5.5 Power, Energy, and Energy-Delay

So far, we have seen that the static CMOS inverter with its almost ideal VTC—symmetrical shape, full logic swing, and high noise margins—offers a superior robustness, which simplifies the design process considerably and opens the door for design automation. Another major attractor for static CMOS is the almost complete absence of power consumption in steady-state operation mode. It is this combination of robustness and low static power that has made static CMOS the technology of choice of most contemporary digital designs. The power dissipation of a CMOS circuit is instead dominated by the dynamic dissipation resulting from charging and discharging capacitances.

5.5.1 Dynamic Power Consumption

Dynamic Dissipation due to Charging and Discharging Capacitances

Each time the capacitor C_L gets charged through the PMOS transistor, its voltage rises from 0 to V_{DD}, and a certain amount of energy is drawn from the power supply. Part of this energy is dissipated in the PMOS device, while the remainder is stored on the load capacitor. During the high-to-low transition, this capacitor is discharged, and the stored energy is dissipated in the NMOS transistor.[3]

A precise measure for this energy consumption can be derived. Let us first consider the low-to-high transition. We assume, initially, that the input waveform has zero rise and fall times, or, in other words, that the NMOS and PMOS devices are never on simultaneously. Therefore, the equivalent circuit of Figure 5.25 is valid. The values of the energy E_{VDD}, taken from the supply during the transition, as well as the energy E_C, stored on the capacitor at the end of the transition, can be derived by integrating the instantaneous power over the period of interest. The corresponding waveforms of $v_{out}(t)$ and $i_{VDD}(t)$ are pictured in Figure 5.26.

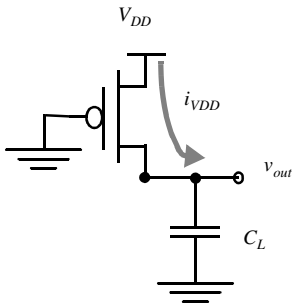


Figure 5.25 Equivalent circuit during the low-to-high transition.

$$E_{VDD} = \int_0^\infty i_{VDD}(t) V_{DD} dt = V_{DD} \int_0^\infty C_L \frac{dv_{out}}{dt} dt = C_L V_{DD} \int_0^{V_{DD}} dv_{out} = C_L V_{DD}^2 \qquad (5.39)$$

[3] Observe that this model is a simplification of the actual circuit. In reality, the load capacitance consists of multiple components some of which are located between the output node and GND, others between output node and V_{DD}. The latter experience a charge-discharge cycle that is out of phase with the capacitances to GND, i.e. they get charged when V_{out} goes low and discharged when V_{out} rises. While this distributes the energy delivery by the supply over the two phases, it does not impact the overall dissipation, and the results presented in this section are still valid.

$$E_C = \int_0^\infty i_{VDD}(t)v_{out}dt = \int_0^\infty C_L\frac{dv_{out}}{dt}v_{out}dt = C_L\int_0^{V_{DD}} v_{out}dv_{out} = \frac{C_L V_{DD}^2}{2} \tag{5.40}$$

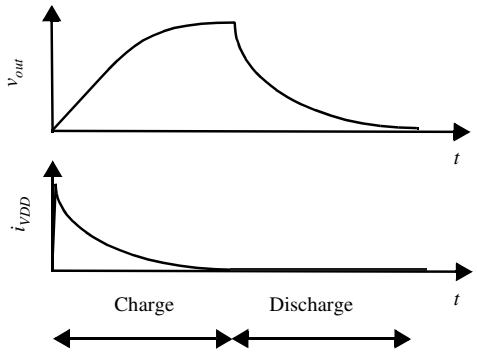


Figure 5.26 Output voltages and supply current during (dis)charge of C_L.

These results can also be derived by observing that during the low-to-high transition, C_L is loaded with a charge $C_L V_{DD}$. Providing this charge requires an energy from the supply equal to $C_L V_{DD}^2$ $(= Q \times V_{DD})$. The energy stored on the capacitor equals $C_L V_{DD}^2/2$. This means that only half of the energy supplied by the power source is stored on C_L. The other half has been dissipated by the PMOS transistor. Notice that this energy dissipation is independent of the size (and hence the resistance) of the PMOS device! During the discharge phase, the charge is removed from the capacitor, and its energy is dissipated in the NMOS device. Once again, there is no dependence on the size of the device. In summary, each switching cycle (consisting of an L→H and an H→L transition) takes a fixed amount of energy, equal to $C_L V_{DD}^2$. In order to compute the power consumption, we have to take into account how often the device is switched. If the gate is switched **on and off** $f_{0\to1}$ times per second, the power consumption equals

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1} \tag{5.41}$$

$f_{0\to1}$ represents the frequency of energy-consuming transitions, this is $0 \to 1$ transitions for static CMOS.

Advances in technology result in ever-higher of values of $f_{0\to1}$ (as t_p decreases). At the same time, the total capacitance on the chip (C_L) increases as more and more gates are placed on a single die. Consider for instance a 0.25 μm CMOS chip with a clock rate of 500 Mhz and an average load capacitance of 15 fF/gate, assuming a fanout of 4. The power consumption per gate for a 2.5 V supply then equals approximately 50 μW. For a design with 1 million gates and assuming that a transition occurs at every clock edge, this would result in a power consumption of 50 W! This evaluation presents, fortunately, a pessimistic perspective. In reality, not all gates in the complete IC switch at the full rate of 500 Mhz. The actual activity in the circuit is substantially lower.

Example 5.11 Capacitive power dissipation of inverter

The capacitive dissipation of the CMOS inverter of Example 5.4 is now easily computed. In Table 5.2, the value of the load capacitance was determined to equal 6 fF. For a supply voltage of 2.5 V, the amount of energy needed to charge and discharge that capacitance equals

$$E_{dyn} = C_L V_{DD}^2 = 37.5 \text{ fJ}$$

Assume that the inverter is switched at the maximum possible rate ($T = 1/f = t_{pLH} + t_{pHL} = 2\,t_p$). For a t_p of 32.5 psec (Example 5.5), we find that the dynamic power dissipation of the circuit is

$$P_{dyn} = E_{dyn}/(2t_p) = 580\ \mu\text{W}$$

Of course, an inverter in an actual circuit is rarely switched at this maximum rate, and even if done so, the output does not swing from rail-to-rail. The power dissipation will hence be substantially lower. For a rate of 4 GHz ($T = 250$ psec), the dissipation reduces to 150 μW. This is confirmed by simulations, which yield a power consumption of 155 μW.

Computing the dissipation of a complex circuit is complicated by the $f_{0\to1}$ factor, also called the *switching activity*. While the switching activity is easily computed for an inverter, it turns out to be far more complex in the case of higher-order gates and circuits. One concern is that the switching activity of a network is a function of the nature and the statistics of the input signals: If the input signals remain unchanged, no switching happens, and the dynamic power consumption is zero! On the other hand, rapidly changing signals provoke plenty of switching and hence dissipation. Other factors influencing the activity are the overall network topology and the function to be implemented. We can accommodate this by another rewrite of the equation, or

$$P_{dyn} = C_L V_{DD}^2 f_{0\to1} = C_L V_{DD}^2 P_{0\to1} f = C_{EFF} V_{DD}^2 f \tag{5.42}$$

where f now presents the maximum possible event rate of the inputs (which is often the clock rate) and $P_{0\to1}$ the probability that a clock event results in a $0 \to 1$ (or power-consuming) event at the output of the gate. $C_{EFF} = P_{0\to1} C_L$ is called the *effective capacitance* and represents the average capacitance switched every clock cycle. For our example, an activity factor of 10% ($P_{0\to1} = 0.1$) reduces the average consumption to 5 W.

Example 5.12 Switching activity

Consider the waveforms on the right where the upper waveform represents the idealized clock signal, and the bottom one shows the signal at the output of the gate. Power consuming transitions occur 2 out of 8 times, which is equivalent to a transition probability of 0.25 (or 25%).

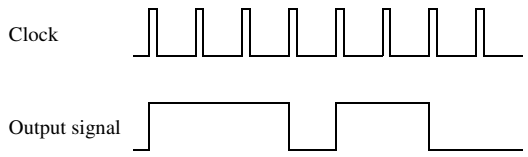


Figure 5.27 Clock and signal waveforms

Low Energy/Power Design Techniques

With the increasing complexity of the digital integrated circuits, it is anticipated that the power problem will only worsen in future technologies. This is one of the reasons that lower supply

voltages are becoming more and more attractive. **Reducing V_{DD} has a quadratic effect on P_{dyn}.** For instance, reducing V_{DD} from 2.5 V to 1.25 V for our example drops the power dissipation from 5 W to 1.25 W. This assumes that the same clock rate can be sustained. Figure 5.17 demonstrates that this assumption is not that unrealistic as long as the supply voltage is substantially higher than the threshold voltage. An important performance penalty occurs once V_{DD} approaches 2 V_T.

When a lower bound on the supply voltage is set by external constraints (as often happens in real-world designs), or when the performance degradation due to lowering the supply voltage is intolerable, the only means of reducing the dissipation is by lowering the effective capacitance. This can be achieved by addressing both of its components: the physical capacitance and the switching activity.

A *reduction in the switching activity* can only be accomplished at the logic and architectural abstraction levels, and will be discussed in more detail in later Chapters. *Lowering the physical capacitance* is an overall worthwhile goal, which also helps to improve the performance of the circuit. As most of the capacitance in a combinational logic circuit is due to transistor capacitances (gate and diffusion), it makes sense to keep those contributions to a minimum when designing for low power. This means that transistors should be kept to *minimal size* whenever possible or reasonable. This definitely affects the performance of the circuit, but the effect can be offset by using logic or architectural speed-up techniques. The only instances where transistors should be sized up is when the load capacitance is dominated by extrinsic capacitances (such as fan-out or wiring capacitance). This is contrary to common design practices used in cell libraries, where transistors are generally made large to accommodate a range of loading and performance requirements.

The above observations lead to an interesting design challenge. Assume we have to minimize the energy dissipation of a circuit with a specified lower-bound on the performance. An attractive approach is to lower the supply voltage as much as possible, and to compensate the loss in performance by increasing the transistor sizes. Yet, the latter causes the capacitance to increase. It may be foreseen that at a low enough supply voltage, the latter factor may start to dominate and cause energy to increase with a further drop in the supply voltage.

Example 5.13 Transistor Sizing for Energy Minimization

To analyze the transistor-sizing for minimum energy problem, we examine the simple case of a static CMOS inverter driving an external load capacitance C_{ext}. To take the input loading effects into account, we assume that the inverter itself is driven by a minimum-sized device (Figure 5.28). The goal is to minimize the energy dissipation of the complete circuit, while maintaining a lower-bound on performance. The degrees of freedom are the size factor f of the inverter and the supply voltage V_{dd} of the circuit. The propagation delay of the optimized circuit should not be larger than that of a reference circuit, chosen to have as parameters $f = 1$ and $V_{dd} = V_{ref}$.

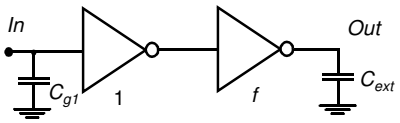


Figure 5.28 CMOS inverter driving an external load capacitance C_{ext}, while being driven by a minimum sized gate.

Using the approach introduced in Section 5.4.3 (*Sizing a Chain of Inverters*), we can derive an expression for the propagation delay of the circuit,

$$t_p = t_{p0}\left(\left(1+\frac{f}{\gamma}\right)+\left(1+\frac{F}{f\gamma}\right)\right) \tag{5.43}$$

with $F = (C_{ext}/C_{g1})$ the overall effective fanout of the circuit t_{p0} is the intrinsic delay of the inverter. Its dependence upon V_{DD} is approximated by the following expression, derived from Eq. (5.21).

$$t_{p0} \sim \frac{V_{DD}}{V_{DD}-V_{TE}} \tag{5.44}$$

The energy dissipation for a single transition at the input is easily found once the total capacitance of the circuit is known, or

$$E = V_{dd}^2 C_{g1}((1+\gamma))(1+f)+F) \tag{5.45}$$

The performance constraint now states that the propagation delay of the scaled circuit should be equal (or smaller) to the delay of the reference circuit (f=1, $V_{dd} = V_{ref}$). To simplify the subsequent analysis, we make the simplifying assumption that the intrinsic output capacitance of the gate equals its gate capacitance, or $\gamma = 1$. Hence,

$$\frac{t_p}{t_{pref}} = \frac{t_{p0}\left(2+f+\frac{F}{f}\right)}{t_{p0ref}(3+F)} = \left(\frac{V_{DD}}{V_{ref}}\right)\left(\frac{V_{ref}-V_{TE}}{V_{DD}-V_{TE}}\right)\left(\frac{2+f+\frac{F}{f}}{3+F}\right) = 1 \tag{5.46}$$

Eq. (5.46) establishes a relationship between the sizing factor f and the supply voltage, plotted in Figure 5.29a for different values of F. Those curves show a clear minimum. Increasing the size of the inverter from the minimum initially increases the performance, and hence allows for a lowering of the supply voltage. This is fruitful until the optimum sizing factor of $f = \sqrt{F}$ is reached, which should not surprise careful readers of the previous sections. Further increases in the device sizes only increase the self-loading factor, deteriorate the performance, and require an increase in supply voltage. Also observe that for the case of F=1, the reference case is the best solution; any resizing just increases the self-loading.

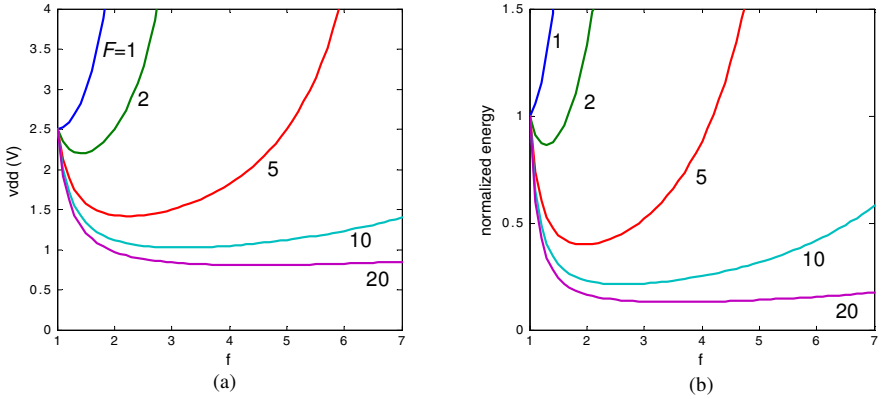


Figure 5.29 Sizing of an inverter for energy-minimization. (a) Required supply voltage as a function of the sizing factor f for different values of the overall effective fanout F; (b) Energy of scaled circuit (normalized with respect to the reference case) as a function of f. V_{ref} = 2.5V, V_{TE} = 0.5V.

With the $V_{DD}(f)$ relationship in hand, we can derive the energy of the scaled circuit (normalized with respect to the reference circuit) as a function of the sizing factor f.

$$\frac{E}{E_{ref}} = \left(\frac{V_{DD}}{V_{ref}}\right)^2\left(\frac{2+2f+F}{4+F}\right) \tag{5.47}$$

Finding an analytical expression for the optimal sizing factor is possible, but yields a complex and messy equation. A graphical approach is just as effective. The resulting charts are plotted in Figure 5.29b, from which a number of conclusions can be drawn:

- **Device sizing, combined with supply voltage reduction, is a very effective approach in reducing the energy consumption of a logic network.** This is especially true for networks with large effective fanouts, where energy reductions with almost a factor of 10 can be observed. But the gain is also sizable for smaller values of F. The only exception is the F=1 case, where the minimum size device is also the most effective one.
- Oversizing the transistors beyond the optimal value comes at a hefty price in energy. This is unfortunately a common approach in many of today's designs.
- The optimal sizing factor for energy is smaller than the one for performance, especially for large values of F. For example, for a fanout of 20, f_{opt}(energy) = 3.53, while f_{opt}(performance) = 4.47. Increasing the device sizes only leads to a minimal supply reduction once V_{DD} starts approaching V_{TE}, hence leading to very minimal energy gains.

Dissipation Due to Direct-Path Currents

In actual designs, the assumption of the zero rise and fall times of the input wave forms is not correct. The finite slope of the input signal causes a direct current path between V_{DD} and *GND* for a short period of time during switching, while the NMOS and the PMOS transistors are conducting simultaneously. This is illustrated in Figure 5.30. Under the (reasonable) assumption that the resulting current spikes can be approximated as triangles and that the inverter is symmetrical in its rising and falling responses, we can compute the energy consumed per switching period,

$$E_{dp} = V_{DD}\frac{I_{peak}t_{sc}}{2} + V_{DD}\frac{I_{peak}t_{sc}}{2} = t_{sc}V_{DD}I_{peak} \tag{5.48}$$

as well as the average power consumption

$$P_{dp} = t_{sc}V_{DD}I_{peak}f = C_{sc}V_{DD}^2 f \tag{5.49}$$

The direct-path power dissipation is proportional to the switching activity, similar to the capacitive power dissipation. t_{sc} represents the time both devices are conducting. For a linear input slope, this time is reasonably well approximated by Eq. (5.50) where t_s represents the 0-100% transition time.

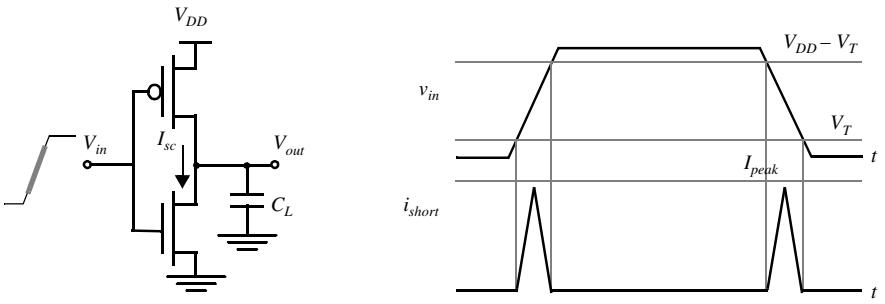
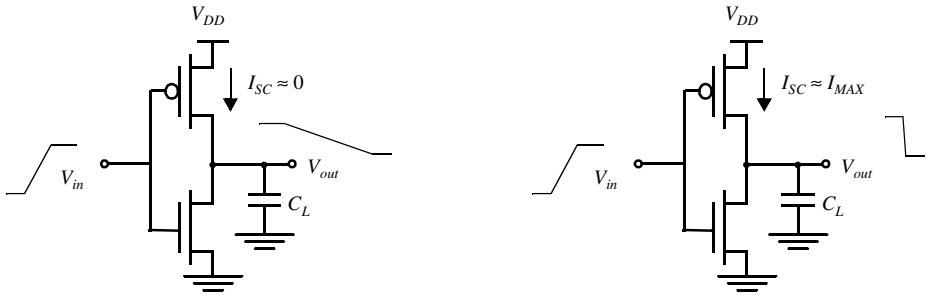


Figure 5.30 Short-circuit currents during transients.

$$t_{sc} = \frac{V_{DD} - 2V_T}{V_{DD}} t_s \approx \frac{V_{DD} - 2V_T}{V_{DD}} \times \frac{t_{r(f)}}{0.8} \tag{5.50}$$

I_{peak} is determined by the saturation current of the devices and is hence directly proportional to the sizes of the transistors. The peak current is also **a strong function of the ratio between input and output slopes**. This relationship is best illustrated by the following simple analysis: Consider a static CMOS inverter with a $0 \rightarrow 1$ transition at the input. Assume first that the load capacitance is very large, so that the output fall time is significantly larger than the input rise time (Figure 5.31a). Under those circumstances, the input



(a) Large capacitive load

(b) Small capacitive load

Figure 5.31 Impact of load capacitance on short-circuit current.

moves through the transient region before the output starts to change. As the source-drain voltage of the PMOS device is approximately 0 during that period, the device shuts off without ever delivering any current. The short-circuit current is close to zero in this case. Consider now the reverse case, where the output capacitance is very small, and the output fall time is substantially smaller than the input rise time (Figure 5.31b). The drain-source voltage of the PMOS device equals V_{DD} for most of the transition period, guaranteeing the maximal short-circuit current (equal to the saturation current of the PMOS). This clearly

represents the worst-case condition. The conclusions of the above analysis are confirmed in Figure 5.32, which plots the short-circuit current through the NMOS transistor during a low-to-high transition as a function of the load capacitance.

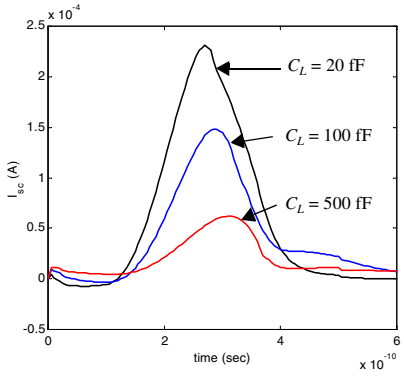


Figure 5.32 CMOS inverter short-circuit current through NMOS transistor as a function of the load capacitance (for a fixed input slope of 500 psec).

This analysis leads to the conclusion that the short-circuit dissipation is minimized by making the output rise/fall time larger than the input rise/fall time. On the other hand, making the output rise/fall time too large slows down the circuit and can cause short-circuit currents in the fan-out gates. This presents a perfect example of how local optimization and forgetting the global picture can lead to an inferior solution.

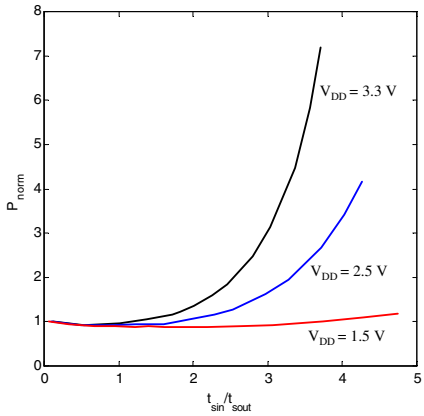
Design Techniques

A more practical rule, which optimizes the power consumption in a global way, can be formulated (Veendrick84]):

> The power dissipation due to short-circuit currents is minimized by matching the rise/fall times of the input and output signals. At the overall circuit level, this means that rise/fall times of all signals should be kept constant within a range.

Making the input and output rise times of a gate identical is not the optimum solution for that particular gate on its own, but keeps the overall short-circuit current within bounds. This is shown in Figure 5.33, which plots the short-circuit energy dissipation of an inverter (normalized with respect to the zero-input rise time dissipation) as a function of the ratio r between input and output rise/fall times. When the load capacitance is too small for a given inverter size ($r > 2…3$ for $V_{DD} = 5$ V), the power is dominated by the short-circuit current. For very large capacitance values, all power dissipation is devoted to charging and discharging the load capacitance. When the rise/fall times of inputs and outputs are equalized, most power dissipation is associated with the dynamic power and only a minor fraction (< 10%) is devoted to short-circuit currents.

Observe also that the impact of **short-circuit current is reduced when we lower the supply voltage**, as is apparent from Eq. (5.50). In the extreme case, when $V_{DD} < V_{Tn} + |V_{Tp}|$, short-circuit dissipation is completely eliminated, because both devices are never on simultaneously. With threshold voltages scaling at a slower rate than the supply voltage, short-circuit power dissipation is becoming of a lesser importance in deep-submicron technologies.



$W/L|_P$ = 1.125 µm/0.25 µm
$W/L|_N$ = 0.375 µm/0.25 µm
C_L = 30 fF

Figure 5.33 Power dissipation of a static CMOS inverter as a function of the ratio between input and output rise/fall times. The power is normalized with respect to zero input rise-time dissipation. At low values of the slope ratio, input-output coupling leads to some extra dissipation.

At a supply voltage of 2.5 V and thresholds around 0.5 V, an input/output slope ratio of 2 is needed to cause a 10% degradation in dissipation.

Finally, its is worth observing that the short-circuit power dissipation can be modeled by adding a load capacitance $C_{sc} = t_{sc}I_{peak}/V_{DD}$ in parallel with C_L, as is apparent in Eq. (5.49). The value of this short-circuit capacitance is a function of V_{DD}, the transistor sizes, and the input-output slope ratio.

5.5.2 Static Consumption

The static (or steady-state) power dissipation of a circuit is expressed by Eq. (5.51), where I_{stat} is the current that flows between the supply rails in the absence of switching activity

$$P_{stat} = I_{stat}V_{DD} \tag{5.51}$$

Ideally, the static current of the CMOS inverter is equal to zero, as the PMOS and NMOS devices are never on simultaneously in steady-state operation. There is, unfortunately, a leakage current flowing through the reverse-biased diode junctions of the transistors, located between the source or drain and the substrate as shown in Figure 5.34. This contribution is, in general, very small and can be ignored. For the device sizes under consideration, the leakage current per unit drain area typically ranges between 10-100 pA/µm^2 at room temperature. For a die with 1 million gates, each with a drain area of 0.5 µm^2 and operated at a supply voltage of 2.5 V, the worst-case power consumption due to diode leakage equals 0.125 mW, which is clearly not much of an issue.

However, be aware that the junction leakage currents are caused by thermally generated carriers. Their value increases with increasing junction temperature, and this occurs in an exponential fashion. At 85°C (a common junction temperature limit for commercial hardware), the leakage currents increase by a factor of 60 over their room-temperature val-

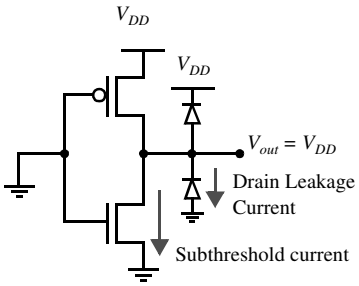


Figure 5.34 Sources of leakage currents in CMOS inverter (for V_{in} = 0 V).

ues. Keeping the overall operation temperature of a circuit low is consequently a desirable goal. As the temperature is a strong function of the dissipated heat and its removal mechanisms, this can only be accomplished by limiting the power dissipation of the circuit and/or by using chip packages that support efficient heat removal.

An emerging source of leakage current is the subthreshold current of the transistors. As discussed in Chapter 3, an MOS transistor can experience a drain-source current, even when V_{GS} is smaller than the threshold voltage (Figure 5.35). The closer the threshold voltage is to zero volts, the larger the leakage current at V_{GS} = 0 V and the larger the static power consumption. To offset this effect, the threshold voltage of the device has generally been kept high enough. Standard processes feature V_T values that are never smaller than 0.5-0.6V and that in some cases are even substantially higher (~ 0.75V).

This approach is being challenged by the reduction in supply voltages that typically goes with deep-submicron technology scaling as became apparent in Figure 3.40. We concluded earlier (Figure 5.17) that scaling the supply voltages while keeping the threshold voltage constant results in an important loss in performance, especially when V_{DD} approaches 2 V_T. One approach to address this performance issue is to scale the device thresholds down as well. This moves the curve of Figure 5.17 to the left, which means that the performance penalty for lowering the supply voltage is reduced. Unfortunately, the threshold voltages are lower-bounded by the amount of allowable subthreshold leakage current, as demonstrated in Figure 5.35. The choice of the threshold voltage hence represents a trade-off between performance and static power dissipation. The continued scaling of the supply voltage predicted for the next generations of CMOS technologies however forces the threshold voltages ever downwards, and makes subthreshold conduction a dominant source of power dissipation. Process technologies that contain devices with sharper turn-off characteristic will therefore become more attractive. An example of the latter is the SOI (Silicon-on-Insulator) technology whose MOS transistors have slope-factors that are close to the ideal 60 mV/decade.

Example 5.14 Impact of threshold reduction on performance and static power dissipation

Consider a minimum size NMOS transistor in the 0.25 μm CMOS technology. In Chapter 3, we derived that the slope factor *S* for this device equals 90 mV/decade. The off-current (at V_{GS} = 0) of the transistor for a V_T of approximately 0.5V equals 10^{-11}A (Figure 3.22). Reducing the threshold with 200 mV to 0.3 V multiplies the off-current of the transistors with a factor of 170! Assuming a million gate design with a supply voltage of 1.5 V, this translates into a static power dissipation of $10^6 \times 170 \times 10^{-11} \times 1.5$ = 2.6 mW. A further reduction of the thresh-

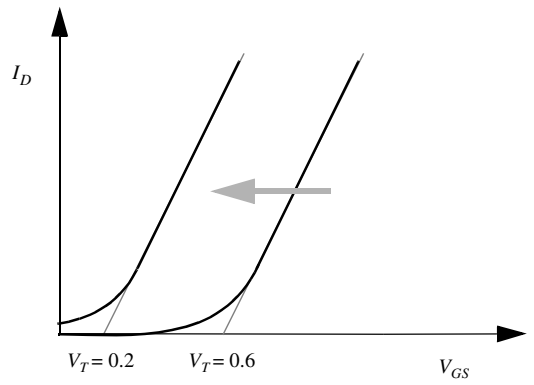


Figure 5.35 Decreasing the threshold increases the subthreshold current at $V_{GS} = 0$.

old to 100 mV results in an unacceptable dissipation of almost 0.5 W! At that supply voltage, the threshold reductions correspond to a performance improvement of 25% and 40%, respectively.

This lower bound on the thresholds is in some sense artificial. The idea that the leakage current in a static CMOS circuit has to be zero is a preconception. Certainly, the presence of leakage currents degrades the noise margins, because the logic levels are no longer equal to the supply rails. As long as the noise margins are within range, this is not a compelling issue. The leakage currents, of course, cause an increase in static power dissipation. This is offset by the drop in supply voltage, that is enabled by the reduced thresholds at no cost in performance, and results in a quadratic reduction in dynamic power. For a 0.25 µm CMOS process, the following circuit configurations obtain the same performance: 3 V supply–0.7 V V_T; and 0.45 V supply–0.1 V V_T. The dynamic power consumption of the latter is, however, 45 times smaller [Liu93]! Choosing the correct values of supply and threshold voltages once again requires a trade-off. The optimal operation point depends upon the activity of the circuit. In the presence of a sizable static power dissipation, it is essential that non-active modules are *powered down*, lest static power dissipation would become dominant. Power-down (also called *standby*) can be accomplished by disconnecting the unit from the supply rails, or by lowering the supply voltage.

5.5.3 Putting It All Together

The total power consumption of the CMOS inverter is now expressed as the sum of its three components:

$$P_{tot} = P_{dyn} + P_{dp} + P_{stat} = (C_L V_{DD}^2 + V_{DD} I_{peak} t_s) f_{0\rightarrow 1} + V_{DD} I_{leak} \tag{5.52}$$

In typical CMOS circuits, the capacitive dissipation is by far the dominant factor. The direct-path consumption can be kept within bounds by careful design, and should hence not be an issue. Leakage is ignorable at present, but this might change in the not too distant future.

The Power-Delay Product, or Energy per Operation

In Chapter 1, we introduced the *power-delay product* (PDP) as a quality measure for a logic gate.

$$PDP = P_{av}t_p \tag{5.53}$$

The PDP presents a measure of energy, as is apparent from the units (Wsec = Joule). Assuming that the gate is switched at its maximum possible rate of $f_{max} = 1/(2t_p)$, and ignoring the contributions of the static and direct-path currents to the power consumption, we find

$$PDP = C_L V_{DD}^2 f_{max} t_p = \frac{C_L V_{DD}^2}{2} \tag{5.54}$$

The PDP stands for the **average energy consumed per switching event** (this is, for a 0→1, or a 1→0 transition). Remember that earlier we had defined E_{av} as the average energy per switching cycle (or per energy-consuming event). As each inverter cycle contains a 0→1, and a 1→0 transition, E_{av} hence is twice the PDP.

Energy-Delay Product

The validity of the PDP as a quality metric for a process technology or gate topology is questionable. It measures the energy needed to switch the gate, which is an important property for sure. Yet for a given structure, this number can be made arbitrarily low by reducing the supply voltage. From this perspective, the optimum voltage to run the circuit at would be the lowest possible value that still ensures functionality. This comes at the major expense in performance, at discussed earlier. A more relevant metric should combine a measure of performance and energy. The energy-delay product (EDP) does exactly that.

$$EDP = PDP \times t_p = P_{av}t_p^2 = \frac{C_L V_{DD}^2}{2} t_p \tag{5.55}$$

It is worth analyzing the voltage dependence of the EDP. Higher supply voltages reduce delay, but harm the energy, and the opposite is true for low voltages. An optimum operation point should hence exist. Assuming that NMOS and PMOS transistors have comparable threshold and saturation voltages, we can simplify the propagation delay expression Eq. (5.21).

$$t_p \approx \frac{\alpha C_L V_{DD}}{V_{DD} - V_{Te}} \tag{5.56}$$

where $V_{Te} = V_T + V_{DSAT}/2$, and α technology parameter. Combining Eq. (5.55) and Eq. (5.56), [4]

[4] This equation is only accurate as long as the devices remain in velocity saturation, which is probably not the case for the lower supply voltages. This introduces some inaccuracy in the analysis, but will not distort the overall result.

$$EDP = \frac{\alpha C_L^2 V_{DD}^3}{2(V_{DD} - V_{TE})} \tag{5.57}$$

The optimum supply voltage can be obtained by taking the derivative of Eq. (5.57) with respect to V_{DD}, and equating the result to 0.

$$V_{DDopt} = \frac{3}{2} V_{TE} \tag{5.58}$$

The remarkable outcome from this analysis is the low value of the supply voltage that simultaneously optimizes performance and energy. For sub-micron technologies with thresholds in the range of 0.5 V, the optimum supply is situated around 1 V.

Example 5.15 Optimum supply voltage for 0.25 μm CMOS inverter

From the technology parameters for our generic CMOS process presented in Chapter 3, the value of V_{TE} can be derived.

$$V_{Tn} = 0.43 \text{ V}, V_{Dsatn} = 0.63 \text{ V}, V_{TEn} = 0.74 \text{ V}.$$
$$V_{Tp} = -0.4 \text{ V}, V_{Dsatp} = -1 \text{ V}, V_{TEp} = -0.9 \text{ V}.$$
$$V_{TE} \approx (V_{TEn} + |V_{TEp}|)/2 = 0.8 \text{ V}$$

Hence, $V_{DDopt} = (3/2) \times 0.8 \text{ V} = 1.2 \text{ V}$. The simulated graphs of Figure 5.36, plotting normalized delay, energy, and energy-delay product, confirm this result. The optimum supply voltage is predicted to equal 1.1 V. The charts clearly illustrate the trade-off between delay and energy.

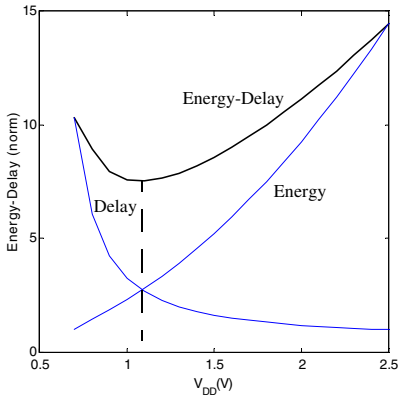


Figure 5.36 Normalized delay, energy, and energy-delay plots for CMOS inverter in 0.25 μm CMOS technology.

WARNING: While the above example demonstrates that there exists a supply voltage that minimizes the energy-delay product of a gate, this voltage does not necessarily represent the optimum voltage for a given design problem. For instance, some designs require a minimum performance, which requires a higher voltage at the expense of energy. Similarly, a lower-energy design is possible by operating by circuit at a lower voltage and by