Standby and Active Leakage Current Control and Minimization in CMOS VLSI Circuits

Farzan Fallah

Fujitsu Labs. of America, Inc.
Advanced CAD Technology Group
San Jose, California

Massoud Pedram

University of Southern California
Dept. of Electrical Engineering
Los Angeles, California

Abstract: In many new high performance designs, the leakage component of power consumption is comparable to the switching component. Reports indicate that 40% or even higher percentage of the total power consumption is due to the leakage of transistors. This percentage will increase with technology scaling unless effective techniques are introduced to bring leakage under control. This article focuses on circuit optimization and design automation techniques to accomplish this goal. The first part of the article provides an overview of basic physics and process scaling trends that have resulted in a significant increase in the leakage currents in CMOS circuits. This part also distinguishes between the standby and active components of the leakage current. The second part of the article describes a number of circuit optimization techniques for controlling the standby leakage current, including power gating and body bias control. The third part of the article presents techniques for active leakage control, including use of multiple-threshold cells, long channel devices, input vector design, transistor stacking to switching noise, and sizing with simultaneous threshold and supply voltage assignment.

1. INTRODUCTION

With the rapid progress in semiconductor technology, chip density and operation frequency have increased, making the power consumption in battery-operated portable devices a major concern. High power consumption reduces the battery service life. The goal of low-power design for battery-powered devices is thus to extend the battery service life while meeting performance requirements. Reducing power dissipation is a design goal even for non-portable devices since excessive power dissipation results in increased packaging and cooling costs as well as potential reliability problems.

Portable electronic devices tend to be much more complex than a single VLSI chip. They contain many components, ranging from digital and analog to electro-mechanical and electro-chemical. Much of the power dissipation in a portable electronic device comes from non-digital components. Dynamic power management – which refers to a selective, shut-off or slow-down of system components that are idle or underutilized – has proven to be a particularly effective technique for reducing power dissipation in such systems. Incorporating a dynamic power management scheme in the design of an already-complex system is a difficult process that may require many design iterations and careful debugging and validation.

IC power dissipation consists of different components depending on the circuit operating mode. First, the switching or dynamic power component dominates during the active mode of operation. Second, there are two primary leakage sources, the *active* component and the *standby* leakage component. The standby leakage may be made significantly smaller than the active leakage by changing the body bias conditions or by power-gating.

Voltage scaling is perhaps the most effective method of saving power due to the square law dependency of digital circuit active power on the supply voltage. Regrettably, scaling V_{DD} also reduces the circuit speed since the gate drive, $V_{GS} – V_T$, is reduced. To deal with this, systems may exploit dynamic voltage scaling to allow the lowest V_{DD} necessary to meet the circuit speed requirements while saving the energy used for the computation [1]-[5].

The current trend of lowering the supply voltage with each new technology generation has helped reduce the dynamic power consumption of CMOS logic gates. Supply voltage scaling increases the gate delays unless the threshold voltage of the transistors is also scaled down. The unfortunate effect of decreasing the threshold voltage is a significant increase in the leakage current of the transistors. Therefore, there is a clear tradeoff between the off-state leakage and the active power for a given application, leading to methodical selection of V_T and V_{DD} for performing a fixed task [6].

As device integration has led to a wide mixture of functions on a single die, it is increasingly difficult to find an optimal point applicable to all circuit blocks on a die. Consequently, design techniques, which can vary by circuit block, combined with device design, can result in higher quality designs.

2. SOURCES OF LEAKAGE POWER

There are four main sources of leakage current in a CMOS transistor (see Figure 1):

1. Reverse-biased junction leakage current (I_{REV})
2. Gate induced drain leakage (I_{GIDL})
3. Gate direct-tunneling leakage (I_G)
4. Subthreshold (weak inversion) leakage (I_{SUB})

as described next.

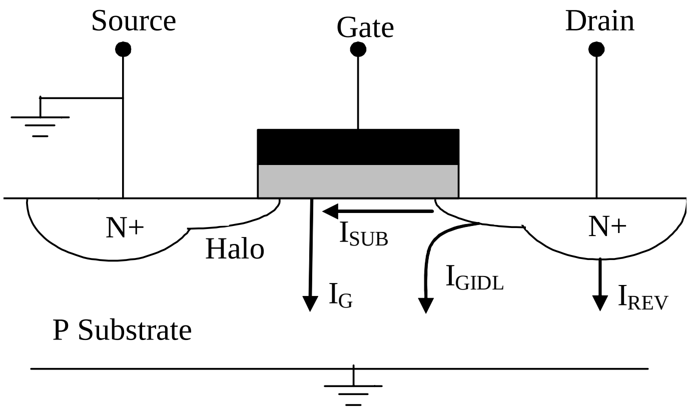


Figure 1: Leakage current components in an NMOS transistor.

2.1 Junction Leakage

The *junction leakage* occurs from the source or drain to the substrate through the reverse-biased diodes when a transistor is OFF. A reverse-biased pn junction leakage has two main components: one is minority carrier diffusion/drift near the edge of the depletion region; the other is due to electron-hole pair generation in the depletion region of the reverse-biased junction [7]. For instance, in the case of an inverter with low input voltage, the NMOS is OFF, the PMOS is ON, and the output voltage is high. Subsequently, the drain-to-substrate voltage of the OFF NMOS transistor is equal to the supply voltage. This results in a leakage current from the drain to the substrate through the reverse-biased diode. The magnitude of the diode's leakage current depends on the area of the drain diffusion and the leakage current density, which is in turn determined by the doping concentration. If both n and p regions are heavily doped, band-to-band tunneling (BTBT) dominates the pn junction leakage [8]. Junction leakage has a rather high temperature dependency (i.e., as much as 50 – 100 x/100 ^{o}C, but it is generally inconsequential except in circuits designed to operate at high temperatures (> 150^{o}C.) Junction reverse-bias leakage components from both the source-drain diodes and the well diodes are generally negligible with respect to the other three leakage components.

2.2 Gate-Induced Drain Leakage

The *gate induced drain leakage* (GIDL) is caused by high field effect in the drain junction of MOS transistors. For an NMOS transistor with grounded gate and drain potential at V_{DD}, significant band bending in the drain allows electron-hole pair generation through avalanche multiplication and band-to-band tunneling. A deep depletion condition is created since the holes are rapidly swept out to the substrate. At the same time, electrons are collected by the drain, resulting in GIDL current. This leakage mechanism is made worse by high drain to body voltage and high drain to gate voltage. Transistor scaling has led to increasingly steep *halo* implants, where the substrate doping at the junction interfaces is increased (cf. Figure 1), while the channel doping is low. This is done mainly to control punch-through and drain-induced barrier lowering while having a low impact on the carrier mobility in the channel. The resulting steep doping profile at the drain edge increases band to band tunneling currents there, particularly as V_{DB} is increased. Thinner oxide and higher supply voltage increase GIDL current. As an example, with a V_{DG}=3V and T_{ox} of 4nm, there is roughly a 10 fold increase in the GIDL current when V_{DB} is increased from 0.8V to 2.2V.

2.3 Gate Direct Tunneling Leakage

The gate leakage flows from the gate thru the "leaky" oxide insulation to the substrate. In oxide layers thicker than 3–4 nm, this kind of current results from the Fowler-Nordheim tunneling of electrons into the conduction band of the oxide layer under a high applied electric field across the oxide layer. For lower oxide thicknesses (which are typically found in 0.15μm and lower technology nodes), however, direct tunneling through the silicon oxide layer is the leading effect. Mechanisms for direct tunneling include electron tunneling in the conduction band (ECB), electron tunneling in the valence band (EVB), and hole tunneling in the valence band (HVB), among which ECB is the dominant one. The magnitude of the *gate direct tunneling current* increases exponentially with the gate oxide thickness T_{ox} and supply voltage V_{DD}. In fact, for relatively thin oxide thicknesses (in the order of 2-3 nm), at a V_{GS} of 1V, every 0.2nm reduction in T_{ox} causes a tenfold increase in I_G [9]. Gate leakage increases with temperature at only about 2x/100^{o}C. Note that the gate leakage for a PMOS device is typically one order of magnitude smaller than that of an NMOS device with identical T_{ox} and V_{DD} when using SiO_2 as the gate dielectric.

As transistor length and supply voltage are scaled down, gate oxide thickness must also be reduced to maintain effective gate control over the channel region. Unfortunately this results in an exponential increase in the gate leakage due to direct tunneling of electrons through the gate oxide [10]. An effective approach to overcome the gate leakage currents while maintaining excellent gate control is to replace the currently-used silicon dioxide gate insulator with high-K dielectric material such as TiO_2 and Ta_2O_5. Use of the high-k dielectric will allow a less aggressive gate dielectric thickness reduction while maintaining the required gate overdrive at low supply voltages [11]. According to the 2003 International Technology Roadmap for Semiconductors (ITRS-03) [12], high-K gate dielectric is required to control the direct tunneling current for low standby power devices in process technology nodes below 90 nm. High-K gate dielectrics are expected to be introduced in 2006.

2.4 Subthreshold Leakage

The *subthreshold leakage* is the drain-source current of a transistor operating in the weak inversion region. Unlike the strong inversion region in which the drift current dominates, the subthreshold conduction is due to the diffusion current of the minority carriers in the channel for a MOS device.[1] For instance, in the case of an inverter with a low input

[1] When the surface potential at the source end of the channel is sufficient to form an inversion layer, but the band bending is less than what is needed to reach strong inversion, a MOSFET is said to operate in weak inversion. When an n-channel MOSFET is in weak inversion, the drain current is determined by diffusion

voltage, the NMOS is turned OFF and the output voltage is high. In this case, although V_{GS} is 0V, there is still a current flowing in the channel of the OFF NMOS transistor due to the V_{DD} potential of the V_{DS}. The magnitude of the subthreshold current is a function of the temperature, supply voltage, device size, and the process parameters out of which the threshold voltage (V_T) plays a dominant role.

In current CMOS technologies, the subthreshold leakage current, I_{SUB}, is much larger than the other leakage current components [12]. This is mainly because of the relatively low V_T in modern CMOS devices. I_{SUB} is calculated by using the following formula:

$$I_{SUB} = \frac{W}{L}\mu v_{th}{}^{2} C_{sth} e^{\frac{V_{GS}-V_T+\eta V_{DS}}{n v_{th}}} (1-e^{\frac{-V_{DS}}{v_{th}}})$$

where W and L denote the transistor width and length, μ denotes the carrier mobility, $v_{th}=kT/q$ is the thermal voltage at temperature T, $C_{sth}=C_{dep}+C_{it}$ denotes the summation of the depletion region capacitance and the interface trap capacitance both per unit area of the MOS gate , and η is the drain-induced barrier lowering (DIBL) coefficient [13]. n is the *slope shape factor* and is calculated as:

$$n = 1 + \frac{C_{sth}}{C_{ox}}$$

where C_{ox} denotes the gate input capacitance per unit area of the MOS gate. When a long-channel transistor with V_{DS} larger than a few v_{th} is in the OFF state (V_{GS}=0), we have:

$$I_{SUB} = \frac{W}{L}\mu v_{th}{}^{2} C_{sth} 10^{\frac{-V_T}{S}}$$

where S denotes the subthreshold swing parameter, which is defined as the inverse of the slope of the $\log_{10}(I_{DS})$ versus V_{GS} characteristic and is equal to $n v_{th} ln(10)$. S is equal to the subthreshold voltage decrease required to increase I_{SUB} by a factor of ten.

It is highly desirable to have a subthreshold swing as small as possible since this is the parameter that determines the amount of voltage swing necessary to switch a MOSFET from OFF to ON state (typical values of S for bulk CMOS devices are 70-110 mV/decade; the theoretical lower bound is 60 mV/decade corresponding to n=1.) This is especially important for modern MOSFETs with supply voltages reaching sub-one volt region. To minimize S, the thinnest possible gate oxide (since it increases C_{ox}) and the lowest possible doping concentration in the channel (since it decreases C_{dep}) must be used. Higher temperature results in larger S value, and hence, an increase in the OFF leakage current.

of electrons from source to drain. This is because the drift current is negligibly small due to the low lateral electric field and small electron concentration in weak inversion.

In long channel devices, the influence of source and drain on the channel depletion layer is negligible. However, as channel lengths are reduced, overlapping source and drain depletion regions cause the depletion region under the inversion layer to increase. The wider depletion region is accompanied by a larger surface potential, which attracts more electrons to the channel. Therefore, a smaller amount of charge on the gate is needed to reach the onset of strong inversion and the threshold voltage decreases. This effect is worsened when there is a larger bias on the drain since the depletion region becomes even wider. More precisely, when a high drain voltage is applied to a short-channel device, it lowers the barrier for electrons between the source and the channel, resulting in further decrease of the threshold voltage. The source then injects carriers into the channel surface (independent of gate voltage), causing an increase in I_{OFF}. This phenomenon, which can be thought of as a lowering of V_T as V_{DS} increases, is the DIBL effect. There is yet another phenomenon known as the "V_T Rolloff" whereby the V_T of a MOSFET decreases as the channel length is reduced. In such a case, the subthreshold swing parameter degrades and the impact of drain bias on $V_{T.}$ increases. Finally, there is the well-known "body effect," which causes an increase in V_T as the body of the transistor is reverse-biased (i.e., V_{SB} of an NMOS transistor is increased).

Clearly, decreasing the threshold voltage increases the leakage current exponentially. In fact decreasing the threshold voltage by 100mV increases the leakage current by a factor of 10. Decreasing the length of transistors increases the leakage current as well. Therefore, in a chip, transistors that have smaller threshold voltage and/or length due to process variation contribute more to the overall leakage.

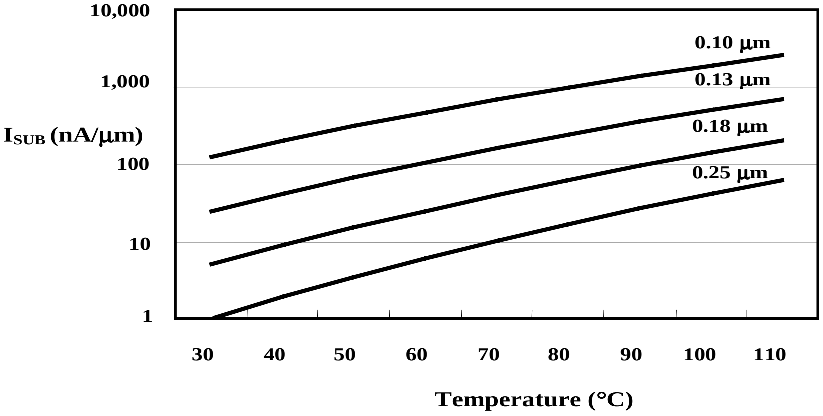


Figure 2: I_{SUB} (V_{GS}=0) trend as a function of temperature. Courtesy of Vivek De, Intel.

The subthreshold leakage current increases with temperature. Figure 2 shows the leakage current for several technologies for different temperatures. As one can see, I_{OFF} grows in each generation. Furthermore, in a given technology, the leakage current increases with the temperature. I_{off} has a temperature sensitivity of 8-12 x/100^{o}C.

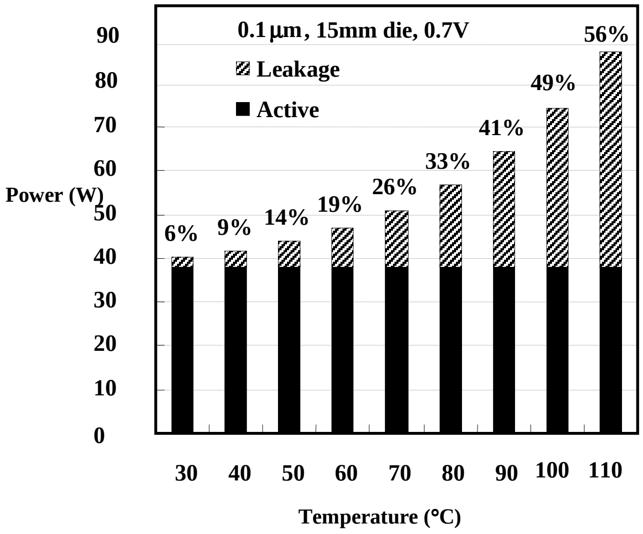


Figure 3: Power consumption of a die as a function of temperature. Courtesy of Vivek De, Intel.

Figure 3 shows the power consumption of a 15mm die fabricated in a $0.1\mu m$ technology with a supply voltage of 0.7V. Although the leakage power is only 6% of the total power consumption at 30°C, it becomes 56% of the total power at 110°C. This clearly shows the necessity of using leakage power reduction techniques in current designs.

2.5 Putting It All Together

Let I_{OFF} denote the leakage of an OFF transistor (V_{GS}=0V for an NMOS device.) From the above discussion, we know that:

$$I_{OFF} = I_{REV} + I_{GIDL} + I_{SUB}\ .$$

Clearly, I_{REV} and I_{GIDL} are maximized when $V_{DB} = V_{DD}$. Similarly, for short-channel devices, I_{SUB} increases with V_{DB} because of the DIBL effect. Note the I_G is not a component of the OFF current, since the transistor gate must be at a high potential with respect to the source and substrate for this current to flow. Among the three components of I_{OFF}, I_{SUB} is clearly the dominant component. So the remainder of this paper focuses on I_{SUB}. More precisely, in the next two sections, methods are presented for decreasing the subthreshold leakage currents in circuits that are in STANDBY or ACTIVE state.

3. LEAKAGE CONTROL IN STANDBY CIRCUITS

Most microelectronic systems spend considerable time in a standby state. The energy consumed by the logic and the DC-DC converter to enter or exit a low power mode must be considered carefully. If the cost of transitioning to and from a low standby power state is low enough then the greedy policy of entering the low power state as soon as the system is idle may be adopted. Otherwise, the expected duration of the standby state must be accurately calculated and taken into account when devising a power management policy.

3.1 Power Gating and Multi-Threshold CMOS

The most natural way of lowering the leakage power dissipation of a VLSI circuit in the STANDBY state is to turn off its supply voltage. This can be done by using one PMOS transistor and one NMOS transistor in series with the transistors of each logic block to create a virtual ground and a virtual power supply as depicted in Figure 4. Notice that in practice only one transistor is necessary. Because of their lower on-resistance, NMOS transistors are usually used.

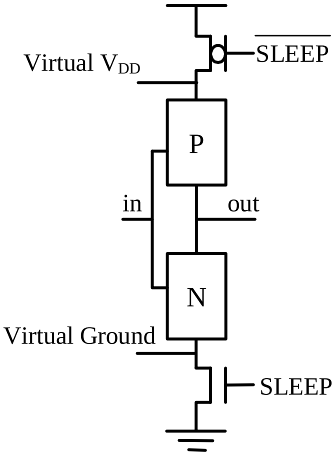


Figure 4: Power gating circuit.

In the ACTIVE state, the sleep transistor is on. Therefore, the circuit functions as usual. In the STANDBY state, the transistor is turned off, which disconnects the gate from the ground. Note that to lower the leakage, the threshold voltage of the sleep transistor must be large. Otherwise, the sleep transistor will have a high leakage current, which will make the power gating less effective. Additional savings may be achieved if the width of the sleep transistor is smaller than the combined width of the transistors in the pull-down network. In practice, Dual V_T CMOS or Multi-Threshold CMOS (MTCMOS) is used for power gating [14][15]. In these technologies there are several types of transistors with different V_T values. Transistors with a low V_T are used to implement the logic, while high-V_T devices are used as sleep transistors.

To guarantee the proper functionality of the circuit, the sleep transistor has to be carefully sized to decrease its voltage drop while it is on. The voltage drop on the sleep transistor decreases the effective supply voltage of the logic gate. Also, it increases the threshold of the pull-down transistors due to the body effect. This increases the high-to-low transition delay of the circuit. This problem can be solved by using a large sleep transistor. On the other hand, using a large sleep transistor increases the area overhead and the dynamic power consumed for turning the transistor on and off. Note that because of this dynamic power consumption, it is not possible to save power for short idle periods. There is a minimum duration of the idle time below which power saving is impossible. Increasing the size of the sleep transistors increases this minimum duration.

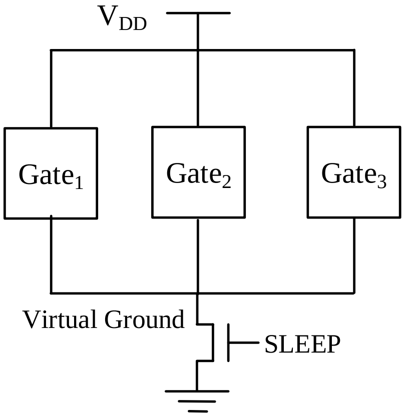


Figure 5: Using one sleep transistor for several gates.

Since using one transistor for each logic gate results in a large area and power overhead, one transistor may be used for each group of gates as depicted in Figure 5. Notice that the size of the sleep transistor in this figure ought to be larger than the one used in Figure 4. To find the optimum size of the sleep transistor, it is necessary to find the vector that causes the worst case delay in the circuit. This requires simulating the circuit under all possible input values, a task that is not possible for large circuits.

In [15], the authors describe a method to decrease the size of sleep transistors based on the mutual exclusion principle. In their method, they first size the sleep transistors to achieve delay degradation less than a given percentage for each gate. Notice that this guarantees that the total delay of the circuit will be degraded by less than the given percentage. In fact the actual degradation can be as much as 50% smaller. The reason for this is that NMOS sleep transistors degrade only the high-to-low transitions and at each cycle only half of the gates switch from high to low. If two gates switch at different times (i.e., their switching windows are non-overlapping), then their corresponding sleep transistors can be shared.

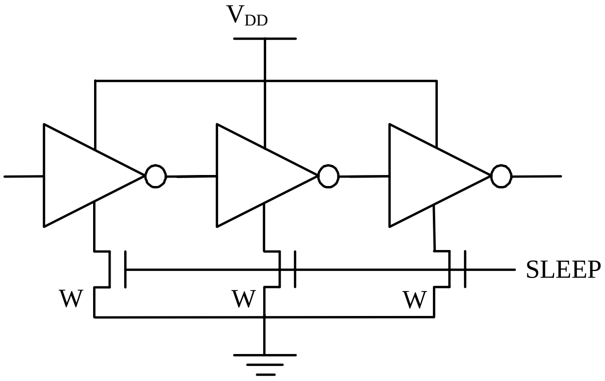


Figure 6: Sleep transistor sharing.

Consider the inverters in Figure 6. These inverters switch at different times due to their propagation delays. Therefore, it is possible to combine their sleep transistors and use one transistor instead of three. In general, if there are *n* logic gates whose output transition windows are non-overlapping, and each has a sleep transistor whose width is W_i, then these sleep transistors may be replaced with a single transistor whose width is $W_{eq} = Max\ W_i$ for $1 \leq i \leq n$. Notice that this will decrease the delay degradation of the logic gates whose corresponding sleep transistors are narrower than W_{eq}. Furthermore, if there are several sleep transistors corresponding to some logic gates with overlapping output transition windows, then these sleep transistors may be replaced by a single transistor whose width is,

$$W_{eq} = \sum_i W_i \,.$$

Using mutual exclusion at the gate level is not practical for large circuits. To handle large circuits, the mutual exclusion principle may be used at a larger level of granularity. In this case, a single sleep transistor is used for each module or logic block. The size of this sleep transistor is calculated according to the number of logic gates and complexity of the block. Next the sleep transistors for different blocks are combined as described before. This method enables one to “hide” the details of the blocks thus large circuits can be handled. However, in this case, the sizes of sleep transistors may be sub-optimal.

Mutual exclusion method gives an upper bound on the required size of the sleep transistor. In many cases this bound is not tight. For example for a chain of three inverters, mutual exclusion method overestimates the required width of the sleep transistor by 60%! This is partly because mutual exclusion does not take into account the fact that only about half of gates in a circuit switch from high to low in a given cycle [16]. To improve the technique it is possible to use logical information instead of structural information to find mutual exclusion. Furthermore, instead of restricting the delay degradation of each gate or path of the circuit, it is sufficient to limit the degradation of the delay of the circuit as a whole. For example in Figure 7 the delay of Output1 is less than the delay of Output2. Therefore, a 10% delay degradation for the circuit can be achieved even when the delay degradation of Output1 is more than 10%.

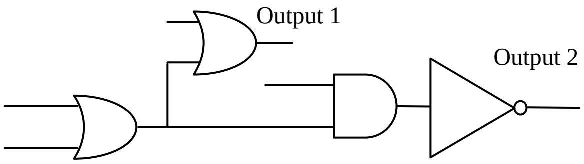


Figure 7 A circuit for which mutual exclusion does not perform well.

Anis et al. [17] propose a clustering method to improve the mutual exclusion technique. They simulate a circuit under different input vectors to find the current profile for each logic gate in the circuit. Next they use a bin-packing based algorithm to assign gates to different clusters. They compare their clustering technique with the mutual exclusion technique and the standard technique of sizing the sleep transistor of each gate to ensure a given switching speed degradation for the gate. They report factors of eight and one hundred reductions in the total width of sleep transistors for their clustering technique compared to these two other techniques, respectively. In addition, they report that the leakage current is reduced by factors of seven and twenty, respectively. Because the clustering technique does not take into account the physical distance of gates in the placed circuit netlist, the routing overhead can be very high. The authors suggest another algorithm based on set partitioning which uses the distance between gates in a circuit while doing the clustering. The new algorithm reduces the routing overhead, but at the cost of increasing the total width and leakage of the sleep transistors.

Although sleep transistors can be used to disconnect logic gates from ground, using them to disconnect Flip Flops from ground or supply voltage results in the loss of data. The authors of [18] solve this problem by using high threshold transistors for the inverters that hold data and low threshold transistors for other parts of Flip Flops. In the sleep mode, the low threshold transistors are disconnected from the ground, but the two inverters that hold data stay connected to the ground. Since high threshold transistors have been used in the inverters, their leakage is small. Other possibilities for saving data when MTCMOS is applied to a sequential circuit are to utilize high V_T devices placed in parallel with low V_T devices [14], leakage-feedback gates and flip flops [19], balloon latches [20], input-referred conditional cutoff [21], or scan-chain latches [22].

One of the drawbacks of using sleep transistors is that they generate noise in circuits. When a sleep transistor is off, some charge accumulates in the virtual ground. This raises the voltage of the virtual ground. When the sleep transistor is turned on, the resistance between the virtual ground and the ground decreases rapidly and large current flows to the ground of the circuit. This creates noise and may result in malfunctioning of other parts of the circuit. To solve this problem, reference [23] proposes circuit techniques along with sleep signal scheduling to gradually decrease the resistance between the virtual ground and ground. As a consequence, the maximum current flowing to the ground is reduced.

Power gating is a very effective method for decreasing the leakage power. However, it suffers from the following drawbacks:

1. It requires modification in the CMOS technology process to support both a high V_T device (for the sleep transistor) and a low V_T device (for logic gates.)
2. It decreases the voltage swing; therefore, it decreases the DC noise margin.
3. Supply voltage scale-down makes it necessary to decrease the threshold voltage of the sleep transistors in each generation. This means that the leakage current will continue to increase exponentially with each generation.
4. Scaling-down the supply voltage decreases the drive on all transistors. As a result, the on-resistance of the transistors increases. This increase is greater for sleep transistors because of their higher threshold voltage. As a result, larger sleep transistors should be used, which means the area overhead of the approach increases.
5. Sleep transistor sizing is a non-trivial task and requires much effort.

3.2 Body Bias Control and Power Supply Collapse

One of the methods proposed for decreasing the leakage current is using reverse-body bias (RBB) to increase the threshold voltage of transistors in the STANDBY state [24]. The threshold voltage of a transistor can be calculated from the following standard expression,

$$V_T = V_{T0} + \gamma\left(\sqrt{\left|-2\phi_F + V_{SB}\right|} - \sqrt{\left|2\phi_F\right|}\right)$$

where V_{T0} is the threshold voltage for $V_{SB}=0$, Φ_F is the substrate Fermi potential, and the parameter γ is the body-effect coefficient [25]. As one can see, reverse biasing a transistor increases its threshold voltage. Reverse biasing can be done during standby, by applying a strong negative bias to the NMOS bulk via a charge pump and connecting the PMOS bulks (N wells) to the V_{DD} rail. Note that reverse body biasing (RBB) is applied to the PMOS transistors by raising the N-well voltages. This method requires a triple-well technology, which may not always be available. Because the threshold voltage changes with the square root of the reverse bias voltage, a large voltage may be necessary to get a small increase in the threshold voltage. As a result, this method becomes less effective as the supply voltage is scaled down. On the positive side, with RBB, the IC logic state is retained while in the STANDBY mode, allowing operation to resume where it suspended. In other words, it is unnecessary to save the sate of the logic before entering the STANDBY state.

As stated previously, ignoring I_G, the leakage current has three main components: I_{REV}, I_{GIDL} and I_{OFF}. The last component is typically much larger than the first two. Bringing the voltage of NMOS Bulk below zero volts decreases I_{OFF}, but it increases I_{REV} and I_{GIDL}. This is because the strong bulk biasing increases GIDL and drain to bulk tunneling leakages to the point where they become limiting on advanced processes. To avoid this, RBB should use the lowest effective voltages. This also suggests there is an optimum substrate voltage for which the total leakage current is at a minimum. The optimum substrate voltage decreases by a factor of two, and the leakage reduction becomes less effective by a factor of four in each technology generation [26]. Therefore, this method may not be as effective in future technology generations.

Reference [27] proposes a combination of RBB and Power Supply Collapsing. The idea is that V_{SS} is increased to apply RBB while reducing V_{DD} - V_{SS} to approximately 350 mV with a V_{DD} value of 1 V. In addition to the RBB effect on I_{OFF}, this rail-to-rail voltage reduction limits GIDL, drain to bulk tunneling, and gate leakage components while applying approximately 650 mV body bias to the NMOS transistors. Notice that the amount of power supply collapse is limited because an excessive collapse of the core voltage would result in non-state-retentive sleep mode, which is undesirable in many applications. Briefly, state loss occurs when the total leakage current of the transistors holding a logic state exceeds that of the "on" transistors.

Alternatively, it is possible to use forward-body bias to decrease the threshold voltage [28]. In the STANDBY state, zero substrate bias is used to have a high V_T for low leakage. To decrease the gate delays while in the ACTIVE state, the threshold voltage is decreased by using a forward-body bias. This method has been successfully used to reduce the leakage power of a router chip by a factor of 3.5.

Finally, as in dynamic supply voltage scaling, it is possible to use a dynamically-varying threshold voltage. The idea is to adjust the V_T of devices by using dynamic body bias control. It is shown in [29] that the leakage current of an MPEG-4 chip can be driven below 10mA in ACTIVE state and below 10μA in STANDBY state.

4. LEAKAGE CONTROL IN ACTIVE MODE

For circuits whose power consumption is dominated by the leakage power in STANDBY mode, the aforementioned techniques can be used to reduce the power consumption. On the other hand, if a circuit's power consumption is not dominated by the leakage in STANDBY mode, then it is necessary to consider the total power consumption (including switching power dissipation and active mode leakage) and optimize the circuit to reduce it as described next.

4.1 Multiple Threshold Cells

Multiple threshold voltages have been available on many CMOS processes for a number of years [30]. Multiple-Threshold CMOS circuit, which has both high and low threshold transistors in a single chip, can be used to deal with the leakage problem. The high threshold transistors can suppress the subthreshold leakage current, while the low threshold transistors are used to achieve the high performance. Since the standby power is much larger for low V_T transistors compared to the high V_T ones, usage is limited to using low V_T transistors on timing-critical paths, with insertion rates on the order of 20% or less. Since T_{ox} and L_{gate} are the same for high and low V_T transistors, low V_T insertion does not adversely impact the active power component or the design size. Drawbacks are that variation due to doping is uncorrelated between the high and low threshold transistors and extra mask steps incur a process cost.

There are four classes of dual V_T designs based on how low and high threshold voltage transistors are mixed within a logic cell [31]:

1- Class 1: Using the same *type* of transistors (i.e., low threshold or high threshold) in a logic cell.
2- Class 2: Using the same type of transistors in a pull up or pull down network.
3- Class 3: Using the same type of transistors in a stack of transistors.
4- Class 4: Mixing transistors freely.

The technology used for fabricating circuits can restrict the manner in which transistors can be mixed. For example, it may not be possible to use different threshold voltages for transistors in a stack due to their proximity. Furthermore, to simplify the design process and Computer-Aided Design (CAD) algorithms, one may wish to restrict the way transistors are mixed. For example, when transistors of the same type are used in a logic cell, the size of multi-threshold cell library is only twice that of the original (single threshold) cell library. This reduces the library development time as well as the complexity and run time of CAD algorithms and tools that use the library.

In general, one expects that the leakage saving increases as the freedom to mix low and high V_T devices in a logic cell is increased. However, the percentage improvement is usually minor. Compared to Class 1 logic cells, reference [31] reports an average of only 5% additional leakage savings when Class 3 logic cells are used. Therefore, due to their simplicity, in many designs, only Class 1 logic cells are employed.

Although using two threshold voltages instead of one significantly decreases the leakage current in a circuit, using more than two threshold voltages marginally improves the result [32]. This is true even when the threshold values are optimized to minimize the leakage for a given circuit. Thus, in many designs, only two threshold voltages are used.

4.2 Long Channel Devices

Active leakage of CMOS gates can be reduced by increasing their transistor channel lengths [33]. This is because there is a V_T roll-off due to the Short Channel Effect (SCE). Therefore, different threshold voltages can be achieved by using different channel lengths. However, the shape of the V_T roll-off can be very sharp when HALO techniques are used [34]. In such technologies, it is non-trivial to control threshold voltages by using multiple channel lengths. The longer transistor lengths used to achieve high threshold transistors tend to increase the gate capacitance, which has a negative impact on the performance and dynamic power dissipation. Compared with multiple threshold voltages, long channel insertion has similar or lower process cost, taken as the size increase rather than the mask cost [35]. It results in lower process complexity. In addition, the different channel lengths track each other over process variation. The technique can be applied in a greedy manner to an existing design to limit the leakage currents. A potential penalty is that the dynamic power dissipation of the up-sized gate is increased proportional to the effective channel length increase. In general, circuit power dissipation may not be saved unless the activity factor of the affected gates is low. Therefore, the activity factor must be taken into account when choosing gates whose transistor lengths are to be increased.

4.3 Minimum Leakage Vector Method

The leakage current of a logic gate is a strong function of its input values. The reason is that the input values affect the number of OFF transistors in the NMOS and PMOS networks of a logic gate.

Table 1 shows the leakage current of a two-input NAND gate built in a 0.18μm CMOS technology with a 0.2V threshold voltage and a 1.5V supply voltage. Input A is the one closer to the output of the gate.

Table 1. The leakage values of a NAND gate.

Inputs		Output	Leakage Current (nA)
A	B	O	
0	0	1	23.06
0	1	0	51.42
1	0	0	47.15
1	1	0	82.94

The minimum leakage current of the gate corresponds to the case when both its inputs are zero. In this case, both NMOS transistors in the NMOS network are off, while both PMOS transistors are on. The effective resistance between the supply and the ground is the resistance of two OFF NMOS transistors in series. This is the maximum possible resistance. If one of the inputs is zero and the other is one, the effective resistance will be the same as the resistance of one OFF NMOS transistor. This is clearly smaller than the previous case. If both inputs are one, both NMOS transistors will be on. On the other hand, the PMOS transistors will be off. The effective resistance in this case is the resistance of two OFF PMOS transistors in parallel. Clearly, this resistance is smaller than the other cases.

In the NAND gate of *Table 1* the maximum leakage is about three times higher than the minimum leakage. Note that there is a small difference between the leakage current of the A=0, B=1 vector and the A=1, B=0 vector due to the body effect. The phenomenon whereby the leakage current through a stack of two or more OFF transistors is significantly smaller than a single device leakage is called the “stack effect”.

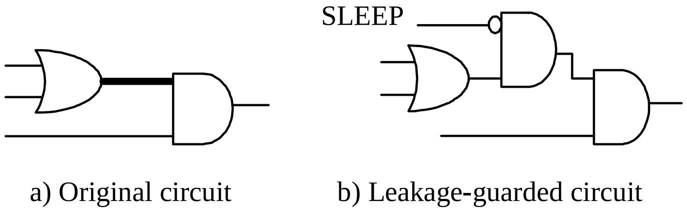
Other logic gates exhibit a similar leakage current behavior with respect to the applied input pattern. Reference [37] reports that the ratio of the maximum to the minimum leakage varies from 1.5 to 6 for circuits in MCNC91 benchmark suite. As a result, the leakage current of a circuit is a strong function of its input values. Abdollahi et al. [36] use this fact to reduce leakage current. They formulate the problem of finding the *Minimum Leakage Vector* (MLV) using a series of Boolean Satisfiability problems. Using the MLV to drive the circuit while in the STANDBY state, they reduce the circuit leakage. It is possible to achieve a moderate reduction in leakage using this technique, but the reduction is not as high as the one achieved by the power gating method. On the other hand, the MLV method does not suffer from many of the shortcomings of the other methods. In particular,

1. No modification in the process technology is required.
2. No change in the internal logic gates of the circuit is necessary.
3. There is no reduction in voltage swing.
4. Technology scaling does not have a negative effect on its effectiveness or its overhead. In fact the stack effect becomes stronger with technology scaling as DIBL worsens.

The first three facts make it very easy to use this method in existing designs.

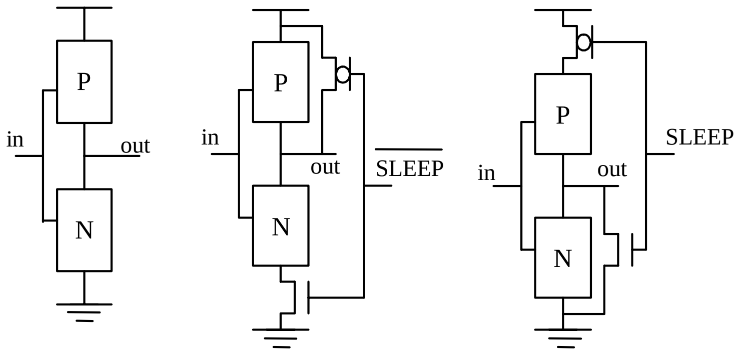
In [37], the authors report between 10% to 55% reduction in the leakage by using the MLV technique. Note that the saving is defined as $(1-\frac{Leakage_{MLV}}{Leakage_{AVG}})\times 100$, where $Leakage_{MLV}$ is the leakage when the minimum leakage vector drives the circuit whereas $Leakage_{AVG}$ is the expected leakage current under an arbitrary input combination (this is used because the input value prior to entering the sleep mode is unknown.)

Further reduction in leakage may be achieved by modifying the internal logic gates of a circuit. Note that due to logic dependencies of the internal signals, driving a circuit with its MLV does not guarantee that the leakage currents of all its logic gates are at minimum values. Therefore, when in the STANDBY state, if, by some means, values of the internal signals are also controlled, even higher leakage savings can be achieved. One way to control the value of an internal signal (line) of a circuit is to replace the line with a 2-to-1 multiplexer [37]. The multiplexer is controlled by the *SLEEP* signal whereas its data inputs are the incoming signal and either a ZERO or ONE value decided by the leakage current minimization algorithm. The output is the outgoing signal. Since one input of the multiplexer is a constant value, the multiplexer can be replaced by an AND or an OR gate. Figure 8 shows a small circuit and its modified version where the internal signal line can explicitly be controlled during the STANDBY state.



a) Original circuit b) Leakage-guarded circuit

Figure 8: Replacing an internal signal line with a two-input AND gate to increase controllability in the STANDBY state.



A) out = g(in) B) out = NAND ($\overline{\text{SLEEP}}$, g(in)) C) out =NOR (SLEEP, g(in))

Figure 9: A CMOS gate and its two modified versions that exhibit higher controllability in the STANDBY state.

In Figure 8*(b),* when the circuit is in the STANDBY state, the output of the AND gate is ZERO; if a ONE on that line is desired, the AND gate has to be replaced by an OR gate. Note that extra gates added to the circuit consume leakage power. Therefore, replacing all internal lines with multiplexers or gates will increase the leakage. The problem of determining which lines to replace and also finding the MLV for primary inputs and the selected internal signals can be formulated using a series of Boolean Satisfiability problems and solved accordingly as shown in [37].

Another way of controlling the value of the internal signals of a circuit is modifying its gates. Figure 9 shows two ways of modifying a CMOS gate. In both cases a transistor is added in series with one of the N or P networks. This decreases the gate leakage because of the transistor stack effect. The percentage of the reduction depends on the type of the gate. In addition, as mentioned before, this modification makes it possible to control the values of the internal lines in the circuit thus decreasing the leakage current of the gates in the fanout of the modified gate.

Clearly, adding transistors to gates increases the delay of the circuit. The problem of finding the minimum leakage vector and the optimal set of gates to be modified in order to minimize the leakage of the circuit under a delay constraint can also be formulated as a series of Boolean Satisfiability problems and solved accordingly [37]. Applying this augmented MLV technique to the circuits in the MCNC91 benchmark results in an additional 15-20% leakage saving.

4.4 Stack Effect-based Method

If the value of the input of a circuit in STANDBY mode is known, some NMOS and PMOS transistors can be added in series with gates to increase the stack effect and reduce the leakage current as a result [38]. In Figure 10, the output of gate (a) is high when it is in the STANDBY mode. This means that the pull down network is OFF. Therefore, putting a transistor in series with the pull down network and keeping it off in the STANDBY mode will not change the value of the output. However, it will increase the resistance between the supply and ground (see circuit (b)). Therefore, the leakage of the logic gate is reduced. Notice that if the output of the gate was low, then adding the transistor and turning it off would make the output of the gate float. This could potentially create a problem as short circuit current flows through the gates in the fanout of a floating node. This is an important consideration any time a logic cell that has been put to sleep is driving some other logic cells that are not in sleep. Reference [38] has captured this constraint by ensuring that no gate in the circuit can have a floating output. The authors report that an average of 65% reduction in the leakage can be achieved by using this method. Higher savings may be achieved if high threshold sleep transistors are used.

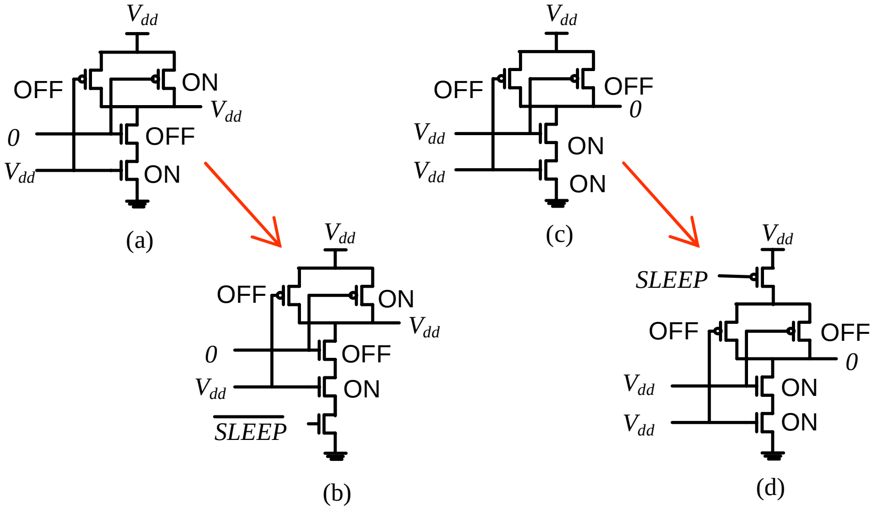


Figure 10 Reducing the leakage of logic gates using stack effect

4.5 Sizing with Simultaneous Threshold and Supply Voltage Assignment

Increasing the threshold voltage of a transistor reduces the leakage current exponentially, but it has a marginal effect on the dynamic power dissipation. On the other hand, reducing the width of a transistor reduces both leakage and dynamic power, but at a linear rate only. Reference [39] reports on an average 60% and 75% reduction in the total power by using sizing and sizing combined with V_T assignment, respectively. The combination of the technique with dual V_{dd} assignment resulted in only a marginal improvement, probably because of the optimization algorithm used by the authors. Combining the three optimizations is currently an active area of research and will enable synthesizing lower power circuits in the near future.

5. SUMMARY

This article reviewed various sources of leakage current in CMOS integrated circuits and described a number of proven circuit optimization and Computer-Aided Design techniques for controlling the OFF current of CMOS circuits in both standby and active modes of circuit operation.

In way of enumerating some of the design challenges that lie ahead, we mention the following:

- Need robust subthreshold leakage control techniques that do not adversely affect the circuit performance and layout cost. This is especially important in light of both statistical process parameter variations (V_T, L, T_{OX}) and environmental changes (temperature, supply voltage) and their impact on leakage currents
- Develop physical design tools that support multiple voltages on the chip, MTCMOS, and adaptive supply voltage and/or body biasing
- Consider power plane integrity in light of sleep transistor insertions, accounting for both DC voltage drop during the active mode, and ground bounce during the wakeup transition from sleep mode.
- Develop RT-level design flows and tools that allow early evaluation and insertion of power gating structures and other leakage reduction mechanisms.

REFRENCES

[1] L.S. Nielsen, C. Niessen, J. Sparso and C.H. Van Berkel, "Low-Power Operation Using Self-Timed Circuits and Adaptive Scaling of the Supply Voltage", IEEE Trans. on VLSI Systems, pp.391-397, Dec. 1994.

[2] T. Ishihara and H. Yasuura, “Voltage scheduling problem for dynamically variable voltage processors,” *Proc. of Int’l Symp. on Low Power Electronics and Design*, August 1999, pp.197-202.

[3] S. Lee and T. Sakurai, “Run-time power control scheme using software feedback loop for low-power real-time applications,” *Proc. of Asia-Pacific Design Automation Conf.*, January 2000, pp.381-386.

[4] D. Shin, J. Kim, and S. Lee, “Low-energy intra-task voltage scheduling using static timing analysis,” Proc. of Design Automation Conf., June 2001, pp. 438-443.

[5] K. Choi, R. Soma and M. Pedram, "Fine-Grained Dynamic Voltage and Frequency Scaling for Precise Energy and Performance Trade-off based on the Ratio of Off-chip Access to On-chip Computation Times," *Proc. of Design Automation and Test in Europe*, February 2004, Vol. 1, pp. 10004.

[6] R. Gonzalez, B. Gordon, M. Horowitz, “Supply and threshold voltage scaling for low power CMOS,” *IEEE Journal of Solid State Circuits*, Vol. 32, August 1997, pp. 1210-1216.

[7] B. Van Zeghbroeck, *Principles of Semiconductor Devices*, http://ece-www.colorado.edu/~bart/book/book/title.htm, *ch.4.*

[8] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998, ch. 2, pp. 94–95.

[9] K. Cao,W.-C Lee,W. Liu, X. Jin, P. Su, S. Fung, J. An, B. Yu, and C. Hu, “BSIM4 gate leakage model including source drain partition,” in *Tech. Dig. Int. Electron Devices Meeting*, 2000, pp. 815–818.

[10] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S. H. Lo, G. Sai-Halasz, R. Viswanathan, and et al., “CMOS scaling into nanometer regime”, Proc. of the IEEE, vol. 85, Apr. 1997, pp. 486–504.

[11] S. Borkar, “Design Challenges of Technology Scaling”, *IEEE Micro*, Vol. 19, Issue 4, Jul.-Aug. 1999

[12] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, 2003 edition, http://public.itrs.net/.

[13] B. Sheu, D. Scharfetter, P. Ko, and M. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE Journal of Solid State Circuits*, Vol. 22, August 1987, pp. 558-566.

[14] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, " 1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold CMOS," *IEEE J. Solid-State Circuits* 30, No. 8, August 1995, pp. 847–854.

[15] J. T. Kao, A. P. Chandrakasan, ``Dual-threshold voltage techniques for low-power digital circuits," *IEEE Journal of Solid-State Circuits*, Vol. 35, July 2000, pp. 1009-1018.

[16] J. Kao, et al., "MTCMOS Hierarchical Sizing Based on Mutual Exclusive Discharge Patterns", *Proc. Design Automation Conference*, June 1998.

[17] M. Anis, et al. "Dynamic and Leakage Power Reduction in MTCMOS Circuits Using an Automated Efficient Gate Clustering Technique", *Proc. Design Automation Conference*, June 2002.

[18] Hyo-Sig Won, et al., "An MTCMOS Design Methodology and Its Application to Mobile Computing", *Proc. of the international symposium on Low power electronics and design*, August 2003.

[19] J. Kao and A. Chandrakasan, "MTCMOS sequential circuits," *Proc. ESSCIRC*, 2001, pp. 332–339.

[20] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, J. Yamada, "A 1-V high-speed MTCMOS circuit scheme for power-down application circuits," *IEEE Journal of Solid State Circuits*, Vol. 32, June 1997, pp. 861-870.

[21] P. van der Meer et al., "Ultra-low standby-currents for deep sub-micron VLSI CMOS circuits: smart series switch," *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 4, 2000, pp. 1-4.

[22] V. Zyuban and S. Kosonocky, "Low power integrated scan-retention mechanism," *Proc. International Symposium on Low Power Electronics and Design*, August 2002, pp. 98-102.

[23] S. Kim, S. V. Kosonocky, Stephen, and D. R. Knebel, "Understanding and minimizing ground bounce during mode transition of power gating structures", *Proc. of the International Symposium on Low Power Electronics and Design*, August 2003.

[24] K. Seta, H. Hara, T. Kuroda, et al., "50% active-power saving without speed degradation using standby power reduction (SPR) circuit," *IEEE International. Solid-State Circuits Conf.*, February 1995, pp. 318-319.

[25] S-M. Kang and Y. Lelebici, *CMOS Digital Integrated Circuits*, Mc Graw Hill, second edition, 1999.

[26] A. Keshavarzi, S. Narendra, S. Borkar, V. De, and K. Roy, "Technology scaling behavior of optimum reverse body bias for standby leakage power reduction in CMOS IC's," *Proc. International Symposium on Low Power Electronics and Design*, August 1999, pp. 252-254.

[27] L. Clark, N. Deutscher, F. Ricci, and S. Demmons, Standby power management for a 0.18 μm microprocessor, *Proc. International Symposium on Low Power Electronics and Design*,, August 2002, pp. 7-12.

[28] V. De and S. Borkar, "Low power and high performance design challenges in future technologies," *Proc. the 10th Great Lakes Symposium on VLSI*, 2000, pp.1-6.

[29] T. Kuroda, T. Fujita, F. Hatori, and T. Sakurai, "Variable threshold-voltage CMOS technology," *IEICE Transactions. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E83-C, November 2000, pp. 1705-1715.

[30] J. Buurma and L. Cooke "Low-power design using multiple Vth ASIC libraries," http://www.sinavigator.com/Low_Power_Design.pdf. .

[31] K. Roy, et al., "Mixed-Vth (MVT) CMOS Circuit Design Methodology for Low Power Applications", *Proc. Design Automation Conference*, June 1999.

[32] A. Srivastava, "Simultaneous Vt Selection and Assignment for Leakage Optimization", *Proc. International Symposium on Low Power Electronics and Design*, August 2003.

[33] L. Wei, K. Roy, and V. De, "Low voltage low power CMOS design techniques for deep submicron ICs," Proc. of Thirteenth International Conference on VLSI Design, 2000, pp. 24-29.

[34] Y. Taur, et al., "High Performance 0.1um CMOS devices with 1.5V power supply," *Proc. International Electron Devices Meeting*, 1993, pp. 127-130.

[35] L.T. Clark, R. Patel, and T.S. Beaty, "Managing standby and active mode leakage power in deep sub-micron design," *Proc. International Symposium on Low Power Electronics and Design*,, August 2004.

[36] A. Abdollahi, F. Fallah, M. Pedram, "Minimizing leakage current in VLSI circuits," *Technical Report*, Department of Electrical Engineering, University of Southern California, No. 02-08, May 2002.

[37] A. Abdollahi, F. Fallah, M. Pedram, "Runtime mechanisms for leakage current reduction in CMOS VLSI circuits," *Proc. International Symposium on Low Power Electronics and Design*, August 2002.

[38] K. Roy, et al., "Leakage Control with Efficient use of Transistor Stacks in Single Threshold CMOS", *Proc. Design Automation Conference*, June 1999.

[39] K. Keutzer, et al., "Minimization of Dynamic and Static Power Through Joint Assignment of Threshold Voltages and Sizing Optimization", *Proc. International Symposium on Low Power Electronics and Design*, August 1999.