Optimization of power consumption is one of the biggest challenges IC designers face today. Although power optimization has always been critical for battery-operated designs, the continued growth of system performance with each new generation of semiconductor technology, along with the increasing emphasis on "green" and "clean" technical applications, has made power optimization essential even for wall-powered designs. Effective power management involves selection of the right technology, the use of optimized libraries and IP (intellectual property), and design methodology (Figure 1). It also means optimizing both active dynamic power and static leakage power. This article examines the various approaches to effective power management.

![Figure 1](image-url) Effective power management requires proper technology selection, library, and IP design, and chip-design methods.

Power consumption is becoming more important in electronics. With increasing emphasis on reducing energy consumption in products, system designers must take more care in managing their power budgets. As a result, managers slash chip power budgets and maintain cost and performance goals. Unfortunately, the migration to finer technologies is compounding the problem. First, leakage power increases significantly with finer geometries and is becoming a key component of the total power. In addition, the finer geometries do not provide the voltage scaling that previous generations of technology enjoyed. As a result, the power saving due to voltage scaling is no longer substantial. In addition, from a design perspective, chips in the new generation of technology normally see a major increase in features and functions. Taking all these factors into account makes power management a significant challenge for most designers. So designers need an intelligent approach for optimizing power consumptions in designs.

**Basic Operation of MOS Transistors** To understand power, begin with the classic MOS-transistor...
equations for the drain current. Although these equations are accurate only for older technologies and do not take into account various effects that the submicron geometries in modern technologies introduce, they provide an understanding of the overall behavior of the transistor.

In digital circuits, when the transistor is on, it is in the saturation region, in which the following equation governs the drain-to-source current, $I_{ds}$ (Figure 2):

$$I_{ds} = K \frac{1}{T_{ox}} \frac{W}{L} (V_{gs} - V_{th})^2.$$  \hspace{1cm} (1)

where $T_{ox}$ is the gate-oxide thickness, $W$ is the channel width of the transistor, $L$ is the channel length of the transistor, $V_{gs}$ is the voltage between the gate and the source of the transistor, $V_{th}$ is the threshold voltage, and $K$ is dependent on the process technology. The following equation governs the threshold voltage:

$$V_{th} = V_{fb} + \Theta_s + \gamma \sqrt{V_{sb} + \Theta_s}.$$  \hspace{1cm} (2)

where $V_{sb}$ is the back-bias voltage between the source and the substrate; $V_{fb}$ is the flat-band voltage, which depends on the process technology; and $\gamma$ and $\Theta_s$ are parameters that also depend on the process technology.

![Figure 2](image)

**Figure 2** An NMOS FET shows the voltages you apply at its terminals.

If the drain-to-drain voltage is the power-supply voltage—that is, the maximum voltage that can be between the gate and the source, you can use the following equation to calculate the on current:

$$I_{on} = K \frac{1}{T_{ox}} \frac{W}{L} (V_{dd} - V_{th})^2.$$  \hspace{1cm} (3)

You can then express the active power as:

$$P_{active} = I_{on} V_{dd} = K \frac{1}{T_{ox}} \frac{W}{L} (V_{dd} - V_{th})^2 V_{dd}.$$  \hspace{1cm} (4)

**Leakage Power** The main components of leakage in a MOS transistor are junction leakage, gate leakage, gate-induced gate leakage, and subthreshold conduction. Junction leakage occurs when the PN junction between the drain and the substrate or the source and the substrate becomes negatively biased when the transistor is off, resulting in a leakage current due to the presence of the reverse-
bias diode. Gate leakage occurs in the presence of a high electric field in the gate oxide, causing electrons to tunnel through the gate into the substrate and resulting in gate leakage. As transistor geometries shrink, the gate-oxide thickness also decreases, making it more prone to tunneling. However, new high-k dielectric materials for the gate oxide have managed to control and minimize this leakage.

Gate-induced drain leakage occurs when the high electric fields in the gate-to-drain overlap region cause band-to-band tunneling and result in gate-induced drain-leakage current. Subthreshold conduction occurs when the transistor is off; it is not truly off but conducts due to weak inversion. Subthreshold conduction is the main contributor of leakage current. You can express this current as:

\[
I_{DS} = K_i \frac{W}{L} \exp \left( \frac{V_{GS} - V_{TH} - \eta V_{DS}}{NV_T} \right) \left( 1 - \exp \left( -\frac{V_{DS}}{V_T} \right) \right),
\]

where \(K_i\), \(\gamma\), \(\eta\), and \(N\) are technology-dependent, \(V_T\) is the thermal voltage, and \(K_i\) is a function of the gate-oxide thickness. You can obtain the off current or leakage current in a transistor by setting the gate-to-source voltage at 0V and setting the drain-to-source voltage to equal the power-supply voltage, \(V_{DD}\). Under these conditions, you can approximate the term

\[
1 - \exp \left( \frac{-V_{DS}}{V_T} \right)
\]

to 1 because the power-supply voltage is much greater than the thermal voltage, leading to

\[
I_{OFF} = K_i \frac{W}{L} \exp \left( \frac{-V_{TH} - \gamma V_{DD} + \eta V_{DD}}{NV_T} \right).
\]

You can now write the leakage power as

\[
P_{LEAKAGE} = I_{OFF} V_{DD} = K_i \frac{W}{L} \exp \left( \frac{-V_{TH} - \gamma V_{DD} + \eta V_{DD}}{NV_T} \right) V_{DD}.
\]

From this result, you can see that the main parameters controlling the power are the threshold voltage, the oxide thickness, the transistor length and width, the power-supply voltage, and the back-gate bias. Because active power varies as the square of the power-supply voltage, reducing the power-supply voltage has the most impact on reducing active power. The reduction in power is twice the amount of reduction in voltage—that is, a 20% reduction in power-supply voltage yields a 40% reduction in active power. The remaining parameters affect active power only linearly. Any significant change in length, width, or threshold voltage has adverse effects on the performance of the transistor. As a result, designers can change these parameters only by small amounts, and they therefore have only a small role on reducing active power. However, they do have a significant impact in reducing leakage power because they are exponentially related to it. From Equation 5, you can see that

\[
\frac{I_{DS}(V_{GS} + \Delta V_{GS})}{I_{DS}(V_{GS})} = \exp \left( \frac{\Delta V_{GS}}{V_T} \right).
\]

If \(\Delta V_{GS} = -\eta V_T\), the equation becomes
meaning that the subthreshold current decreases by a factor of 2.71828 for every NVₜ reduction in effective gate-to-source voltage. N is typically 1 to 2.5 for a technology, and the threshold voltage is 26 mV at room temperature, so, for every 50- to 75-mV change in gate-to-source voltage, you can see a reduction of 2.7 in subthreshold current. Increasing the threshold voltage has the same effect. Thus, for every 50- to 75-mV increase in threshold voltage, the leakage current decreases by a factor of 2.7. A 100- to 150-mV increase in threshold voltage reduces leakage by a factor of 7.4.

You can further reduce leakage current by increasing the back-gate bias. The gain is less significant due to the presence of the body-bias coefficient, γ. Reducing the power-supply voltage also helps to reduce leakage current. Increasing the channel length of the transistor not only directly reduces leakage current, as in Equation 5, but also helps to increase the threshold voltage, as in Equation 2.

The subthreshold current has an exponential dependence on temperature. Because the term NVₜ appears in the denominator of the negative exponent, as the temperature increases, the current increases significantly. This increase poses a major challenge because leakage power becomes a significant component of total power at high temperatures. So you must consider the total power at high temperature for fast-process-corner devices for worst-case power analysis.

Now that you understand the parameters that affect active and leakage power, you should examine how you can control these parameters using process-technology and design methods.

**The Role of Technology Selection** Proper technology selection is one of the key aspects of power management. The goal of each technological advancement is to improve performance, density, and power consumption. The typical approach in developing a new generation of technology is to apply constant-electric-field scaling. Process designers scale both the applied voltage and the oxide thickness to maintain the same electric field. This approach reduces power by about 50% with every new technology node. However, as the voltage gets smaller, the threshold voltage also must scale down to meet the performance targets of that technology. This scaling unfortunately increases the subthreshold current and hence the leakage power. To overcome this constraint, process engineers no longer apply constant-field scaling for processes of 65 nm or smaller; instead, they used a more generalized form of scaling.

Because it is impossible to optimize a technology for both performance and leakage at once, each technology usually has two variants. One variant aims for high performance, and the other shoots for low leakage. The primary differences between the two are in the oxide thickness, supply voltage, and threshold voltage. The technology variant with the thicker gate oxide aims for low-leakage design and must support a higher voltage to achieve a reasonable performance.

Equation 2 shows technology-dependent parameters γ and Θₛ, which you can manipulate to control the threshold voltage. These parameters depend on the doping concentration, which process designers can adjust by using an additional implant mask. This adjustment allows you to use one technology to create devices with multiple threshold voltages. You can then use this method to control leakage power in your designs.

When selecting a technology to optimize the power for a given design, you must take both aspects into consideration: the need to use a smaller geometry to reduce active power and the need to use a low-leakage variant to reduce leakage. There is a trade-off, however, with cost and risk.
Smaller geometries require a higher initial investment in mask costs and other NRE (nonrecurring-engineering) expenditures. Although they provide an advantage in unit cost because more units are available per wafer, they also pose a higher risk in process and design maturity. The design risk can be high if the design contains complex circuits such as SERDES (serializers/deserializers) or other sensitive blocks that are new in that process. The process risk depends on how long the technology has been in full production at the foundry. A new technology usually takes a year or more of production to iron out all the process kinks and to provide stable yields.

Should you aim for high performance or low leakage, and which is suitable for optimizing power? The answer to this question depends on both the nature of the power and the end application. If the end application is battery-powered, then you must minimize leakage. This constraint might automatically lead you to select a low-leakage technology, but that scenario need not always occur. For example, if you can turn off the design in standby mode, your design doesn't need a low-leakage process because you can turn off circuits in a high-performance system and also achieve the benefits of low leakage.

Low-leakage processes also use higher voltages and typically have larger areas and therefore consume more active power for the same performance. Leakage power is therefore the primary driver of the selection of a low-leakage process. Selecting a low-leakage process meets the requirements when leakage power becomes a significant component of the total power in a design during its active operation or when a design has a stringent requirement for the power in the standby mode that leakage dominates. In most other situations, you may select the standard process along with a variety of circuit-design techniques to optimize power.

**Circuit-Design Techniques** Once you select a technology, you can focus on the design techniques with which to optimize power. Start with the basic building block in a digital circuit: the logic gate. Logic gates are typically parts of a standard cell library. Each gate in a standard cell library uses the smallest transistors. Each type of gate has multiple versions with different drive strengths that employ wider transistors or multiple stages for more drive current. Because the main parameter for controlling active power is the power-supply voltage, cell designers typically design and characterize the gates to operate at voltages as much as 30% lower than the power-supply voltage. This voltage has performance implications. Lowering the power-supply voltage produces smaller currents, resulting in a longer time for charging and discharging the same capacitance. As a result, the design gets slower. However, this slowdown is acceptable if the design is not pushing the edges of a given technology.

Increasing the threshold voltage reduces the leakage current in the device. You can control leakage power by designing logic gates with multiple-threshold-voltage devices, including standard-, high-, and low-threshold-voltage devices. It is now common practice to design standard cell libraries with multiple-voltage-threshold devices. There is a trade-off between leakage and performance for a NAND gate that you implement with standard-, high-, and low-threshold-voltage devices (Figure 3). You can mix cells from these threshold-voltage libraries to optimize power in a chip.
The next factor is channel length. Cell designers create the logic gates in a standard cell library with minimum-channel-length devices. By increasing the channel length, you can reduce the leakage current in the device, but doing so also reduces the on current in the transistor and slows it down, so you can do this task only in small increments. Standard-cell-library providers recently created standard cells with multiple channel lengths. A combination of multiple-threshold-voltage devices and multiple channel lengths provides a rich standard-cell library for power management.

Another technique is back biasing. Traditionally, digital designers have viewed a MOS transistor as a three-terminal device in which the substrate ties to the source. As a result, the back-bias voltage is always 0V. By making the substrate available as a separate terminal and applying a reverse bias, you can increase the threshold voltage and lower leakage. You connect N-channel device substrates to a high negative voltage and P-channel device substrates to a high positive voltage. You need a large voltage for a small change in threshold because of the square-root relationship of the back-bias voltage to the threshold voltage as well as the presence of the body bias coefficient, γ. However, you apply back bias only in standby mode so that you don’t affect the performance of the device.

The same techniques also apply to memory design. Memories can have high-threshold-voltage devices in both their bit cells and their peripheral circuitry as well as reverse-bias control to manage leakage in the off state. Using different combinations of threshold-voltage devices for the bit cell and peripheral circuitry provides extensive control of memory leakage and various levels of performance. Lowering the supply voltage in memories has a significant performance penalty. For this reason, memories typically require dual power supplies—a higher voltage for the bit cell and a lower voltage for the peripheral circuitry.

**Power Management** After examining power-management techniques at the circuit level, you can look at techniques at the chip level. The first is the use of power switches to shut down circuits when they are not in operation. In shutdown mode, the circuit consumes only leakage power and consumes no active power. You can further reduce leakage-power consumption by applying a back-gate bias. You shut down power by using MOSFETs as switches that connect to the power supply and ground rails (Figure 4). When implementing shutdown, you must consider how the circuit wakes up and you must sometimes preserve the state of the design. In this case, you can use retention flip-flops to store the state. These flip-flops remain on in the shutdown state so that they can recover the state of the circuit when it wakes up. There is a small penalty in the form of wake-up recovery time.
You can use header and footer switches to turn off logic to save active power. You can use power switches to provide multiple levels of granularity in controlling which parts of the design need to shut down. You can switch power at the gate level, accompanying each gate by header and footer switches that connect to the power supplies. Alternatively, you can use header and footer switches with clusters of logic or at the block level with power islands. You can also use power islands without power switches by simply connecting the islands to different power supplies, which the design then turns on or off externally. These power supplies can have the same or different values. Power islands require the use of isolation cells at their boundaries. These cells ensure that the inputs to the island that is shut down are also off, so that there are no spurious currents.

A multiple-power-supply design has power islands with different values (Figure 5). This technique allows slower blocks of logic to run at lower voltages, thereby saving power. For multiple-power-supply designs, you must insert level-shifter cells at the island boundaries. These cells translate the logic into the appropriate levels of the island to which they interface. The UPF (Unified Power Format) language enables chip designers to describe designs with power gating and multiple power supplies. It allows the definition of power-supply domains for multiple-power-supply operation. It also allows the definition of isolation cells, level shifters, and power-gating switches. CPF (Common Power Format), a similar language, has the same purpose. These languages are currently competing to become the standard for defining power management in designs.
Figure 5  In a chip with multiple power domains, the low-performance portions use lower-power supplies to reduce power. Level shifters properly interface the logic in the different domains. Today's EDA tools effectively support these power-management techniques. They also provide additional power savings during implementation. Because the clock network and the flip-flops they drive consume a significant amount of power, you can achieve power savings by turning off clocks when you don't need them to be running—that is, gating the clocks. Clock gating eliminates the clock activity in a flip-flop during cycles when its input is inactive (Figure 6). Clock gating can achieve active-power savings of more than 30%.

Figure 6 Clock gating eliminates the clock activity in a flip-flop during cycles when its input is inactive.

You can also optimize power in clock-distribution networks. By using cloning techniques, you break up the clock tree into smaller sections, thereby reducing the total capacitance in the clock network and lowering its power. The physical-optimization process also takes power into account. Once you meet the timing constraints, physical optimization downsizes the gates in the noncritical paths to reduce power without affecting timing.

Leakage Optimization The main approach for optimizing leakage power is the use of standard cell libraries with multiple-voltage-threshold devices. Many tools allow a designer to use multiple libraries during physical implementation and automatically select cells from the appropriate library to optimize leakage power and meet performance targets. However, use this feature carefully because a design's area can sometimes become larger. Higher-threshold-voltage cells are weak, so your design may need larger cells to meet timing. In a mixed-threshold-voltage design, 80% of the cells typically have high threshold voltages, and the remaining 20% have either standard or low threshold voltages. You should use low-threshold-voltage devices sparingly and only in areas in which performance is critical because they contribute to leakage current. You can combine multiple-channel-length libraries with multiple-threshold-voltage devices to provide additional flexibility.

Another possibility is to use the TSMC (Taiwan Semiconductor Manufacturing Co) Power-Trim service, which varies the channel length of transistors in noncritical paths, virtually without affecting the layout of the design. The technique applies a bias to the polysilicon mask, which instructs the mask-making process to make adjustments to increase the effective channel length of the transistor. Power-Trim does this task as a postprocessing step during manufacturing, and it has the advantage of not affecting the design schedule.

Once the design has met its performance targets, Power-Trim uses software that Tela acquired from
Blaze DFM to analyze the design and tag the transistors whose channel lengths could increase. Typically, these devices are in the noncritical paths of the design. The tool increases the channel lengths in predefined increments, which has a precharacterized standard-cell library. The tool performs a timing analysis with the modified gates to ensure that there is no impact on the chip’s performance. This technique can provide an additional 20 to 30% savings in leakage power. Because this technique modifies only the transistors in the standard-cell library, it is meaningful only in designs in which digital logic dominates and leakage power is a significant component of the total power.

Another aspect of power management that engineers sometimes overlook is power integrity. Power integrity affects both the core and the I/O power in a chip. You must take care to distribute the power in the core, especially in the case of a multiple-power-supply design and when the external power delivery is by means of a wire-bond package. In a typical flip-chip design, the availability of a large number of bumps, especially in the core region of the chip, enables distribution of power to the core with minimal IR (current/resistance) drop and minimal signal-integrity effects. For a wire-bond package, however, you must perform careful analysis to ensure that you have allocated enough I/O buffers for power and ground to accommodate the core power requirements.

IR drop and EM (electromigration) are other key areas of concern in the core region. You must ensure that the worst-case supply voltage in the core region does not fall below 10% of the nominal value, meaning that the total variation in power supply across the package and the die should not exceed 10%. The external supply itself typically has 5% tolerance, which means that you typically need an IR drop of 5% or better on the die. Otherwise, you must have a smaller tolerance on the external supply, which significantly increases the cost of its voltage regulator. This requirement often dictates the number of power and ground I/O buffers on the die and the choice of the thickness and width of the top metal layers on which you will design the power mesh.

In addition to IR drop, you must satisfy the EM criteria. The EM current-density limit is the current density above which metal migration occurs, resulting in irreversible damage to the metal layer and eventually causing an open connection. The EM current densities are significantly tighter at higher temperatures. For example, a 10° increase in operating temperature from 110 to 120°C requires a doubling of the metal-trace width because the EM-current-density limit at 120°C is only one-half of the limit at 110°C. So you must take into account the EM criteria at the maximum operating temperature of the die when you determine the number of I/O buffers for power and ground.

Finally, you will need to insert decoupling capacitors in the core-and sometimes in the package-to smooth out large peaks in the core current. Also, when a chip includes multiple power domains in which large blocks of logic switch off and on, a key design consideration is to ensure that there is enough decoupling capacitance or phase management to ensure the integrity of the turn-on operation during any sudden surge in operating currents. There is generally plenty of room available for such devices on the core. Most standard-cell libraries provide decoupling-capacitor cells that you can place in the unused portion of the standard cell regions in a chip. In addition, you can build custom cells for use in other areas of the chip. Note that you must design decoupling-capacitor cells with low leakage, however, because they can otherwise contribute to significant additional leakage. This leakage can be a challenge because lower leakage also means lower capacitance.

**Author Information**
Prasad Subramaniam, PhD, is vice president of design technology at eSilicon Corp, where he is responsible for developing the technology platforms for IC design. He has a wide range of experience in ASIC, custom, and mixed-signal design. Subramaniam is a senior member of the IEEE and has published more than 40 papers in technical conferences and journals. He received a doctorate in electrical engineering from the State University of New York-Stony Brook.

References