Leakage Reduction Techniques for Nanometer Scale CMOS Circuits

Cassondra Neau and Kaushik Roy
Purdue University Department of Electrical & Computer Engineering
1285 Electrical Engineering Building
West Lafayette, IN 47907
cneau@purdue.edu

ABSTRACT

As we enter the nanoscale regime, power reduction is of increasing importance, but the established leakage reduction techniques will become less effective. Sub-threshold leakage is being joined by band-to-band tunneling and gate leakage as the primary leakage mechanisms. The increased significance of these leakage components threatens the usefulness of some traditional leakage reduction techniques such as stacking and reverse body bias. We present modified versions of these techniques that result in lower total leakage (sub-threshold + gate + band-to-band tunneling) across future technology generations. These modified techniques consider the significance of each leakage mechanism in a given technology generation and make trade-offs accordingly. A novel stacking technique of input vector selection based on the relative contributions of sub-threshold leakage and gate leakage results in up to a 44% reduction in leakage compared to traditional stacking in 50nm devices. An optimal body bias selection technique based on the relative contributions of sub-threshold leakage and band-to-band tunneling results in up to a 42% savings in leakage compared to the zero body bias case for 50nm devices. Since these techniques weigh the relative contributions of leakage components, they reflect the changing requirements as technologies scale.

1. INTRODUCTION

As the physical dimensions of MOS transistors are scaled down, the supply voltage is reduced to maintain device reliability and limit power consumption. This necessitates a commensurate reduction in threshold voltage to preserve the performance gains sought with transistor scaling. This reduction in threshold voltage is responsible for an increase in sub-threshold leakage. Due to this increase in sub-threshold leakage, the power consumed by a microprocessor in the idle state (or leakage power) is rapidly increasing and is a significant portion of the total power consumption of the chip. Therefore, several sub-threshold leakage reduction techniques have been developed to lower the leakage power.

But as devices continue to scale, sub-threshold leakage’s position as the primary idle-state leakage mechanism is beginning to be challenged. As the lateral dimensions of transistors scale, the vertical dimensions, including the gate oxide thickness, must also be scaled to limit the short channel effect. [The short channel effect is the reduction in threshold voltage as the channel length decreases, particularly at high drain biases.] As a result of the thinning of gate oxides, carriers can tunnel through these gate oxides giving rise to significant gate oxide tunneling current. In addition, the doping profiles of sub-0.18um devices include highly doped halo implants and/or retrograde wells. These non-uniform doping profiles control the short channel effect, but result in considerable band-to-band tunneling current between the highly doped p- and n-regions, which are becoming increasingly proximal.

With the increased importance of leakage mechanisms other than sub-threshold leakage, leakage reduction techniques must be reevaluated. In the next section we will describe the three primary leakage mechanisms in nanometer scale CMOS devices: sub-threshold leakage, gate leakage, and band-to-band tunneling. There are other sources of leakage in CMOS devices such as gate-induced drain leakage and punchthrough current, but the three components discussed in section 2 are the most significant in the normal modes of device operation. In the following two sections we will discuss two common sub-threshold leakage reduction techniques and explain how they must be modified in light of the dominant leakage mechanisms of the future. In section 3 we will discuss transistor stacking and in section 4 we will discuss reverse body bias for sub-threshold leakage reduction in the idle mode. Finally, in section 5, we will draw some conclusions and discuss implications for low power design in the future.

2. CMOS LEAKAGE COMPONENTS

2.1. Sub-threshold Leakage

Sub-threshold leakage is the current that flows between the drain and source of a MOS transistor when the gate voltage is below the threshold voltage. In the sub-threshold region, the minority carrier density is exponentially dependent on the gate voltage therefore the sub-threshold leakage is most significant when the gate leakage is just below \( V_{th} \). Since threshold voltages have been rapidly scaled this weak inversion sub-threshold current typically dominates off-state leakage and its importance continues to grow with scaling. Sub-threshold leakage has been extensively modeled and most efforts to reduce leakage focus on this leakage component.
Sub-threshold leakage is modified by drain-induced barrier lowering (DIBL) and the body effect. DIBL is the reduction of threshold voltage as the depletion region of the drain interacts with that of the source, particularly at high drain biases. The body effect is the increase in threshold voltage with reverse body bias due to a widened bulk depletion region. These effects are modeled in [1]:

$$I_{subth} = A \times e^{\eta} \left( \frac{V_{ds}}{V_{th}} \right) \times \left( 1 - e^{-\frac{V_{ds}}{V_{th}}} \right)$$  \hspace{1cm} (1)$$

where

$$A = \mu_0 C_{ox} W \gamma V_{app} \eta$$

$$\Delta V_{th}$$ is the zero bias threshold voltage, and $\nu = KT/q$ is the thermal voltage. The body effect for small values of source to bulk voltages is represented by the term $\gamma' V_{app}$ where $\gamma'$ is the linearized body effect coefficient and $V_{app}$ is the applied reverse body bias. $\eta$ is the DIBL coefficient, $C_{ox}$ is the gate oxide capacitance, $\mu_0$ is the zero bias mobility, and $m$ is the sub-threshold swing coefficient of the transistor. $\Delta V_{th}$ is a term introduced to account for transistor-to-transistor leakage variations.

### 2.2. Gate Leakage

With device scaling the gate oxides are becoming progressively thinner. Thin gate oxides and the resultant high electric fields across the oxides allow direct tunneling (electrons tunneling from the inverted silicon surface directly to the gate through the forbidden energy gap of the SiO$_2$ layer [2]) to occur. The three mechanisms of direct tunneling in MOS devices are: (1) Conduction Band Electron Tunneling (CBET) where electrons tunnel from the conduction band of the substrate to the conduction band of the gate (or vice-versa); (2) Valence Band Electron Tunneling (VBET) where electrons tunnel from the valence band of the substrate to the conduction band of the gate (with generation of free holes in the substrate); and (3) Valence Band Hole Tunneling (VBHT) where holes tunnel from the valence band of the substrate to the conduction band of the gate (or vice-versa) [3] (Figure 1). Direct tunneling current density is modeled as [4]

$$J_{DT} = A \left( \frac{V_{ox}/T_{ox}}{V_{th}} \right)^2 \exp \left( -B \left( \frac{V_{ox}}{T_{ox}} / \phi_h \right)^{0.3} \right)$$  \hspace{1cm} (2)$$

where $J_{DT}$ is the direct tunneling current density, $V_{ox}$ is the potential drop across the oxide, $\phi_h$ is the barrier height for the tunneling particle (electron or hole), and $T_{ox}$ is the oxide thickness. $A$ and $B$ are physical parameters given by [4]:

$$A = \frac{q^3}{16 \pi^2 h \phi_h} \text{ and } B = \frac{4 \sqrt{2} m^* \phi_h^{3/2}}{3 h q}$$

where, $m^*$ is the effective mass of the tunneling particle, $q$ is the electronic charge, and $h$ is the reduced Plank’s constant. Figure 3 shows the variation of tunneling current density with $V_{ds}$ based on (2) [5]. The tunneling current increases exponentially with decreasing oxide thickness and with increasing potential drop across the oxide.

Gate tunneling is generally divided into different categories based on where it occurs in the device [3]. Edge direct tunneling (EDT) components flow in the region where the gate overlaps with the source or drain regions ($I_{gso}$ & $I_{gdo}$). There is also a gate-to-channel tunneling component, part of which flows to the source and part to the drain ($I_{gcd}$ & $I_{gdo}$). Finally there is a gate-to-substrate tunneling component ($I_{gbo}$). EDT is more significant when the device is in accumulation (i.e. for NMOS $V_{th} < V_{gs} \leq 0$, where, $V_{th}$ is the flat-band voltage), but the gate to channel current is dominant for $V_{gs} > 0$ [6].

![Figure 1. Direct Tunneling Mechanisms.](image)

![Figure 2. Variation in Tunneling Current Density](image)
In modern MOS devices, both the p- and n-regions are heavily doped due to strongly doped shallow source/drain regions and halo implants which limit the short channel effect. This gives rise to a significant band-to-band tunneling (BTBT) current that dominates the reverse biased pn-junction current. High electric fields across the reverse biased pn junction result in electrons tunneling from the valence band of the p-side to the conduction band of the n-side as shown in figure 4 [2]. For BTBT to occur, the total voltage drop across the junction must be greater than the bandgap. This tunneling current density is given by [2]:

\[
J_{BTBT} = A \frac{E_{app}}{E_g^{1/2}} \exp \left(-\frac{B E_{app}^{1/2}}{E_g}\right)
\]

(3)

where \(m^*\) is effective mass of electron; \(E_g\) is the energy band-gap; \(V_{app}\) is the applied reverse bias; and \(E\) is the electric field at the junction. Assuming a step junction, the electric field at the junction is given by [2]:

\[
E = \sqrt{\frac{2qN_a N_d (V_{app} + V_{bi})}{\varepsilon_s (N_a + N_d)}}
\]

where \(N_a\) and \(N_d\) are the doping in the p- and n- side, respectively; \(\varepsilon_s\) is the permittivity of silicon; and \(V_{bi}\) is the built in voltage across the junction. In scaled devices, high doping concentrations and abrupt doping profiles cause significant BTBT current through the drain-well junction, particularly if the well is reverse biased.

When both transistors in the NMOS stack are turned off there is a small non-zero drain current flowing through M1 and M2. This leakage current raises the intermediate node voltage, \(V_{bat}\), to a small positive voltage. This positive voltage has the following three effects on transistor M1: the gate to source voltage of M1 (\(V_{gs1}\)) becomes negative; the negative body to source potential of M1 (\(V_{bs1}\)) increases the body effect; and the drain to source potential of M1 (\(V_{ds1}\)) is reduced [1]. As shown in (1), all of these effects decrease the sub-threshold leakage through M1 and hence through the transistor stack.

Transistor stacks are natural by-products of the design process and therefore they fortuitously reduce the sub-threshold leakage of all devices. However the leakage reduction is increased if transistor stacks are intentionally exploited. The first way of doing this is a technique of optimum input vector selection [7]. Due to the stacking effect, the leakage current through a logic gate depends on the applied input vector. By manipulating the primary inputs of a design such that more logic gates have favorable input vectors, the total leakage of a design can be reduced in the “idle” state. This technique reduces leakage by up to 50% by maximizing the number of “off” gates in transistor stacks.

Further leakage reduction can be achieved by a technique called “forced stacking” in which additional series connected transistors are appended to leakage paths where it is only possible to turn off a single transistor to decrease leakage. Although this technique has a small area penalty associated with it, by turning off these additional transistors in the transistor stacks, leakage is reduced by an additional 35% to 90% over input vector selection [7]. MTCMOS [8] which uses high \(V_{bi}\) “sleep” transistors to decouple the logic from the power supply or ground lines during long idle periods can be considered an extension of the concept of forced stacking.

Although transistor stacking reduces sub-threshold leakage, it can result in increased gate leakage as shown in the following example. The minimum sub-threshold leakage condition for the 2input NAND gate in figure 4 occurs when both M1 and M2 have a “0” input. But this input combination results in a large gate current between the drain and the gate of M1. This is primarily due to the large edge direct tunneling current in the gate-drain overlap region (\(I_{gdo}\)). If gate leakage dominates sub-threshold leakage, this could result in an increase in total leakage.

The minimum gate-leakage condition for the 2-input NAND gate occurs when the top gate is turned on and the bottom gate is turned off [5]. Therefore the minimum total leakage condition depends on the ratio of gate leakage to sub-threshold leakage. As shown in figure 5, for larger devices with thicker gate oxides where sub-threshold current dominates the gate leakage, the input vector with all zeros gives the minimum sub-threshold current and therefore the minimum overall leakage. But for devices with thinner oxides, gate leakage can become the dominant

\[\text{Figure 3. Band-to-Band Tunneling.}\]

\[\text{Figure 4. 2-input NAND gate.}\]

3. TRANSISTOR STACKING

When two or more transistors are connected in series, the sub-threshold leakage through this stack of transistors is minimized when the maximum number of MOS transistors is turned “off” in each leakage path [1]. To illustrate this concept, consider the 2 input NAND gate in figure 4.
component of leakage. In this case, turning on the top transistor will give the minimum total leakage condition. Therefore the stacking technique must be modified to prefer turning on the top transistor in transistor stacks if gate leakage is dominant. The leakage savings of this modified technique are shown in table I.

![Figure 5. Standby Power for Various Input Vectors and Transistor Sizes for a 2-input Stack.](image)

**Table I. Leakage Savings for Modified Stacking**

<table>
<thead>
<tr>
<th>Tech.</th>
<th>2T stack savings over ‘01’</th>
<th>2T stack savings over ‘00’</th>
<th>3T stack savings over ‘10’</th>
<th>3T stack savings over ‘00’</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 nm</td>
<td>62%</td>
<td>0%</td>
<td>74%</td>
<td>0%</td>
</tr>
<tr>
<td>50 nm</td>
<td>78%</td>
<td>42%</td>
<td>87%</td>
<td>44%</td>
</tr>
<tr>
<td>25 nm</td>
<td>58%</td>
<td>36%</td>
<td>69%</td>
<td>37%</td>
</tr>
</tbody>
</table>

4. BODY BIAS TO MINIMIZE LEAKAGE

Reverse substrate (or body) bias in the “off”-state is another leakage reduction technique that has been successfully employed to reduce sub-threshold leakage [9]. By selectively applying reverse body bias (RBB) in the “off”-state, the threshold voltage is raised via the body effect, reducing the sub-threshold leakage in the “off”-state without sacrificing performance in the “on”-state.

The scalability of RBB has recently been called into question [10]. The problem with RBB in ultra-small technologies is an increase in the short channel effect. RBB increases drain-induced barrier lowering (DIBL) and with highly doped substrates leads to significant band-to-band tunneling (BTBT) current at the source/drain junctions. These current components can eliminate any power-saving benefits from reverse body bias and even increase leakage in future technology generations. In addition, a fixed RBB leads to an increased sensitivity to process variations. To determine the body bias that lead to minimum leakage it is necessary to consider the trade-off between sub-threshold leakage (1) and band-to-band tunneling leakage (3). As shown in [11] and figure 6 the total leakage is minimized at the body bias for which these two current components are approximately equal. This body bias depends on the device structure and doping profile and changes with process variations. By applying this optimal body bias, it is possible to reduce leakage by over 40% for nominal 44nm and 70nm devices and to reduce the effects of process variation on leakage [11].

5. CONCLUSION

Since there is no longer a single dominant leakage mechanism, designers must consider the effects of leakage reduction techniques on all major leakage components and make tradeoffs accordingly to minimize the total leakage. Since the relative magnitudes of these leakage mechanisms change with each technology generation, leakage reduction techniques must adapt to continue to be effective.

6. REFERENCES


