Effectiveness and Scaling Trends of Leakage Control Techniques for Sub-130nm CMOS Technologies

Bhaskar Chatterjee, Manoj Sachdev, Steven Hsu*, Ram Krishnamurthy*, Shekhar Borkar*
Dept. of Electrical and Computer Engineering
University of Waterloo
Waterloo, Ontario, Canada
bhaskar@vlsi.uwaterloo.ca

Abstract
This paper compares the effectiveness of different leakage control techniques in deep submicron (DSM) bulk CMOS technologies. Simulations show that the 3-5x increase in I_{off}/µm per generation is offsetting the savings in switching energy obtained from technology scaling. We compare both the transistor I_{off} reduction and I_{on} degradation due to each technique for the 130nm-70nm technologies. Our results indicate that the effectiveness of leakage control techniques and the associated energy vs. delay tradeoffs depend on the ratio of switching to leakage energies for a given technology. We use our findings to design a 7nm low power word line driver scheme for a 256 entry, 64-bit register file (RF). As a result, the leakage (total) energy of the word line drivers is reduced by 3x(2.5x) and for the RF by up to 35%(25%) respectively.

Categories and Subject Descriptors
C.4 [Performance of Systems]—modeling techniques, B.7.1 [Integrated Circuits]: Types and Design Styles—advanced technologies, VLSI (very large scale integration)

General Terms
Performance, Design, and Theory

Keywords
DSM leakage control and scaling trends, high performance RF design, non-minimum L_e, RBB.

1. Introduction
The need to achieve improved performance for high-end microprocessors, portable and wireless devices has resulted in aggressive technology scaling over the past two decades. As this trend continues into the future, it is expected that both the device geometry and transistor threshold voltage ($V_{th}$) will be further scaled. This will lead to degraded short channel effects (SCE) and increased transistor OFF-state ($I_{off}$) current. Figure 1 shows the plots for $I_{on}/I_{off}$ ratio, and threshold voltages ($V_{th}$) for low and high $V_{th}$ n-MOS transistors for the 130nm, 100nm and 70nm technologies [1]. The transistor $I_{off}/µm$ is increasing by 3-5x per generation resulting in the degradation of the $I_{on}/I_{off}$ ratio with technology scaling. This will result in excessive leakage currents for the 70nm generation and offset the reduction in switching energy obtained from scaling. In addition, aggravated leakage current is causing thermal hot spots and thermal run away problems during burn-in and adversely affecting long-term reliability of high-end microprocessors [2-4, 6, 18]. Several different leakage control techniques have been proposed in the past [4-11] reducing transistor OFF-state current.

In this paper, we consider four of the most effective techniques:
- Reducing supply voltage ($V_{cc}$) [2, 6, 11]
- Non-minimum channel length ($L_e$) transistors [7]
- Stack effect [10, 14]
- Reverse body bias (RBB) [8, 9]

and compare their effectiveness in reducing leakage current for the 130nm-70nm technologies. In addition, we compare the degradation in $I_{on}$ associated with each of these leakage control techniques. This enables us to choose the most efficient technique offering maximum leakage current savings with minimum degradation in transistor saturation current ($I_{on}$) and circuit performance.

Figure 1: $I_{on}/I_{off}$ and $V_{th}$ scaling for sub-130nm generations

Register files (RF) are performance critical building blocks of high-end microprocessors requiring single cycle read/write latency [12, 13, 21]. In this paper, we design low power word-line (WL) drivers using the above mentioned leakage control techniques for a 256-entry 64-bit high-performance RF. We discuss the energy vs. delay tradeoffs associated with each WL driver implementation. Supply voltage reduction allows switching and leakage energy savings and is most effective in reducing total energy for the 130nm generation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. ISSPDE 03, August 25–27, 2003, Seoul, Korea.
Copyright 2003 ACM 1-58113-682-X/03/0008...$5.00.
RBB and non-minimum \( L_e \) techniques reduce transistor \( I_{\text{OFF}} \) with minimum \( I_{\text{ON}} \) degradation. Consequently, RBB and non-minimum \( L_e \) techniques become more efficient leakage control techniques for high performance digital circuits in the sub-130nm regime.

Rest of the paper is organized as follows: Section 2 discusses various leakage control techniques and models their impact on transistor \( I_{\text{OFF}} \). In Section 3, we present data showing the degradation in transistor \( I_{\text{ON}} \) caused by the leakage control techniques. In Section 4, implementation of the leakage control techniques for a 256-entry 64-bit RF is described. Section 5 discusses some implementation overheads and Section 6 summarizes our conclusions and discusses future work.

## 2. Leakage Control Techniques

In this section, we discuss the leakage control mechanism of the four above-mentioned techniques and present simulation results showing the reductions in \( I_{\text{OFF}} \) for the 130nm generation. The transistor \( I_{\text{OFF}} \) comprises of several different components \([10, 17]\) of which, the weak inversion and drain-induced barrier lowering (DIBL) currents are the most important. These two dominant leakage current components can be modeled using Eq. 1 \([10, 14-16]\) as shown below:

\[
I_{\text{OFF}} = Ae^\left( \frac{V_{GS}-V_{TH}-\eta V_{SB}+\gamma V_{DS}}{n\gamma V_{T}} \right) \left( 1 - e^{-\frac{V_{DS}}{V_T}} \right)
\]

where, \( A = \mu_0 C_{ox} W / L_{eff} \gamma V_T^2 e^{1.8} \), \( \mu_0 \) is the zero bias carrier mobility, \( C_{ox} \) is the gate-oxide capacitance, \( L_{eff} \) is the transistor effective channel length, \( W \) is the transistor width, \( \gamma \) is the DIBL coefficient, \( \eta \) is the linearized body effect coefficient, \( n \) is the transistor sub-threshold swing coefficient and \( V_T \) is the thermal voltage given by \( kT/q \) (33mV at 110°C). In addition, \( V_{TH0}, V_{GS}, V_{SB} \) and \( V_{DS} \) denote the transistor zero-bias threshold voltage, gate-source, source-body, drain-source voltages respectively. We determined the worst-case transistor leakage by simulating the device OFF-state current at 110°C for \( V_{GS} = 0V \) and \( V_{DS} = V_{DD} \) as shown in Figure 2(a-d).

**Figure 2: Leakage control techniques: simulation setup**

In this study, we present data for the n-MOS transistor only. Results for the p-MOS transistor show similar trend and hence are not shown. Based on the leakage current expression in Eq. 1 and simulation setup shown in Figure 2, it is possible to establish closed form approximate expressions modeling \([16, 19]\) the reduction in leakage current for each of the techniques. We use the term \( \Delta I_{\text{OFF}} / I_{\text{OFF}} \) to quantify the leakage current reduction, where:

\[
\frac{\Delta I_{\text{OFF}}}{I_{\text{OFF}}} = \frac{I_{\text{OFF} - final}}{I_{\text{OFF}} - I_{\text{OFF} - final}} = 1 - \frac{I_{\text{OFF} - final}}{I_{\text{OFF}}}
\]

and, \( I_{\text{OFF}} \) and \( I_{\text{OFF} - final} \) represent the transistor OFF-state current without and with leakage control, respectively. In order to simplify the modeling using Eq. 1, we assume \( e^{-\frac{V_{DS}}{V_T}} \approx 0 \). This approximation is justified since in this study, the ratio of \( V_{DS} / V_T \geq 20 \). Thus, the simplified transistor OFF-state current is given by:

\[
I_{\text{OFF}} = Ae^\left( \frac{V_{GS}-V_{TH}-\eta V_{SB}+\gamma V_{DS}}{n\gamma V_{T}} \right)
\]

Supply voltage reduction lowers the transistor drain-source voltage consequently reducing the DIBL current (Figure 2(a)). In addition, since \( V_{SB} = 0V \), the corresponding term in Eq. 3 is equal to zero. Thus, the reduction in leakage current obtained by lowering of the supply voltage is given by:

\[
\frac{\Delta V_{CC}}{I_{\text{OFF}}} = 1 - e^{-\frac{\eta V_{DS}}{V_T}}
\]

Non-minimum channel length transistors (Figure 2(b)) reduce the \( I_{\text{OFF}} \) \([7]\) by increasing the zero-bias threshold voltage (\( V_{TH0} \)). The impact of transistor channel length on \( V_{TH0} \) is shown in Figure 3.

**Figure 3: \( V_{TH} \) vs. channel length for 130nm n-MOS transistor**

In the region of interest (Figure 3), the threshold voltage increases almost linearly for small increases in the drawn channel length (\( L_e \)). As a result, the increase in the transistor zero bias threshold voltage
can be approximated using \( \Delta V_{THI} = V_{THI} \left( \frac{L_c}{L_{\text{eff}}} \right) \). In addition, the channel mobility remains approximately constant due to velocity saturation in DSM transistors. Therefore, the reduction in leakage current using non-minimum channel length transistors can be modeled as:

\[
\frac{\Delta I_{OFF}}{I_{OFF}} \approx 1 - \frac{1}{1 + \frac{\Delta L_{\text{eff}}}{L_{\text{eff}}}} e^{-\frac{\Delta V_{TH}}{nV_T}} \tag{5}
\]

where, \( \Delta L_{\text{eff}} \) is the change in effective channel length (\( L_{\text{eff}} \)) while all other terms have their usual meanings.

Similarly, the reduction in leakage current using stack effect can be explained with the help of Figure 2(c). The intermediate node voltage (\( V_N \)) reaches a steady state DC value for the two stack n-MOS pull down when both transistors are OFF. This value is within an order of magnitude of the thermal voltage (\( V_T \)) and can be calculated using the generalized equations derived in [14]. Simulations in 130nm CMOS technology for our specific circuit indicate that \( V_T \) reaches a steady state value of 95mV. This results from the \( V_{GS} \) voltage drop across the bottom transistor (N2). As a result, a negative \( V_{GS} \) gate drive voltage appears across the top n-MOS transistor (N1) of the stack. Furthermore, there is a reduction in \( V_{DS} \) (DIBL current suppression) and appearance of negative \( V_{BS} \) (RBB) across the top transistor (N1) because of stack effect. Thus, the leakage current reduction using stack effect is:

\[
\frac{\Delta I_{OFF}}{I_{OFF}} \approx 1 - e^{-\frac{\Delta V_{GS}}{nV_T}} \tag{6}
\]

Finally, the leakage current can be suppressed by reverse body biasing (RBB) the transistor (Figure 2(d)). The body of the n-MOS transistor is connected to a negative voltage with respect to the source terminal. The reduction in leakage current is proportional to the extent of the applied reverse bias voltage (\( V_{SB} \)). However, recent research indicates [8, 9] that beyond a certain optimal RBB voltage the transistor OFF-state current starts to increase due to increased gate induced drain lowering (GIDL) leakage as shown in Figure 4.

For the range of RBB voltages in the region of interest, Eq. (3) can be used to model the leakage current reduction as follows:

\[
\frac{\Delta I_{OFF}}{I_{OFF}} \approx 1 - e^{-\frac{\Delta V_{GS}}{nV_T}} \tag{7}
\]

The reduction in n-MOS transistor worst-case \( I_{OFF} \) was modeled for each leakage control technique using Eqs. (4-7). Table 1 demonstrates, that the theoretical models track the simulation results for all the four above-mentioned techniques. However, the models consistently under estimate the reductions in leakage current. This is because the simplified equations do not account for all the leakage current components and associated reductions when using leakage control techniques. These results indicate that stack effect reduces leakage current by up to 12x while using non-minimum \( L_c \) (\( L_c \) increased 30%) reduces leakage by 9.3x. On the other hand, supply voltage reduction (\( V_{CC} \) reduced by 30%) or RBB equal to 30% of \( V_{CC} \) reduces leakage by 2.2x-2.3x.

**Table 1: Leakage current reductions for 130nm technology**

<table>
<thead>
<tr>
<th>Technique</th>
<th>Simulation Results</th>
<th>Theoretical Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply Voltage (30% ( V_{CC} ) reduction)</td>
<td>2.2x</td>
<td>1.9x</td>
</tr>
<tr>
<td>Non minimum ( L_c ) (( L_c ) +30%)</td>
<td>9.3x</td>
<td>8.7x</td>
</tr>
<tr>
<td>Stack Effect</td>
<td>12.0x</td>
<td>11.5x</td>
</tr>
<tr>
<td>RBB (30% reverse bias)</td>
<td>2.3x</td>
<td>2.1x</td>
</tr>
</tbody>
</table>

### 3. Performance Impact and Scaling Trends

An efficient leakage control technique is one that allows large reduction in \( I_{OFF} \) with minimum \( I_{ON} \) degradation. This helps minimize the adverse performance impact of the leakage control techniques when used in high-end digital circuits. We quantify the \( I_{OFF} \) vs. \( I_{ON} \) tradeoffs for each technique in Figures 5 (a) and 5(b) for the 130nm and 70nm generations.
The data for the 100nm generation shows similar trend, and is therefore not presented. These figures indicate that both RBB and non-minimum \( L_e \) techniques result in lesser degradation of transistor \( I_{ON} \) than supply \( V_{CC} \) reduction or stack effect. Consequently, both RBB and non-minimum \( L_e \) techniques have steeper gradients in the \( I_{OFF}-I_{ON} \) plane making them more efficient in reducing leakage current for the 130nm-70nm generations. The extent to which \( I_{OFF} \) can be reduced using RBB is however limited by the optimal reverse bias voltage as discussed in Section 2.

In addition to examining the performance trade-offs, we study the scalability of these techniques for the sub-130nm generations. For this purpose, we compare the degradation of the normalized \( I_{OFF}/I_{ON} \) ratio as shown below:

\[
\xi = \left( \frac{\partial I_{OFF}}{\partial I_{ON}} \right) \left( \frac{I_{OFF}}{I_{ON}} \right)
\]

Figure 5(b): \( I_{OFF} \) vs. \( I_{ON} \) plots for 70nm technology

Table 2 shows the value of \( \xi \) and scaling trends for each of the leakage control techniques. It should be noted that a technique with higher value of \( \xi \) is more efficient in reducing leakage current.

Table 2. Normalized \( I_{OFF}/I_{ON} \) degradation: scaling trends

<table>
<thead>
<tr>
<th>( \xi )</th>
<th>130nm</th>
<th>100nm</th>
<th>70nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply Voltage (30% ( V_{CC} ) reduction)</td>
<td>1.1</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>Non minimum ( L_e ) (( L_e+30% ))</td>
<td>3.1</td>
<td>3.1</td>
<td>2.8</td>
</tr>
<tr>
<td>Stack Effect (30% reverse bias)</td>
<td>2.2</td>
<td>2.1</td>
<td>1.9</td>
</tr>
<tr>
<td>RBB</td>
<td>20.0</td>
<td>9.0</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Our results indicate that RBB followed by non-minimum \( L_e \) have the highest \( \xi \) values. On the other hand, the degradation in \( I_{ON} \) resulting from techniques like supply voltage reduction and stack effect (2 series transistors) make their respective \( \xi \) values lower. In fact, the reduction in \( I_{ON} \) more than offsets the savings in leakage current when supply voltage is reduced by 30% for the 70nm generation.

4. Low Power Word Line Driver for RFs

Wide bit-width register files (RF) are performance-critical components of microprocessor integer/FPU execution cores and demand single cycle read/write latency. In this section, we use the leakage control techniques discussed in the previous sections to implement low power word line (WL) drivers for a 2-read, 1-write ported 256-entry 64-bit high performance RF whose organization is shown in Figure 6. Each of the 256 RF entries is uniquely selected using an 8:256 decoder scheme that generates the read and write select (RS/WS) signals. In order to read/write from an entry, only one WL driver signal/port switches high (active) while the rest of the 255 drivers are inactive and are leaking. The read port WL drivers are on the RF critical path and are hence upsized to drive the 32 local bit-cells on each side of this partitioned RF. This results in increased leakage for the WL drivers of the 255 deselected entries.

Figure 7 shows the contribution of the WL drivers to the RF total energy as a function of technology scaling. In addition, we show the breakup of the WL driver total energy in terms of switching and leakage energy. The normalized WL driver energy contribution increases from 18% (130nm) to 27% (70nm) of RF total energy with scaling. The WL driver switching energy reduces by about 50% per generation due to capacitance and voltage scaling as indicated in Figure 7.
The WL driver leakage energy however, increases from 22% (130nm) to 83% (70nm) of the total energy due to 3-5x increase in $I_{off}/\mu\text{m}$ per generation. This provides the basis for our investigation into efficient leakage control techniques for the RF word line drivers. Figure 8 shows the layout of the WL drivers and the local bit-line (LBL) interconnects of the 256 entry 64 bit RF for the 130nm generation.

![Figure 8: WL driver layout for 256-entry, 64-bit 130nm RF](image)

Figures 9(a) and 9(b) show the WL driver total energy (switching + leakage) vs. delay tradeoffs for each of the leakage control techniques for the 130nm and 70nm generations respectively.

![Figure 9(a): Total energy vs. delay for 130nm WL drivers](image)

![Figure 9(b): Total energy vs. delay for 70nm WL drivers](image)

Switching energy dominates the WL driver total energy (switching energy $\propto V_{cc}^2$) for the 130nm generation. Thus, a 30% supply voltage reduction allows a 2.4x reduction in total energy and a 3.6x reduction in leakage energy. The reduction in leakage energy using RBB is however limited by the optimal reverse body bias voltage discussed earlier. In addition, RBB does not have significant impact on switching energy. Consequently, 30% RBB allows a 12% reduction in total energy while reducing leakage energy by 1.9x for the 130nm technology (Figure 9(a)).

However, as the technology is scaled to 70nm, leakage energy becomes 83% of WL driver total energy and RBB becomes more effective than supply scaling as indicated in Figure 9(b). Thus, a 30% RBB allows 3x(2.5x) reduction in leakage (total) energy compared to the baseline design (design with no leakage control). We compare the RBB and $V_{cc}$ scaling curves at 2 points (A and B) in the total energy vs. delay plane. Point A corresponds to the 30% RBB case and allows 37% savings in total energy compared to $V_{cc}$ scaling at the same delay. On the other hand, point B corresponds to the case with 30% $V_{cc}$ scaling and indicates that RBB allows 20% improved performance for the same total energy. It should be noted that the energy vs. delay tradeoffs for the different techniques is dependent on the switching and leakage energy breakup of the total energy and the $V_{cc}/V_{TH}$ ratio for a given technology.

Figure 10 shows the leakage energy, total energy and delay tradeoffs for the 70nm, 2 read-1 write ported 256 entry 64-bit RF using the different low power word line drivers. The leakage (total) energy reduction for the RF is between 30-35% (13-25%) while the worst-case RF read delay increases by 4-21%. Our results indicate that RBB followed by usage of non-minimum $L_e$ transistors offer better energy savings with less delay penalties for the 70nm RF design and are more efficient than techniques involving reduced supply voltage or stack effect.

![Figure 10: Normalized leakage, total energy, delay: 70nm RF](image)

5. Implementation Issues

In addition to performance degradation, the above-mentioned leakage control techniques have certain implementation overheads. For example, the selective usage of supply voltage reduction for the word line drivers might require a dual-$V_{cc}$ design to mitigate excessive performance penalty [20]. This requires the generation and routing of a second power supply within the RF. Similarly, the usage of RBB requires the generation and routing of extra power supply to the body and well terminals of n- and p-MOS transistors. In addition, it will require the usage of a triple well bulk CMOS process [5] increasing the overall implementation cost. The stack effect and non-minimum $L_e$ techniques do not require any additional power supply. However, stack effect reduces the effective transistor $I_{on}$ and results in excessive performance degradation. Furthermore, the additional stack transistor has to be upsized thereby increasing the overall switching energy and area penalty. On the other hand,
usage of non-minimum $L_e$ would require precise process control over the transistor channel length.

6. Conclusions and Future Work
In this paper, we have compared four different leakage control techniques and discussed their effectiveness in reducing leakage and total energy for the 130nm-70nm technologies. We demonstrated that RBB followed by non-minimum $L_e$ are the most efficient leakage control techniques in the sub-130nm regime. However, the effectiveness of each of these techniques in reducing leakage and total energy depends on the ratio of switching/leakage energy and $V_{CC}/V_{TH}$ for a given technology. Based on the above results we designed a low power word line driver for a 2 read, 1 write ported, 256 entry 64 bit RF which achieved 3x(2.5x) reduction in leakage (total) energy for the 70nm generation. In this paper, we modeled the impact of leakage control techniques on the sub-threshold current. However, it is expected that the gate leakage will become a significant component of the total transistor leakage current in the sub-100nm regime. Both stack effect and non-minimum $L_e$ techniques result in the increase of the overall transistor active area. In addition, RBB results in an increase of the effective gate-body voltage. Thus, a more detailed model needs to account for the impact of leakage control techniques on gate leakage for DSM technologies. Studying the reductions in total leakage energy afforded by various leakage control techniques in the context of high performance digital logic blocks remains the topic of our future research.

7. Acknowledgements
Authors would like to acknowledge Oleg Semenov and Christine Kwong (University of Waterloo), Atila Alvandpour, Srimat Vangal, Vivek De (Intel Corp.) for discussions, and Matt Maycock and Justin Rattner for encouragement and support.

8. References