Unit V

Leakage Power Minimization

9.1 Introduction

Due to aggressive device-size scaling, the very-large-scale integration (VLSI) technology has moved from the millimetre to nanometre era by providing increas-ingly higher performance along the way. Performance improvement has been con-tinuously achieved primarily because of the gradual decrease of gate capacitances. However, as the supply voltage must continue to scale with device-size scaling to maintain a constant field, the threshold voltage of the metal–oxide–semiconductor (MOS) transistors should also be scaled at the same rate to maintain gate overdrive (V_{cc}/V_t) and hence performance. Unfortunately, the reduction of V_t leads to an exponential increase in the subthreshold leakage current. As a consequence, the leakage power dissipation has gradually become a significant portion of the total power dissipation. For example, for a 90-nm technology, the leakage power is 42 % of the total power and for a 65-nm technology, the leakage power is 52 % of the total power. This has led to vigorous research work to develop suitable approaches for leakage power minimization.

9.2 Fabrication of Multiple Threshold Voltages

The present-day process technology allows the fabrication of metal–oxide–semi-conductor field-effect transistors (MOSFETs) of multiple threshold voltages on a single chip. This has opened up the scope for using dual-V_t CMOS circuits to re-alize high-performance and low-power CMOS circuits. The basic idea is to use high-V_t transistors to reduce leakage current and low-V_t transistors to achieve high performance. Before we discuss dual-V_t (or multiple V_t) circuit design techniques, we shall explore various fabrication techniques [1] used for implementing multiple threshold voltages in a single chip.

9.2.1 Multiple Channel Doping

The most commonly used technique for realizing multiple-V_T MOSFETs is to use different channel-doping densities based on the following expression:

$$V_{th} = V_{fb} + 2\tau_B + \frac{\sqrt{2\varepsilon_s q \cdot Na(2\tau_B + V_{bs})}}{C_{ox}}, \qquad (9.1)$$

where V_{fb} is the flat-band voltage, N_a is the doping density in the substrate, and $\tau_B = kT / q\ (L_x (N_a / x))$.

A higher doping density results in a higher threshold voltage. However, to fabricate two types of transistors with different threshold voltages, two additional masks are required compared to the conventional single-V_t fabrication process. This makes the dual-V_t fabrication costlier than single-V_t fab-rication technology. Moreover, due to the non-uniform distribution of the doping density, it may be difficult to achieve dual threshold voltage when these are very close to each other.

9.2.2 Multiple Oxide CMOS

The expression for the threshold voltage shows a strong dependence on the value of C_{ox}, the unit gate capacitance. Different gate capacitances can be realized by us-ing different gate oxide thicknesses. The variation of threshold voltage with oxide thickness (t_{ox}) for a 0.25-μm device is shown in Fig. 9.3. Dual-V_{th} MOSFETs can be realized by depositing two different oxide thicknesses. A lower gate capacitance due to higher oxide thickness not only reduces subthreshold leakage current but also provides the following benefits:

Reduced gate oxide tunnelling because the oxide tunnelling current exponentially decreases with the increase in oxide thickness. Reduced dynamic power dissipation due to reduced gate capacitance, because of higher gate oxide thickness.Although the increase in gate oxide thickness has the above benefits, it has some adverse effects due to an increase in short-channel ef-fect. For short-channel devices as the gate oxide thickness increases, the aspect ratio (AR), which is defined by AR = lateral dimension/vertical dimension, decreases:

$$\mathrm{AR} = \frac{L}{\left[t_{ox}\left(\frac{\varepsilon_{si}}{\varepsilon_{ox}}\right)\right]^{1/3} W_{dm}^{1/3} x_j^{1/3}}, \qquad (9.2)$$

where ε_{si} and ε_{ox} are silicon and oxide permittivities, L, t_{ox}, W_{dm}, and X_j are chan-nel length, gate oxide thickness, depletion depth, and junction depth, respectively.

9.2.3 Multiple Channel Length

In the case of short-channel devices, the threshold voltage decreases as the channel length is reduced, which is known as V_{th} roll-off. This phenomenon can be exploited to

realize transistors of dual threshold voltages. The variation of the threshold volt-age with channel length is shown in Fig. 9.5. However, for transistors with feature sizes close to 0.1 μm, halo techniques have to be used to suppress the short-channel effects. As the V_{th} roll-off becomes very sharp, it turns out to be a very difficult task to control the threshold voltage near the minimum feature size. For such technologies, longer channel lengths for higher V_{th} transistors increase the gate capacitance, which leads to more a dynamic power dissipation and delay.

9.2.4 Multiple Body Bias

The application of reverse body bias to the well-to-source junction leads to an in-crease in the threshold voltage due to the widening of the bulk depletion region, which is known as *body effect*. This effect can be utilized to realize MOSFETs hav-ing multiple threshold voltages. However, this necessitates separate body biases to be applied to different nMOS transistors, which means the transistors cannot share the same well. Therefore, costly triple-well technologies are to be used for this pur-pose. Another alternative is to use silicon-on-insulator (SoI) technology, where the devices are isolated naturally.

In order to get best of both the worlds, i.e. a smaller delay of low-V_t devices and a smaller power consumption of high-V_t devices, a balanced mix of both low-V_t and high-V_t devices may be used. The following two approaches can be used to reduce leakage power dissipation in the standby mode.

9.3 VTCMOS Approach

We have observed that low supply voltage along with low-threshold voltage provides a reduced overall power dissipation without a degradation in performance. However, the use of low-V_t transistors inevitably leads to increased subthreshold leakage current, which is of major concern when the circuit is in standby mode. In many recent applications, such as cell phones, personal digital assistants (PDAs), etc., a major part of the circuit remains in standby mode most of the time. If the standby current is not low, it will lead to a shorter battery life.

VTCMOS circuits make use of the body effect to reduce the subthreshold leakage current, when the circuit is in normal mode. We know that the threshold voltage is a function of the voltage difference between the source and the substrate. The substrate terminals of all the n-channel metal–oxide–semiconductor (nMOS) transistors are connected to the ground potential and the substrate terminals of all the p-channel metal–

oxide–semiconductor (pMOS) transistors are connected to V_{dd}, as shown in Fig. 9.6. This ensures that the source and drain diffusion regions always remain reversed-biased with respect to the substrate and the threshold voltages of the transistors are not significantly influenced by the body effect. On the other hand, in the case of VTCMOS circuits, the substrate bias voltages of nMOS and pMOS transistors are controlled with the help of a substrate bias control circuit, as shown in Fig. 9.7.

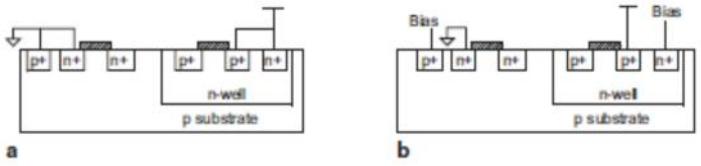


Fig. 9.6 Physical structure of a CMOS inverter **a** without body bias, **b** with body bias. *CMOS* complementary metal–oxide–semiconductor

Although the VTCMOS technique is a very effective technique for controlling threshold voltage and reducing subthreshold leakage current, it requires a twin-well or triple-well CMOS fabrication technology so that different substrate bias voltages can be applied to different parts of the chip. Separate power pins may also be required if the substrate bias voltage levels are not generated on chip. Usually, the additional area required for the substrate bias control circuitry is negligible compared to the overall chip area.

9.4 Transistor Stacking

When more than one transistor is in series in a CMOS circuit, the leakage current has a strong dependence on the number of turned off transistors. This is known as the *stack effect* [4–6]. The mechanism of the stack effect can be best understood by

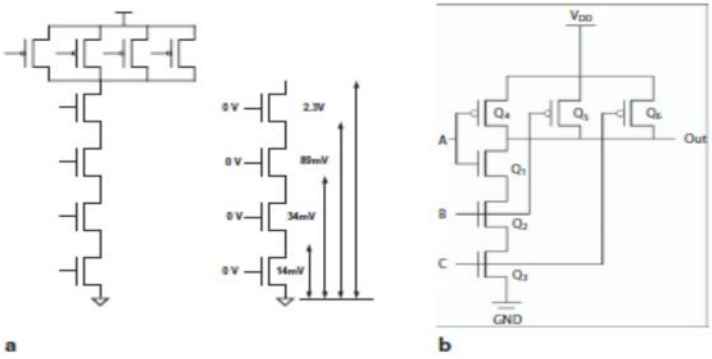


Fig. 9.8 **a** Source voltages of the nMOS transistors in the stack, **b** A 4-input NAND gate. *nMOS* n-channel metal–oxide–semiconductor

considering the case when all the transistors in a stack are turned off. Figure 9.8a shows four nMOS devices of a four-input NAND gate in a stack. The source and drain voltages of the MOS transistors obtained by simulation are shown in the figure. These voltages are due to a small drain current passing through the circuit. The source voltages of the three transistors on top of the stack have positive values. Assuming all gate voltages are equal to zero, the gate-to-source voltages of the three transistors are negative. Moreover, the drain-to-source potential of the MOS transistors is also reduced. The following three mechanisms come into play to reduce the leakage current:

i Due to the exponential dependence of the subthreshold current on gate-to-source voltage, the leakage current is greatly reduced because of negative gate-to-source voltages.

ii The leakage current is also reduced due to body effect, because the body of all the three transistors is reverse-biased with respect to the source.

iii As the source-to-drain voltages for all the transistors are reduced, the subthreshold current due to drain-induced barrier lowering (DIBL) effect will also be lesser. As a consequence, the leakage currents will be minimum when all the transistors are turned off, which happens when the input vector is 0000. The leakage cur-rent passing through the circuit depends on the input vectors applied to the gate and it will be different for different input vectors. For example, for a three-input NAND gate shown in Fig. 9.8b, the leakage current contributions for different input vectors are given in Table 9.1. It may be noted that the highest leakage current is 99 times the lowest leakage current. The current is lowest when all the transistors in series are OFF, whereas the leakage current is highest when all the transistors are ON.

Table 9.1 Input vectors and corresponding leakage currents for the three-input NAND gate

State (ABC)	Leakage current (nA)	Leaking transistors
000	0.095	Q_1, Q_2, Q_3
001	0.195	Q_1, Q_2
010	0.195	Q_1, Q_3
011	1.874	Q_1
100	0.184	Q_2, Q_3
101	1.220	Q_2
110	1.140	Q_3
111	9.410	Q_4, Q_5, Q_6

9.5 MTCMOS Approach

In this approach, MOSFETs with two different threshold voltages are used in a single chip. It uses two operational modes—*active* and *sleep* for efficient power management. A basic MTCMOS circuit scheme is shown in Fig. 9.9. The realization of a two-input NAND gate is shown in the figure. The CMOS logic gate is realized with transistors of low-threshold voltage of about 0.2–0.3 V. Instead of connecting the power terminal lines of the gate directly to the power supply lines V_{dd} and GND, here these are connected to the 'virtual' power supply lines (VDDV and GNDV). The real and virtual power supply lines are linked by the MOS transistor Q_1 and Q_2. These transistors have a high-threshold voltage in the range 0.5–0.6 V and serve as sleep control transistors. Sleep control signals SL and$\overline{SL}$ are connected to Q_1 and Q_2, respectively, and used for active/sleep mode control. In the active mode, when SL is set to LOW, both Q_1 and Q_2 are turned ON connecting the real power lines to VDDV and GNDV. In this mode, the NAND gate operates at a high speed cor-responding to the low-threshold voltage of 0.2 V, which is relatively low compared to the supply voltage of 1.0 V. In the sleep mode, SL is set to HIGH to turn both Q_1 and Q_2 OFF, thereby isolating the real supply lines from VDDV and GNDV. As the sleep transistors have a high-threshold voltage (0.6 V), the leakage current flowing through these two transistors will be significantly smaller in this mode. As a consequence, the leakage power consumption during the standby period can be dramatically reduced by sleep control.

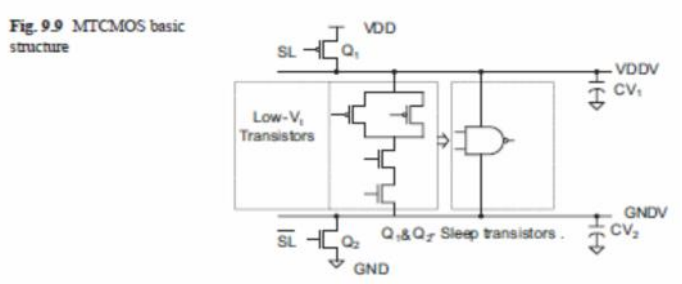


Fig. 9.9 MTCMOS basic structure

Two other factors that affect the speed performance of an MTCMOS circuit are: the width of the sleep control transistors and the capacitances of the virtual power line. The sleep transistors should have their widths large enough so that the ON resistances are small. It has been established by simulation that W_H/W_L of 5 and C_V/C_0 of 5 leads to the decrease in V_{eff} within 10 % of V_{dd} and the degradation in gate delay time within 15 % compared to a pure low-V_{th} CMOS. So far as C_V is concerned, the condition is satisfied by the intrinsic capacitances present and no external capacitance needs to be added.

Leakage Power minimization:

9.6 Power Gating

The basic approach of the MTCMOS implementation has been generalized and extended in the name of power management or power gating. Here, the basic strat-egy is to provide two power modes: an *active mode*, which is the normal operat-ing mode of the circuit, and a low-power mode when the circuit is not in use. The low-power mode is commonly termed as *sleep mode*. At an appropriate time, the circuit switches from one mode to the other in an appropriate manner such that the energy drawn from the power source is maximized with minimum or no impact on the performance.

9.6.1 Clock Gating Versus Power Gating

We have discussed the use of the clock-gating approach for the reduction of switching power. The typical activity profile for a subsystem using a clock-gating strategy is shown in Fig. 9.12a. As shown in the figure, no dynamic power dissipation takes place when the circuit is clock-gated. However, leakage power dissipation takes place even when the circuit is clock-gated. In the early generation of CMOS circuits (above 250 nm), the leakage power was an insignificant portion of the total power. So, the power dissipation of a clock-gated subsystem was negligible. However, the leakage power has grown with every generation of CMOS process technology and it is essential to use power gating to reduce leakage power when the circuit is not in use. The activity profile for the same subsystem using a power-gating strategy is shown in Fig. 9.12b. Power gating can be implemented by inserting power-gating transistors in the stack between the logic transistors and either power or ground, thus creating a virtual supply rail or a virtual ground rail, respectively. The logic block contains all low-V_{th} transistors for fastest switching speeds while the switch transistors, header or footer, are built using high-V_{th} transistors to minimize the leak-age power. Power gating can be implemented without using multiple thresholds, but it will not reduce leakage as much as if implemented with multiple thresholds. MTCMOS refers to the use of transistors with multipl

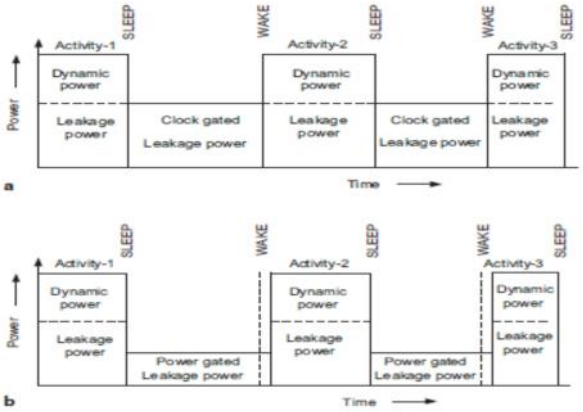


Fig. 9.12 **a** Activity profile for a subsystem with clock gating, **b** activity profile of the same subsystem with power gating

threshold voltages in power-gating circuits. The most common implementations of power gating use a footer switch alone to limit the switch area overhead. High-V_{th} NMOS footer switches are about half the size of equivalent-resistance high-V_{th} PMOS header switches due to differences in majority carrier mobilities. Power gating reduces leakage by reducing the gate-to-source voltage, which in turn, drives the logic transistors deeper into the cutoff region. This occurs because of the stack effect. The source terminal of the bottom-most transistor in the logic stack is no longer at ground, but rather at a voltage somewhat above ground due to the presence of the power-gating transistor. Leakage is reduced due to the reduction of the V_{gs}.

9.6.2 Power-Gating Issues

The clock-gating approach discussed earlier does not affect the functionality of the circuit and does not require changes in the resistor–transistor logic (RTL) representation. But, power gating is much more invasive because, as we shall see later, it affects inter-block interfaces and introduces significant time delays in order to safely enter and exit power-gated modes. The most basic form of power-gating control, and the one with the lowest long-term leakage power, is an externally switched power supply.

Various issues involved in the design of power-gated circuits are listed below:

- Power-gating granularity
- Power-gating topologies
- Switching fabric design
- Isolation strategy
- Retention strategy
- Power-gating controller design

9.6.2.1 Power-Gating Granularity

Two levels of granularity are commonly used in power gating. One is referred to as *fine-grained* power gating and the other one is referred to as *coarse-grained* power gating. In the case of fine-grained power gating, the power-gating switch is placed locally as part of the standard cell. The switch must be designed to supply the worst-case current requirement of the cell so that there is no impact on the performance. As a consequence, the size of the switch is usually large (2 × to 4 × the size of the original cell) and there is significant area overhead.

In the case of coarse-grained power gating, a relatively larger block, say a pro-cessor, or a block of gates is power switched by a block of switch cells. Consider two different implementations of a processor chip. In the first case, a single sleep control signal is used to power down the entire chip. In the second case, separate sleep control signals are used to control different building blocks such as instruction decoder, execution unit, and memory controller. The former design is considered as coarse-grained power gating, whereas the latter design may be categorized as fine-grained power gating.

The choice of granularity has both logical and physical implications. A *power domain* refers to a group of logic with a logically unique sleep signal.

Another advantage of fine-grained power gating is that the timing impact of the current I passing through a through a switch with equivalent resistance R resulting in IR drop across the switch can be easily characterized and it may be possible to use the traditional design flow to deploy fine-grained power gating. On the other hand, the sizing of the coarse-grained switched network is more difficult because the exact switching activity of the logic block may not be known at design time. In

spite of the advantages of fine-grained power gating, the coarse-grained power gating is preferred because of its lesser area overhead.

9.6.2.2 Power-Gating Topologies

Another issue closely related to granularity of power gating is the power-gating topologies. Power-gating topologies can be categorized into three types:

•Global power gating

•Local power gating

•Switch in cell gating

9.6.2.3 Global Power Gating

Global power gating refers to a logical topology in which multiple switches are connected to one or more blocks of logic, and a single virtual ground is shared in common among all the power-gated logic blocks as shown in Fig. 9.14. This topology is effective for large blocks (coarse-grained) in which all the logic is power gated, but is less effective, for physical design reasons when the logic blocks are small. It does not apply when there are many different power-gated blocks, each controlled by a different sleep enable signal.

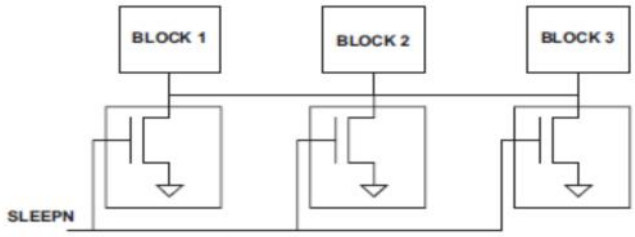


Fig. 9.14 Example of global power gating

9.6.2.4 Local Power Gating

Local power gating refers to a logical topology in which each switch singularly gates its own virtual ground connected to its own group of logic. This arrangement results in multiple segmented virtual grounds for a single sleep domain as shown in Fig. 9.15.

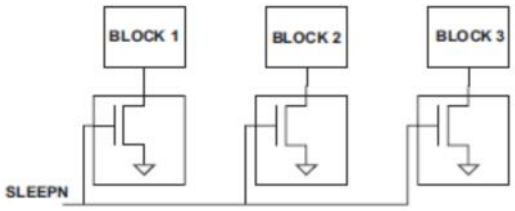
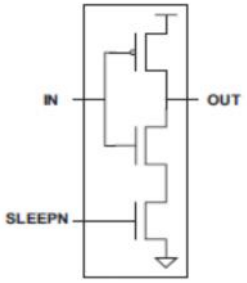


Fig. 9.15 Example of local power gating

9.6.2.5 Switch in Cell

Switch in cell may be thought of as an extreme form of local power-gating im-plementation. In this topology, each logic cell contains its own switch transistor as shown in Fig. 9.16. Its primary advantages are that delay calculation is very straightforward. The area overhead is substantial in this approach.

Fig. 9.16 Example of switch in cell power gating



9.6.2.6 Switching Fabric Design

Although the basic concept of using sleep transistors for power gating is simple, the actual implementation of the switching fabric involves many highly technology-specific issues. First and foremost among the issues is the architectural issue to decide whether to use only header switch using pMOS transistors or use only footer switch using nMOS transistors or use both. Some researchers have advocated the use of both types of switches. However, for designs at 90 nm or smaller than 90 nm, either the header or footer switch is recommended due to the tight voltage margin, significant IR drop and large area, and delay penalties when both types of transistors are used. Various issues to be addressed for switching fabric design are:

•Header-versus-footer switch

•Power-gating implementation style

Figure 9.17a shows a header switch used for power gating. High-V_t PMOS transistors are used to realize the header switch. Similarly, a footer switch used for power gating is shown in Fig. 9.17b, where high-V_t NMOS transistors are used to realize the switch.

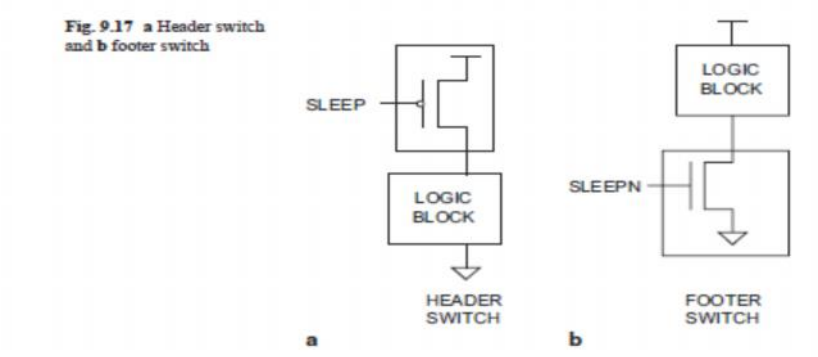


Fig. 9.17 **a** Header switch and **b** footer switch

9.6.2.8 Implementation Styles

Implementation of power-gating switches can be broadly categorized into two types: *ring* and *grid* styles. In ring-style implementation, the switches are placed external to the power-gated block by encapsulating it by a ring of switches as shown in Fig. 9.18. The switches connect VDD to the virtual VVDD of the power-gated block. This is the only style that can be used to supply power to an existing hard block by placing the switches outside it. On the other hand, the switches are distrib-uted throughout the power-gated region as shown in Fig. 9.19. Any one of the styles can be used for the implementation of coarse-grain power gating.

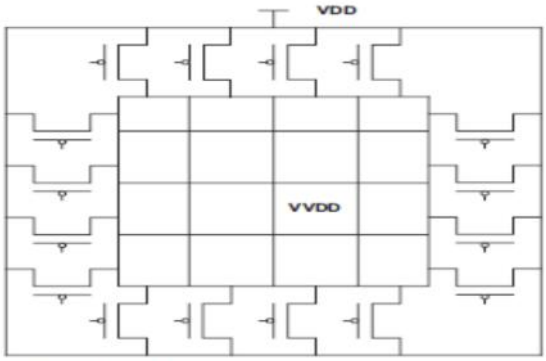
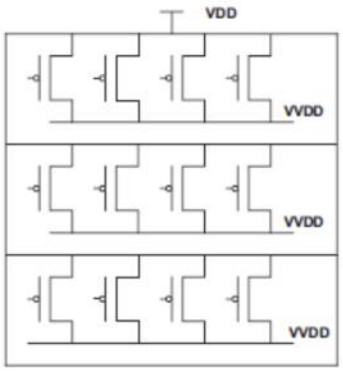


Fig. 9.18 Ring-style switching fabric

Fig. 9.19 Grid-style switching fabric



9.7 Isolation Strategy

Ideally, the outputs and internal nodes of a header-style power-gated block should collapse down towards ground level. Similarly, the outputs and internal nodes of a footer-style power-gated block should collapse down towards supply rail. How-ever, in practice, the outputs and internal nodes may neither discharge to ground level nor fully charge to supply voltage level, because of finite leakage currents passing through the off switches. So, if the output of power-down block drives a power-up block, there is possibility of short-circuit power (also known as crowbar power) in the power-up block. When the power-down block is driving a combinational logic circuit, the output can be clamped to a particular value that reduces the leakage current using stack effect (refer to Sect. 9.4).

Similarly, isolation cell to clamp the output to '1' logic level can be accomplished using an OR gate If we want to clamp the output to the last value, it is necessary to use a latch to hold the last value.

9.8 State Retention Strategy

Given a power switching fabric and an isolation strategy, it is possible to power gate a block of logic. But unless a retention strategy is employed, all state information is lost when the block is powered down. To resume its operation on power up, the block must either have its state restored from an external source or build up its state from the reset condition. In either case, the time and power required can be significant. One of the following three approaches may be used:

•A software approach based on reading and writing registers

• A scan-based approach based on the reuse of scan chains to store state off chip

• A register-based approach that uses retention registers

Software-Based Approach In the software approach, the always-ON CPU reads the registers of the power-gated blocks and stores in the processor's memory. During power-up sequence, the CPU writes back the registers from the memory. Bus traffic slows down the power-down and power-up sequence and bus conflicts may make powering down unviable. Software must be written and integrated into the system's software for handling power down and power up.

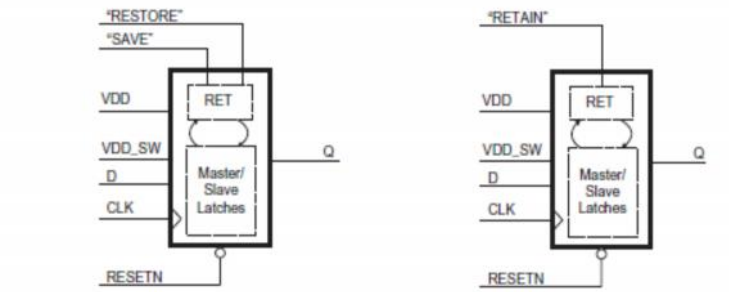


Fig. 9.24 Retention registers used for state retention

Scan-Based Approach Scan chains used for built-in self-test (BIST) can be reused. During power-down sequence, the scan register outputs are routed to an on-chip or off-chip memory. In this approach, there can be significant saving of chip area.

Retention Registers In this approach, standard registers are replaced by retention registers. A retention register contains a shadow register that can preserve the regis-ters state during power down and restore it at power up. High-V_t transistors are used in the slave latch, the clock buffers, and the inverter that connects the master latch to the slave latch as shown in Fig. 9.24. In addition to area penalty, this approach requires more complex power controller.

9.9 Power-Gating Controller

A key concern in controlling the switching fabric is to limit the in-rush of current when power to the block is switched on. An excessive in-rush current can cause voltage spikes on the supply, possibly corrupting registers in the always on blocks, as well as retention registers in the power-gated block. One representative approach is to daisy-chain the control signal to the switches. The result of this daisy chaining is that it takes some time from the assertion of a 'power-up' signal until the block is powered up. A

more aggressive approach to turning on the switching fabric is to use several power-up control signals in sequence. Regardless of the specific control method, during the power-up sequence, it is important to wait until the switching fabric is completely powered up before enabling the power-gated block to resume normal operation.

9.10 Power Management

The basic idea of power management stems from the fact that all parts of a circuit are not needed to function all the time. The power management scheme can identify conditions under which either certain parts of the circuit or the entire circuit can re-main idle and shut them down to reduce power consumption. For example, the most conventional approach used in the X86-compatible processors is to regulate the power consumption by rapidly altering between running the processor at full speed and turning the processor off. A different performance level is achieved by varying the on/off time (duty cycle) of the processor.

10. Adiabatic Logic Circuits

10.1 Introduction

Static complementary metal–oxide–semiconductor (CMOS) circuits are extremely successful in terms of market share because of many advantages such as lower pow-er dissipation, reliable operation and availability of computer-aided design (CAD) synthesis tools. We have seen that all the circuit nodes make a rail-to-rail (0 and V_{dd}) transition for each switching event and the supply voltage V_{dd} remains constant. As a consequence, the output node makes a transition from 0 to V_{dd} with a load capacitance C_L, an energy of $C_L V^2_{dd}$ is drawn from the power supply. Out of this, $1/2C_L V^2_{dd}$ is stored in the capacitor and the remaining half is dissipated in the p-type metal–oxide–semiconductor (pMOS) network. Subsequently, when the output node switches from V_{dd} to 0, the energy that was stored in the capacitor is dissipated in the n-type metal–oxide–semiconductor (nMOS) network. The power dissipation that takes place because of these switching events is converted to heat, which is ultimately released to the environment. This has far-reaching consequences like global warming. To reduce power dissipation, the circuit designers can reduce the supply voltage, decrease the node capacitance or minimize the number of switching events. In the preceding chapters, we have discussed these approaches.

10.2 Adiabatic Charging

To get introduced to the basic concept of adiabatic circuits, first we consider the conventional charging of a capacitor *C* through a resistor *R,* followed by adiabatic charging [2]. Figure 10.1a consists of a resistor *R* and capacitor *C* in series and a

supply voltage V_{dd}. As the switch is closed at time t = 0, current starts flowing. Initially, at time t = 0, the capacitor does not have any charge and therefore the voltage across the capacitor is 0 V and the voltage across the resistor is V_{dd}. So, a cur-rent of V_{dd}/R flows through the circuit. As current flows through the circuit, charge

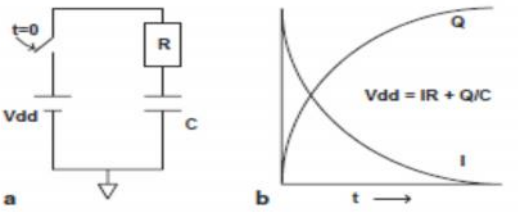


Fig. 10.1 **a** Charging of a capacitor *C* through a resis-tor *R* using a power supply. **b** As charging progresses, current decreases and charge increases

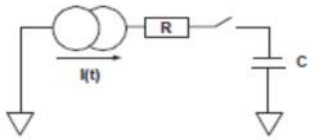


Fig. 10.2 Adiabatic charging of a capacitor

accumulates in the capacitor and voltage builds up. As the time progresses, the voltage across the resistor decreases with a consequent reduction in current through the circuit. At any instant of time, $V_{dd} = IR + Q/C$, where Q is the charge stored in the capacitor. Figure 10.1b shows how conventional charging of a capacitor leads to the dissipation of an energy of $1/2C_L V^2_{dd}$.

Now let us consider the adiabatic charging of a capacitor as shown in Fig. 10.2. Here, a capacitor C is charged through a resistor R using a constant current $I(t)$ in-stead of a fixed voltage V_{dd}. Here also it is assumed that initially at time t = 0, there is no charge in the capacitor. The voltage across the capacitor $V_c(t)$ is a function of time and it can be represented by the following expression:

$$V_c(t) = 1/C \cdot I(t) \cdot t$$
$$or\ I(t) = C \cdot V_c(t)/t$$

Assuming the current is constant, the energy dissipated by the resistor in the time interval 0 to T is given by

$$E_{diss} = R \cdot I^2(t) \cdot dt$$
$$= R \cdot I(T)^2 T$$
$$= (RC/T) \cdot C \cdot V_c(T)^2,$$

where $V_c(T)$ is the voltage across the capacitor at time T.

We may make the following conclusions from Eq. (10.1):

- For $T > 2RC$, the dissipated energy is smaller than $C_L V^2_{dd,}$ which is dissipated in the case of conventional charging using a supply voltage of V_{dd}.

• The energy dissipated can be made arbitrarily small by increasing the time *T*. Longer the value of *T*, smaller is the dissipated energy.

- Contrary to the conventional charging where the energy dissipation is indepen-dent of the value of the resistor *R*, here the energy dissipated is proportional to the value of *R*. This has opened up the possibility of reducing the energy dissipa-tion by decreasing the value of *R*.

- In the case of adiabatic charging, the charge moves from the power supply slow-ly but efficiently (dissipating lesser energy).

Another requirement is that the power supply must generate a non-standard time-varying output contrary to the fixed volt-age output generated by standard power supplies. These power supplies are known as 'pulsed power supplies' having the output characteristic as shown in Fig. 10.3. It may be noted that it has four phases: *precharge hold*, *recover* and *wait*. In the pre-charge phase, the load capacitor is adiabatically charged, in the hold phase necessary computation is performed, in the recover phase the charge is transferred back to the power supply and finally the wait phase before a new pre-charge phase starts.

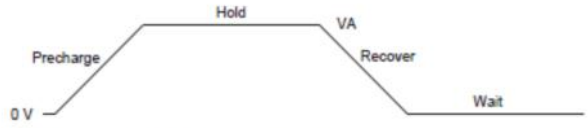
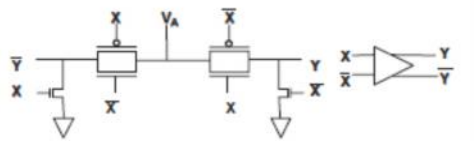


Fig. 10.3 Output waveform of a pulsed power supply

10.3 Adiabatic Amplification

Amplification is a fundamental operation performed by electronic circuits to increase the current or voltage drive. In this section, we shall discuss how it can be done adia-batically to drive capacitive loads. Here, the adiabatic amplification is implement-ed using two transmission gates and the output is dual-rail encoded, which means amplified output along with its complemented output is available. To drive the trans-mission gates, input is also dual-rail encoded as shown in Fig. 10.4. Apart from the transmission gates, two clamping circuits are also used. The steps of operation of the circuit are as follows:

Fig. 10.4 Adiabatic amplification



Step 1: Input X and its complement are applied to the circuit, which remain stable in the following steps.

Step 2: The amplifier is activated by applying V_A, which is a slow ramp voltage from 0 V to V_{dd}.

Step 3: One of the two capacitors which is connected through the transmission gate is adiabatically charged to V_A and the other one is clamped to 0 V in transition time T.

Step 4: After the charging is complete, the output signal pair remains stable and can be used as inputs to the next stage of the circuit.

Step 5: The amplifier is de-energized by ramping the voltage from V_A to 0 V. In this step, the energy that was stored in C is transferred back to the power supply.

Let us consider the energy dissipation that takes place in the above operation. Energy dissipation takes place in steps 3 and 5. As V_A ramps up and down between 0 V and $V_{dd,}$ the states of the two transistors of the transmission gate change. Both the transistors operate in the non-saturated region in the middle part of ramping up and ramping down (between V_{tp} and $V_{dd}-V_{tn}$). Initially, the nMOS transistor is ON and it remains ON till the output reaches the voltage ($V_{dd}-V_{tn}$). On the other hand, the pMOS transistor turns ON when the ramp voltage attains the voltage $|V_{tp}|$ and remains ON till the maximum value.

10.4 Adiabatic Logic Gates

Starting with a static CMOS gate, the adiabatic logic gate for the same Boolean function can be realized using the following steps:

Step 1: Replace each pull-up nMOS network and the pull-down pMOS network of the static CMOS circuit with transmission gates.

Step 2: Use the expanded pull-up network to drive the true output load capacitance.

Step 3: Use the expanded pull-down network to drive the complementary output load capacitance.

Step 4: Replace V_{dd} by a pulsed power supply V_A.

Figure 10.5a shows the schematic diagram of a static CMOS circuit with its pull-up and pull-down blocks. Figure 10.5b shows the transformed adiabatic circuit where both the networks are used to charge and discharge the load capacitances adiabati-cally. Figure 10.6 shows the realization of the adiabatic AND/NAND gate based on

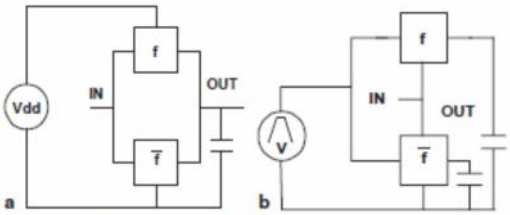
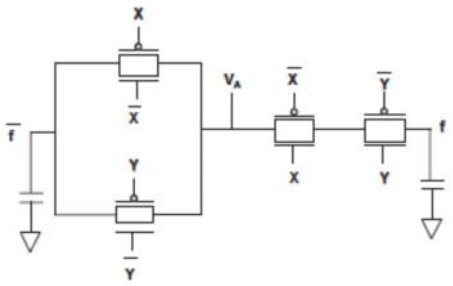


Fig. 10.5 a Static CMOS schematic diagram, b adiabatic circuit schematic diagram

Fig. 10.6 Adiabatic realization of the AND/NAND gate



the above procedure. In this way, the adiabatic realization of any function can be performed. It may be noted that the number of transistors required for the realization of the adiabatic circuit is larger than that of the static CMOS realization of the same function.

10.5 Pulsed Power Supply

Pulse power supply plays an important role in the realization of adiabatic circuits. As we know, adiabatic circuits allow less energy dissipation during charging/discharging of the load capacitance compared to static CMOS circuits. It also allows energy recovery during discharge of the load capacitance and above all it serves as the timing clock for the adiabatic circuits. The recovered node energies are also stored in pulsed power supply. Total energy consumed in an adiabatic switching operation is the sum of the energy consumed by the adiabatic circuit and the pulsed power supply. Therefore, the pulsed power supply should dissipate much less energy to achieve maximum possible energy efficiency from an adiabatic circuit.

The power clock generators can be grouped into two main types: asynchronous and synchronous. Asynchronous power clock generators are free running circuits that use feedback loops to self-oscillate without any external timing signals. Figure 10.7 illustrates two commonly used asynchronous power clock generators: 2N and 2N2P power clock generators. These are simple, dual-rail LC oscillators where the active elements are cross-coupled pairs of NMOS and PMOS transistors.

Asynchronous structures are associated with several problems. Using phase-locked loops or synchronizers such as self-timed first-in-first-out (FIFO) memory devices would not be energy- and area-efficient solutions to this problem. In such cases, the synchronous power clock generators provide a better alternative in terms of efficiency.

Synchronous power clock generators are synchronized to external timing signals usually available in large systems. Figure 10.8 illustrates two synchronous power clock generators similar to the asynchronous counterparts except that the gate control signals are generated externally. The capacitors CE1 and CE2 are external balancing capacitors to achieve better conversion efficiency. The adiabatic module can be easily synchronized to a larger conventional non-adiabatic system by using synchronous power clock generators.

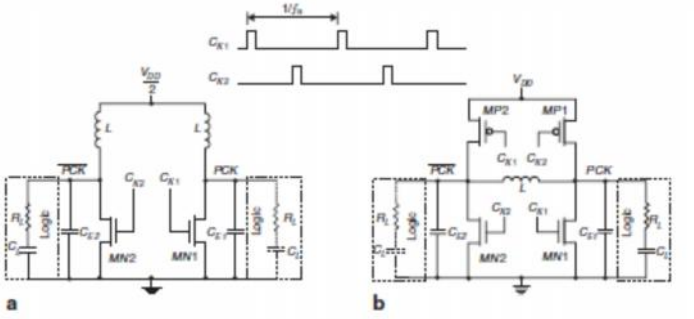


Fig. 10.8 Synchronous two-phase clock generator **a** 2N, **b** 2N2P

10.6 Stepwise Charging Circuits

The general power supply has several difficulties at the circuit level, especially in terms of chip-level integration and overall efficiency. An alternative to using pure voltage ramps is to use stepwise supply voltage waveforms where the output voltage of the power supply is increased and decreased in small increments during charging and discharging.

Figure 10.9 shows a CMOS inverter driven by a stepwise supply voltage wave-form. Assuming that the output voltage is equal to zero initially, the input voltage set to logic low level, the power supply voltage V_A is increased from 0 to V_{dd}, in n equal voltage steps as shown in Fig. 10.10. Since the pMOS transistor is conducting during this transition, the output load capacitance will be charged up in a stepwise manner. The on-resistance of the pMOS transistor can be represented by the linear resistor R. Thus, the output load capacitance is being charged up through a resistor, in small voltage increments.

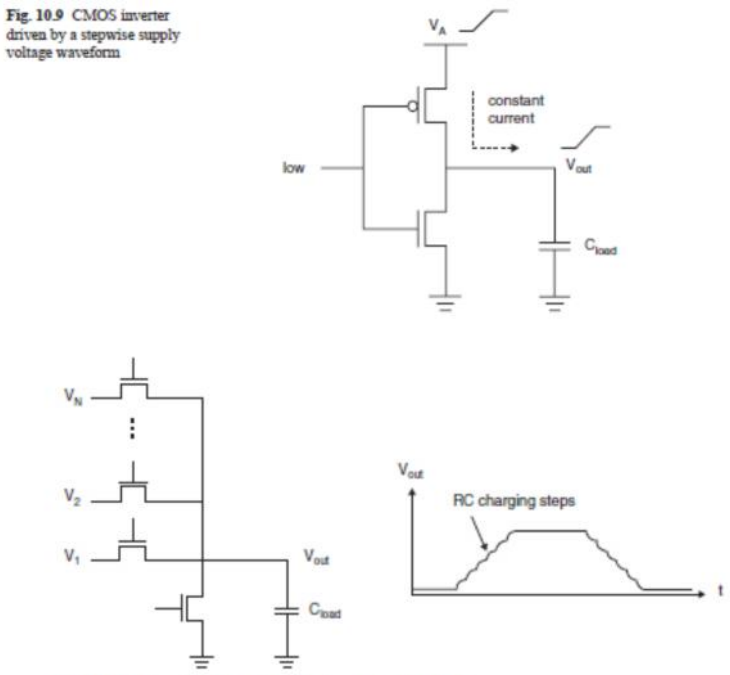


Fig. 10.9 CMOS inverter driven by a stepwise supply voltage waveform

Fig. 10.11 Stepwise driver circuit to charge capacitive loads

The switch devices are shown as nMOS transistors in Fig. 10.11, yet some of them may be successively connected to constant voltage sources V_i through an array of switches replaced by pMOS transistors to prevent

the undesirable threshold volt-age drop problem and the substrate-bias effects at higher voltage levels.

One of the most significant drawbacks of this circuit configuration is the need for multiple supply voltages. A power supply system capable of efficiently generating *n* different voltage levels, would be complex and expensive.

10.7 Partially Adiabatic Circuits

Implementation of fully reversible adiabatic logic circuits has a very large over-head. A fully reversible, bit-level pipelined three-bit adder requires several times as many devices as a conventional one and many times the silicon area. This has motivated researchers to apply the adiabatic technique to realize a partially adiabatic logic circuit. Most of the circuits use crossed coupled devices connecting two nodes that form the true and complementary outputs, i.e. dual-rail encoded. When a volt-age ramp is applied, the outputs settle to one of the states based on the inputs. V_Φ is connected to the pulsed-power supply.

There is non-adiabatic dissipation of approximately $(1/2)C_L V_{th}^2$ for transition from one state to another.

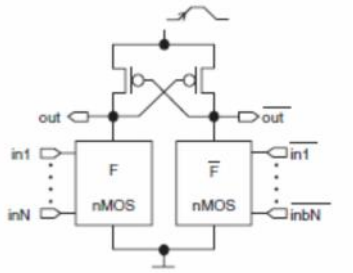
10.7.1 Efficient Charge Recovery Logic

Figure 10.13 shows the generalized schematic for efficient charge recovery logic (ECRL) . It consists of two pMOS transistors connected in a cross-coupled man-ner and two networks of NMOS transistors acting as evaluation networks. The waveforms of the supply clock as well as I/O signals for a NOT gate. In order to recover and to reuse the supplied energy, an ac power supply is also used for ECRL gates. As usual, in adiabatic circuits, the supply voltage also acts as a clock. Both the signals, out and its complement are generated so that the power clock generator can always drive a constant load capacitance, independent of the input signal. If the circuit operates correctly, energy has an oscillatory behaviour, be-cause a large part of the energy supplied to the circuit is given back to the power supply. As usual, for adiabatic logic, the energy behaviour follows the supply voltage. It is also observed that, due to a coupling effect, the low-level output goes to a negative voltage value during the recovery phase (that is, when the supply voltage ramps down).

The dissipated energy can be defined as the difference between the energy that the circuit needs to load the output capacitance, and the energy

that the circuit re-turns back to the power supply during the recovery phase. The dissipated energy value depends on the input sequence and on the switching activity factor. Therefore, the dissipated energy per cycle can be obtained from the mean value of the whole sequence. It can also be seen that a larger energy is dissipated if the input state changes and therefore the output capacitances have to switch from one voltage level to the other.

Fig. 10.13 ECRL generalized schematic diagram



An ECRL realization of an inverter is shown in Fig. 10.14. How the data are transferred from one stage to its succeeding stage is shown in Fig. 10.15. The arrows show when data move from one gate to the consecutive gate in four phases: precharge, hold, recover and wait phases.

Fig. 10.14 ECRL inverter

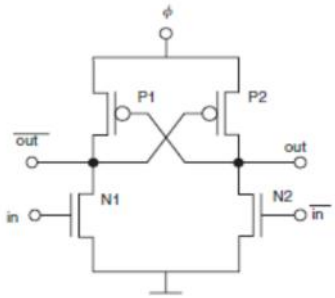
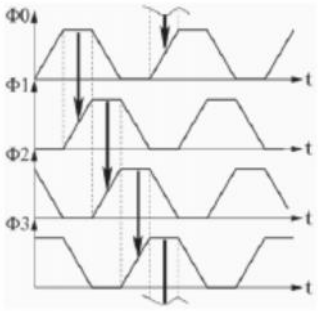


Fig. 10.15 Data transfer in ECRL gates



10.7.2 Positive Feedback Adiabatic Logic Circuits

The structure of a positive feedback adiabatic logic (PFAL) gate is shown in Fig. 10.16. Two nMOS networks are used to realize the logic functions. This logic family also generates both positive and negative outputs. The two major differences with respect to ECRL are that the latch is made by two pMOS and two nMOS FETs, rather than by only two pMOS FETs as in ECRL, and that the functional blocks are in parallel with the transmission pMOS FETs. Thus, the equivalent resistance is smaller when the capacitance needs to be charged. During the recovery phase, the loaded capacitance gives back energy to the power supply and the supplied energy decreases. The input NMOS network is connected in parallel to the PMOS transistors. In a PFAL gate, no output is floating and all outputs have full logic swing; PFAL shows the best performance in terms of energy consumption, useful frequency range and robustness against technology variations.

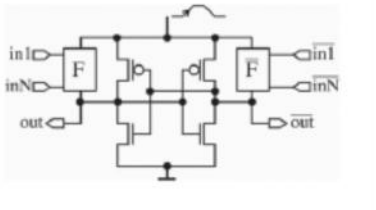


Fig. 10.16 Schematic diagram of a PFAL logic gate

10.7.3 2N−2N2P Inverter/Buffer

This adiabatic logic family was derived from ECRL in order to reduce the coupling effect. Figure 10.18 shows the general schematic diagram. The primary advantage of 2N−2N2P over ECRL is that the cross-coupled nMOSFETs switches result in non-floating outputs for large part of the recovery phase.

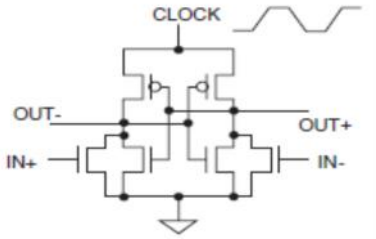


Fig. 10.18 Schematic diagram of a 2N-2N2P logic gate

11. Battery-Aware Systems

11.1 Introduction

Over the years, with increasing usage of mobile devices in everyday life, there is proliferation of portable computing and communication equipments, such as laptops, palmtops, cell phones, etc. The growth rate of the number of these por-table devices is very high compared to the rate of growth of desktop and server systems. It has been observed that the processing capability of the contemporary portable computing devices is becoming comparable to that of desktop computers. Moreover, complexity of these portable computing devices is increasing due to the gradual addition of more and more functionality. However, power dissipation keeps on increasing with the increase in computing complexity. Fortunately, with the advancement of very-large- scale integration (VLSI) technology and power-efficient approaches used in the design, these portable devices do not consume as much power as that of desktop computers. As these devices are battery operated, battery life is of primary concern, and it has put additional constraints. Commercial success of these products depends on weight, cost, and battery run time after each recharge. Unfortunately, the battery technology has not kept up with the energy requirement of the portable equipment. To satisfy larger energy requirement, use of a battery of higher capacity is not a solution because, for portable devices, the size and weight of battery, which are proportional to the battery capacity, have stringent design constraints. This has motivated the designers to consider alternative approaches such as battery-aware synthesis, to satisfy the energy requirement of the portable devices.

11.7 Battery-Driven System Design

Battery-driven system design involves the use of one or more of the following techniques:

Voltage and Frequency Scaling As we mentioned earlier, the power dissipation has square law dependence on the supply voltage and linear dependence on the frequency. Depending upon the performance requirement, the supply voltage V_{dd} and frequency of operation of the circuit driven by the battery can be dynamically adjusted to optimize energy consumption from the battery. Information from a battery model is used to vary the clock frequency dynamically at run time based on the workload characteristics. If the workload is higher, higher voltage and clock frequency are used, and, for lower workload, the voltage and clock frequency can be lowered such that the battery is discharged at a lower rate. This, in turn, improves the actual capacity of the battery.

Dynamic Power Management The state of charge of the battery can be used to frame a policy that controls the operation state of the system.

Battery-Aware Task Scheduling The current discharge profile is tailored to meet battery characteristics to maximize the actual battery capacity.

Battery Scheduling and Management Efficient management of multi-battery systems by using appropriate scheduling of the batteries.

Static Battery Scheduling These are essentially open-loop approaches, such as serial scheduling, random scheduling, round-robin scheduling, where scheduling is done without checking the condition of a battery.

Terminal Voltage-Based Battery Scheduling The scheduling algorithm makes use of the state of charge of the battery.

Discharge Current-Based Battery Scheduling This approach is used when heterogeneous batteries with different rate capacities are used.

Battery-Efficient Traffic Shaping and Routing Network protocols and communication traffic patterns play important roles in determining battery efficiency and lifetime.

11.7.1 Multi-battery System

Instead of having a single battery, it is possible to use multiple batteries in a single system. However, the load can be serviced entirely by a single battery at a time. Here, the goal is to switch the load between cells in such

a way that their lifetime is maximized, and this can result in very diverse load distributions. The control problem that is being used is much more complex because there is no single set point that can be used to improve the behavior of the system. More efficient use of multiple batteries can be achieved by exploiting the phenomenon of recovery effect, which is a consequence of the chemical properties of a battery: as the charge is drawn from a battery, the stored charge is released by a chemical reaction, which takes time to replenish the charge. In general, the charge is drawn from a battery at a faster rate than the reaction can replenish it, and this leads to a battery appearing to become devoid of charge when, in fact, it still contains stored charge. By allow-ing the battery to remain idle, the reaction can replenish the charge, and the battery becomes operational once again as we have already mentioned. Thus, efficient use of multiple batteries involves carefully timing the use and idle periods for a set of batteries. This problem can be considered as a planning problem.

11.7.3 Task Scheduling with Voltage Scaling

Task scheduling can be combined with voltage scaling to maximize the amount of charge that a battery can supply subject to the following constraints:

a. Dependency constraint: task dependencies are preserved;

b. Delay constraint: the profile length is within the delay budget; and

c. Endurance constraint: the battery survives all the tasks.