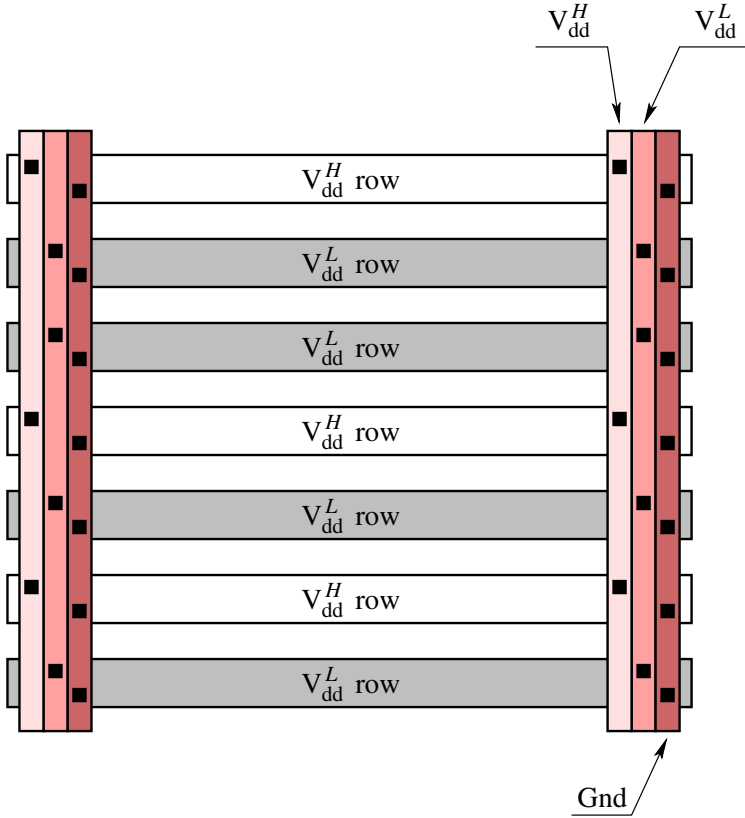
In-row architecture

An improved row-by-row layout architecture is presented in [319]. This architecture is based on a modified cell library [319]. Unlike conventional standard cells, the new standard cell has two power rails and

Fig. 14.7. Layout of an area-by-area architecture with a dual power supply voltage. In this architecture, the V_{dd}^L cells are placed in one area, while the V_{dd}^H cells are separately placed in a different area.

one ground rail. One of the power rails is connected to V_{dd}^L and the other power rail is connected to V_{dd}^H. The modified library supports the allocation of both V_{dd}^L cells and V_{dd}^H cells within the same row, as shown in Fig. 14.9. This layout scheme is therefore referred to as an in-row architecture. Note that the width of the power and ground lines in each cell is reduced, slightly increasing the overall area (a 2.7% area overhead as compared to the original cell) [319]. Since the number of V_{dd}^L cells is typically greater than the number of V_{dd}^H cells, the lower power supply provides higher current. The low voltage power rail is therefore wider than the high voltage power rail to maintain a similar voltage drop within each power rail. Note that the in-row architecture results in a significant reduction in the interconnect length between the V_{dd}^L and V_{dd}^H cells, as compared to a row-by-row scheme [319]. An

Fig. 14.8. Layout of a row-by-row architecture with a dual power supply voltage. In this architecture, the V_{dd}^L cells and V_{dd}^H cells are placed in different rows. Each row consists of only V_{dd}^L cells *or* V_{dd}^H cells.

in-row layout scheme should therefore be utilized in high performance, high complexity ICs to reduce overall power with minimal area and complexity penalties.

14.3 Optimum number and magnitude of available power supply voltages

In low power techniques with multiple power supply voltages, the power reduction is primarily determined by the number and magnitude of the available power supply voltages. The trend in power reduction with a
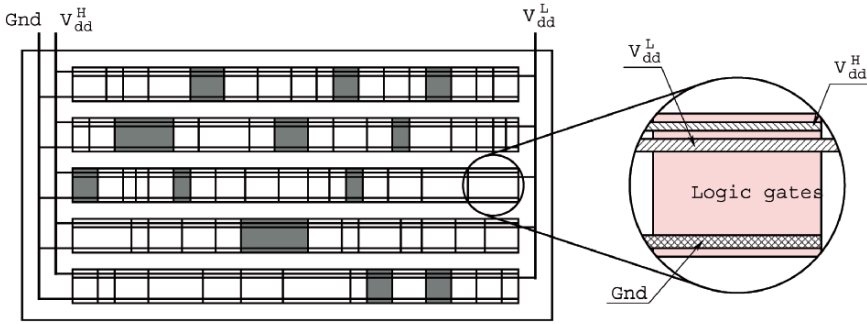
Fig. 14.9. In-row dual power supply voltage scheme. This architecture is based on a modified cell library with two power rails and one ground rail in each cell. The V_{dd}^H cells are shown in grey and the V_{dd}^L cells are white.

multi-voltage scheme as a function of the number of available supply voltages is illustrated in Fig. 14.10. Observe from Fig. 14.10 that if fewer power supplies than the optimum number are available ($n < n_{opt}$), the savings in power can be fairly small. The maximum power savings is achieved with the number of supply voltages close to the optimum number (represented by region $n = n_{opt}$ in Fig. 14.10). If more than the optimum number of power supplies are used, the savings in power becomes smaller, as depicted in Fig. 14.10 for $n > n_{opt}$. This decline in power reduction when the number of supply voltages is greater than the optimum number is due to the increased overhead of the additional power supplies (as a result of the increased area, number of level converters, and design complexity). Any savings in power is also constrained by the magnitude of the available power supplies. A tradeoff therefore exists between the number and magnitude of the available power supplies and the achievable power savings. A methodology is therefore required to estimate the optimum number and magnitude of the available power supply voltages in order to produce the greatest reduction in power. Design techniques for determining the optimum number and magnitude of the available power supplies are the subject of this section.

In systems with multiple power supply voltages (where $V_1 > V_2 > \cdots > V_n$), the power dissipation is [320]

$$P_n = f \left\{ \left(C_1 - \sum_{i=2}^{n} C_i \right) V_1^2 + \sum_{i=2}^{n} C_i V_i^2 \right\}, \qquad (14.1)$$
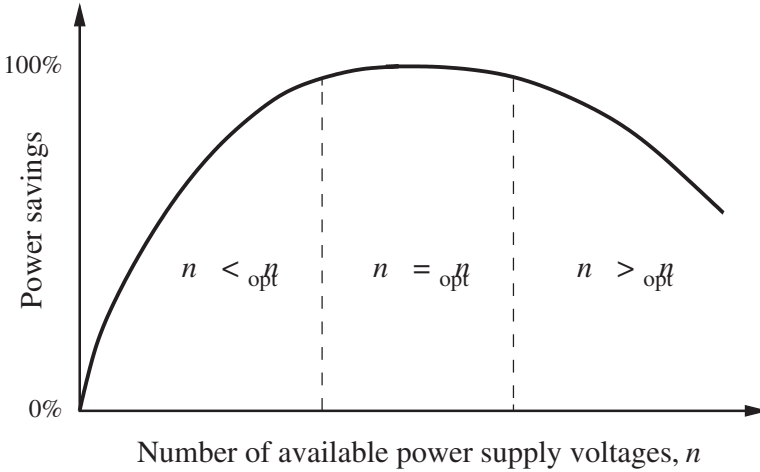
Fig. 14.10. Trend in power reduction with multi-voltage scheme as a function of the number of available supply voltages.

where C_i is the total capacitance of the logic gates and interconnects operating at a reduced supply voltage V_i and f is the operating frequency. The ratio of the power dissipated by a system with multiple power supply voltages as compared to the power dissipation in a single power supply system is

$$K_{V_{\mathrm{dd}}} \equiv \frac{P_n}{P_1} = 1 - \sum_{i=2}^{n} \left[\left(\frac{C_i}{C_1} \right) \left\{ 1 - \left(\frac{V_i}{V_1} \right)^2 \right\} \right]. \qquad (14.2)$$

Since delay is proportional to the total capacitance, $\dfrac{C_i}{C_1}$ is

$$\frac{C_i}{C_1} = \frac{\int\limits_{0}^{1} p(t)\, t_i\, dt}{\int\limits_{0}^{1} p(t)\, t\, dt}, \qquad (14.3)$$

where $p(t)$ is the normalized path delay distribution function and t_i is the total delay of the circuits operating at V_i. For a path with a total delay $t_{i,0} < t < t_{i-1,0}$, where $t_{i,0}$ denotes the path delay at V_1 (equal to the cycle time when all of the circuits operate at V_i), the power dissipation is minimum when (V_i, V_{i-1}) are applied. In this case, t_i is

$$t_i = \begin{cases} \dfrac{t_{i,0}}{t_{i,0} - t_{i+1,0}}(t - t_{i+1,0}) & : \quad t_{i+1,0} \le t \le t_{i,0} \\ \dfrac{t_{i,0}}{t_{i-1,0} - t_{i,0}}(t_{i-1,0} - t) & : \quad t_{i,0} \le t \le t_{i-1,0}, \end{cases} \quad (14.4)$$

where $t_{i,0}$ is

$$t_{i,0} = \left(\frac{V_1}{V_i}\right)\left(\frac{V_i - V_{\text{th}}}{V_1 - V_{\text{th}}}\right)^{\alpha}, \quad (14.5)$$

V_{th} is the threshold voltage, and α is the velocity saturation index [321]. Note that $t_{n+1,0} = 0$. $K_{V_{\text{dd}}}$ can be determined from (14.1) – (14.5) for a specific $p(t)$, V_1, V_i, and V_{th}.

For a lambda-shaped normalized path delay distribution function $p(t)$ (see Fig. 14.11) as determined from post-layout static timing analysis, approximate rules of thumb for determining the optimum magnitude of the power supply voltages have been determined by Hamada *et al.* [320],

$$\text{for } \{V_1, V_2\} \qquad \frac{V_2}{V_1} = 0.5 + 0.5\frac{V_{\text{th}}}{V_1}, \qquad (14.6)$$

$$\text{for } \{V_1, V_2, V_3\} \qquad \frac{V_2}{V_1} = \frac{V_3}{V_2} = 0.6 + 0.4\frac{V_{\text{th}}}{V_1}, \qquad (14.7)$$

$$\text{for } \{V_1, V_2, V_3, V_4\} \qquad \frac{V_2}{V_1} = \frac{V_3}{V_2} = \frac{V_4}{V_3} = 0.7 + 0.3\frac{V_{\text{th}}}{V_1}. \quad (14.8)$$

Criteria (14.6) – (14.8) can be used to determine the magnitude of each power supply voltage based on the total number of available power supply voltages. Note that these rules of thumb result in the optimum power supply voltages where the maximum difference in power reduction is less than 1% as compared to the absolute minimum (as determined from an analytic solution of the system of equations).

Note again that if a greater number of power supplies is used, the total power can be further reduced, reaching a constant power level at some number of power supplies (see Fig. 14.10). As determined in [320], up to three power supply voltages should be utilized to reduce the power consumed by an IC. The reduction in power diminishes as the power supply voltage is scaled and $\frac{V_{\text{th}}}{V_{\text{dd}}}$ increases.

A rule of thumb for two power supply voltages has been evaluated by simulations in [309]. For $V_{\text{dd}}^H = 3.3$ volts, a V_{dd}^L of 1.9 volts is estimated, exhibiting good agreement with (14.6). The dependence of the total power of a dual power supply media processor as a function of the lower
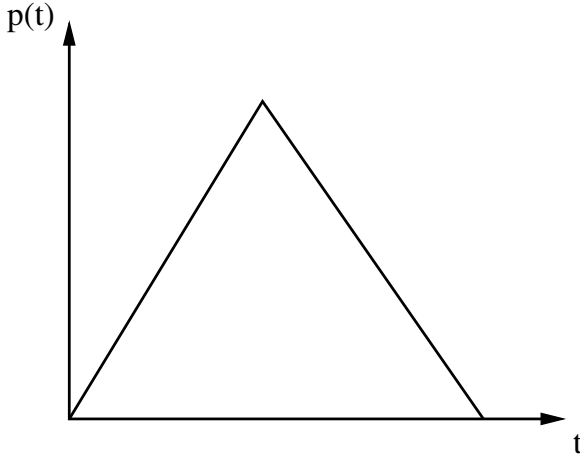
Fig. 14.11. A lambda-shaped normalized path delay distribution function.

power supply V_{dd}^L is depicted in Fig. 14.12. Observe from Fig. 14.12 that the minimum overall power is achieved at $V_{dd}^L = 1.9$ volts.

The minimum overall power of a dual power supply system can be explained as follows. In a dual power supply system, the power reduction is determined by two factors: the reduction in power of a single logic gate due to scaling the power supply voltage from V_{dd}^H to V_{dd}^L, and the number of original V_{dd}^H gates replaced with V_{dd}^L gates. At lower V_{dd}^L, the power dissipated by a V_{dd}^L gate decreases, while the number of original V_{dd}^H gates replaced with V_{dd}^L gates is reduced. This behavior is due to the degradation in performance of the V_{dd}^L gates at a lower V_{dd}^L. As a result, fewer gates can be replaced with lower voltage gates without violating existing timing constraints. Conversely, at a higher V_{dd}^L, the number of gates replaced with V_{dd}^L gates increases, while the reduced power in a single V_{dd}^L gate decreases. The overall power therefore has a minimum at a specific V_{dd}^L voltage, as shown in Fig. 14.12.

Low power techniques with multiple power supply voltages and a single fixed threshold voltage have been discussed in this chapter. Enhanced results are achieved by simultaneously scaling the multiple threshold voltages and the power supply voltages [299], [322], [323]. This approach results in reduced total power with low leakage currents. The total power can also be lowered by simultaneously assigning threshold voltages during gate sizing. Nguyen *et al.* [324] demonstrated power
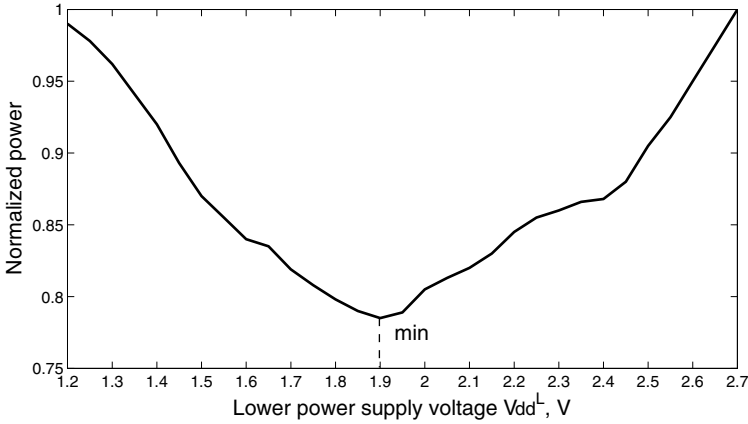
Fig. 14.12. Dependence of the total power of a dual power supply system on a lower power supply voltage V_{dd}^L [309]. The original high power supply voltage $V_{dd}^H = 3.3$ volts.

reductions approaching 32% on average (57% maximum) for the IS-CAS85 benchmark circuits. CVS with variable supply voltage schemes has been presented in [325]. In this scheme, the power supply voltage is gradually scaled based on an accurate model of the critical path delay. Up to a 70% power savings has been achieved as compared to the same circuit without these low power techniques. In [326], a column-based dynamic power supply has been integrated into a high frequency SRAM circuit. The power supply voltage is adaptively changed based on the read/write mode of the SRAM, reducing the total power.

As described in this chapter, power dissipation has become a major factor, limiting the performance of high complexity ICs. Multiple low power techniques should therefore be utilized to achieve significant power savings in modern nanoscale ICs.

14.4 Summary

The discussion of multiple on-chip power supply systems and different low power design techniques can be summarized as follows:

- The total power consumed by an IC can be reduced by utilizing multiple power supply voltages

- In multi-voltage low power techniques, a lower power supply voltage is applied to those logic gates with excessive slack to reduce power consumption

- In a multi-voltage scheme, the gates and flip flops with a lower power supply voltage should be determined such that the overall power and number of level converters are minimized while satisfying existing timing constraints

- CVS and ECVS techniques exploit specific connectivity patterns, reducing the number of level converters

- Various penalties, such as area, power, and design complexity, should be considered during the system design process so as to maximize the savings in power

- The in-row layout scheme reduces overall power with minimum area and design complexity

- A maximum of two or three supply voltages should be employed in low power applications

- Rules of thumb have been described for determining the optimum magnitude of the multiple power supply voltages

- A greater savings in power can be achieved by simultaneously scaling the multiple threshold voltages and power supply voltages