Design Limitations in Deep Sub-0.1µm CMOS SRAM

Robert K. Grube, Qi Wang and Sung-Mo Kang
Low-Power Circuit Design Group, Jack Baskin School of Engineering
University of California, Santa Cruz
rkgrube@soe.ucsc.edu, wangqi@soe.ucsc.edu, kang@soe.ucsc.edu

ABSTRACT

Leakage effects in deep sub-0.1µm CMOS technologies are of critical concern to designers of high-performance integrated circuits. Recent estimates [1] of a 7.5x increase in leakage current per chip generation; along with several proposals for energy-efficient cache architectures that unfortunately do not address static leakage-energy issues [2], [3], [4], have heightened concerns over the functionality and stability of future high-performance SRAM cache designs. Each of these future technology generations, in addition to having increased short-channel effects (SCEs) will now also suffer from gate leakage currents across physically and electrically thin (~1.5nm) SiO_2 gate dielectrics. In this paper we demonstrate the limits of this scaling on the operational behavior of on-chip SRAM cache designs, and briefly discuss the impact of these results on high-performance memory architectures.

Categories and Subject Descriptors

Hardware – Integrated Circuits – Types and Design Styles (B.7.1): **Advanced Technologies**; Hardware – Memory Structures – Semiconductor Memories (B.3.1): **Static Memory (SRAM)**

General Terms

Performance, Design

Keywords

Gate leakage, tunneling currents, GIDL, On-Chip Cache

1. INTRODUCTION

Historically, technology scaling trends seek to improve gate delay by about 30% and the reduction of transition-energy by approximately 30% ~ 65% per generation, typically by scaling supply voltages and/or shrinking the process technology. Maximum supply voltages are limited by gate oxide wear-out; whereas minimum supply voltage levels are typically set by practical noise-margin and performance considerations. Shrinks in process technology, on the other hand, must maintain proper device behavior at smaller and smaller channel lengths and progressively thinner gate dielectrics -- which is in turn dependent on maintaining an adequately large lateral-to-vertical aspect ratio for a device [5] [6]:

GLSVLSI'02, APRIL 18-19, 2002, NEW YORK, NEW YORK, USA.

$$\frac{L_{eff}}{\left(\frac{t_{ox}\cdot\varepsilon_{si}}{\varepsilon_{ox}}\right)^{1/3}(d)^{1/3}(d_j)^{1/3}}$$

In this equation, L_{eff} is the effective channel length; t_{ox} the gate oxide thickness; d the channel depletion depth; d_j the effective junction depth; and ε_{si} and ε_{ox} are the permittivities of silicon and oxide. Thus, the ability to scale SiO_2 gate dielectrics is limited by both the scalability of the supply voltage and the desire to preserve the device's aspect ratio.

SRAM circuits in deep sub-0.1µm CMOS technologies exhibit profound *Read* sensitivities to increased leakage current. Due to the limited scalability in supply voltages in high-performance applications, high electric fields may develop across the thin (~1.5nm) SiO_2 gate oxide. This field distorts the silicon bandgap, such that electrons may more easily travel from the valence to the conduction band, from the gate to the channel and body. The following models for direct tunneling current may be used to better understand this leakage component [7]:

$$J_{DT} = A\left(\frac{V_{ox}}{t_{ox}}\right)^2 e^{\frac{-B\left(1-\left(1-\frac{V_{ox}}{\varphi_B}\right)^{3/2}\right)}{\frac{V_{ox}}{t_{ox}}}}$$

Here, A and B are physical parameters, t_{ox} and V_{ox} are the gate oxide's thickness and voltage potential. The inherent dependence of V_{ox} on the surface potential (ψ_s) can be approximated in the weak- and strong-inversion regimes by the following equations:

$$\psi_{s,weak} = V_{gb} - V_{FB} + \frac{\gamma^2}{2} - \gamma\sqrt{V_{gb} - V_{FB} + \frac{\gamma^2}{4}}$$

$$\psi_{s,strong} = 2\phi_f + V$$

In the equations above, γ is the body factor; and V denotes the electron quasi-Fermi potential, ranging from V_{sb} at the source to V_{db} at the drain.

This tunneling current along with sub-threshold leakage mechanisms, combine to affect the buildup of a voltage differential between the SRAM's bit lines such that the current-sinking behavior of the selected SRAM cell's wordline-NFETs must contend with significant leakage current from the non-selected devices. Due to the inability of the sense-amplifier's offset voltage to scale at a similar 30% rate per technology generation, the

voltage swing on the bit line cannot be scaled at the same rate as the supply voltage. Subsequently, the number of rows per bit line in high-performance SRAM caches has been historically reduced by 2X per technology generation to alleviate the leakage effects from the non-selected devices.

In the following sections we examine speculative 0.06µm and 0.13µm CMOS process technologies for SRAM circuits that utilize differential sensing; and examine the impact of various gate dielectric thicknesses, channel lengths, and doping profiles on the reliability and performance of the SRAM's *Read* operation. Low-V_{th} transistors are used in our SRAM array since they provide faster current conducting capabilities, and also because their usage reflects a general trend within the semiconductor industry to reduce memory-access times in high-performance, on-chip SRAM caches [8]. Differential sensing is implemented rather than single-ended sensing [9], since the degradation in noise margin for single-ended sensing was found to be unworkable at the supply voltage levels one would typically use at the 0.06µm technology generation.

2. SRAM & SENSE-AMP CIRCUITS

6-T SRAM and differential sense-amplifier circuits with devices having an L_{drawn} of 130nm or 60nm and a maximum gate dielectric thickness of 2.5nm were used for our analysis. The W/L ratio of each transistor was designed to guarantee maximum stability of the internal inverter latch structure, which allowed us to perform a more realistic analysis on our speculative processes. Our analysis concerning design trade-offs and scaling behavior, then, is done from the perspective of a well-designed current-sink to compensate as much as possible for parasitic leakages.

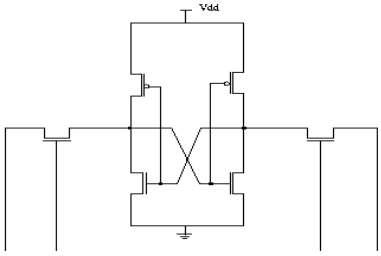

Figure 1. 6-T CMOS SRAM Cell

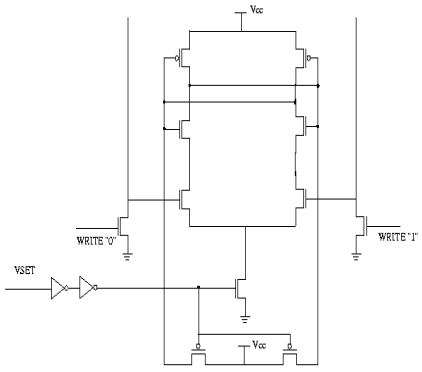

Figure 2. Differential Sense-Amplifier

3. ANALYSIS AND RESULTS

Worst-case data analysis is used to maximize device leakage and its effect on the buildup of a differential voltage across the bitlines for the circuit's *Read* '0' operation. When a data-state of '0' is being read from a particular SRAM cell, worst-case leakage occurs when all of the other bits on that bitline have a data-state of '1'. So, the '0'-bit's wordline-NFETs will be enabled, and that particular SRAM cell will serve as a current-sink for the bitline. All other wordline-NFETs will be disabled, creating resistive current sources. Figure 3 shows this worst-case scenario.

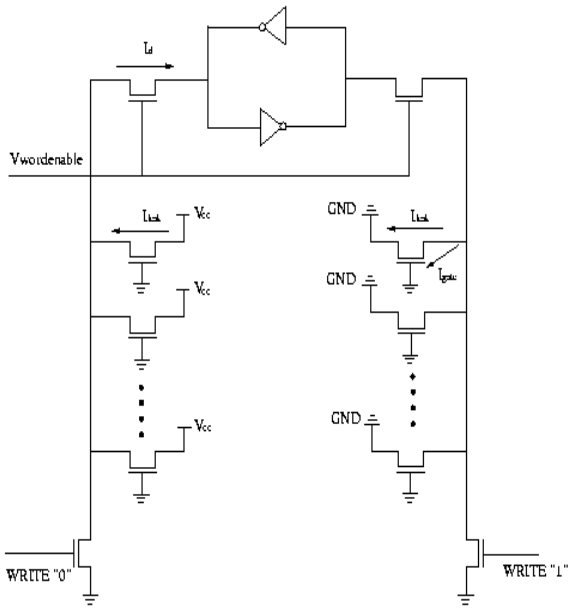

Figure 3. Circuit configuration for worst-case analysis

Since the disabled devices have a data state of '1' on their sources and their gates are grounded, a large static electrical field exists across their gate-to-source overlap areas which enhances direct tunneling – resulting in a large number of enhanced current sources that contend with the enabled current sink.

On the opposite bitline, the enabled device 'sees' V_{cc} or a data value of '1', and the rest of the unselected cells store a data value of '0'. So during the *Read* operation the unselected SRAM cells on this side sink current from the bitline through significant sub-threshold- and gate-leakage currents, which are exacerbated by the existence of a full V_{ds} across these transistors. This combination of leakage effects on both bitlines disrupts the ability of the sense-amplifier to sense a proper voltage differential in some of our simulations.

Separate analyses based on speculative 0.13µm and 0.06µm technologies were performed to quantify leakage trends, and to understand design limitations in SRAM circuits as the transistor's dimensions scale. Berkeley Spice and HSPICE simulation tools were used to conduct our analysis; and accurate device parameters were extracted from ultra-thin gate oxides that were grown using

rapid thermal oxidation (RTO) in work published in [10]. All simulations were performed with 32 SRAM cells per bitline, as we wish to analyze the scalability of modern, high-performance, on-chip SRAM micro-architectures. The results of our studies are summarized in the following graphs below. These simulations indicate the relative impact of the leakage current on the circuit's performance as the oxide shrinks; and indicate the limits of the *Read* operation on 32-bit SRAM circuits within the set of device parameters listed below in Table 1.

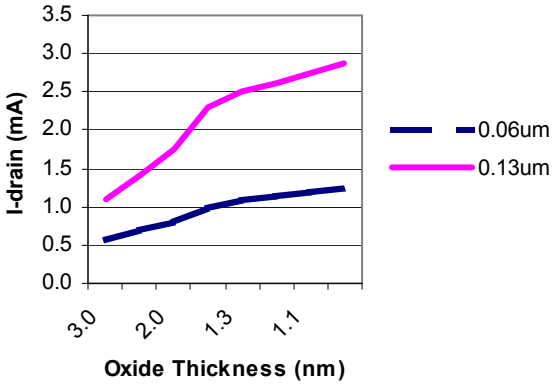

Figure 4. I_{drain}, 0.06µm and 0.13µm processes

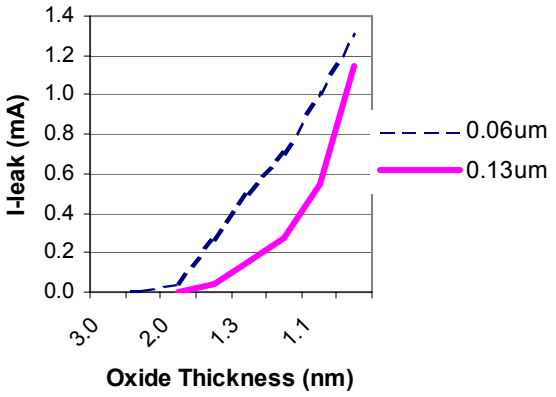

Figure 5. I_{leak}, 0.06µm and 0.13µm processes

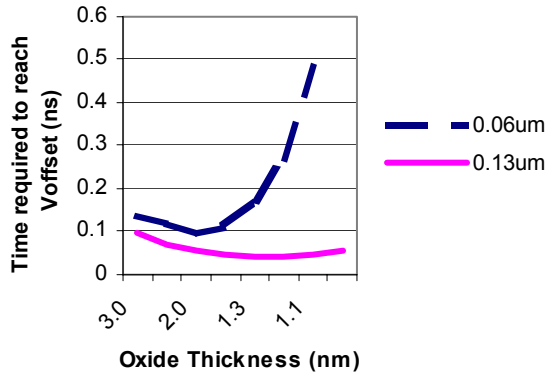

Figure 6. V_{offset}, 0.06µm and 0.13µm processes

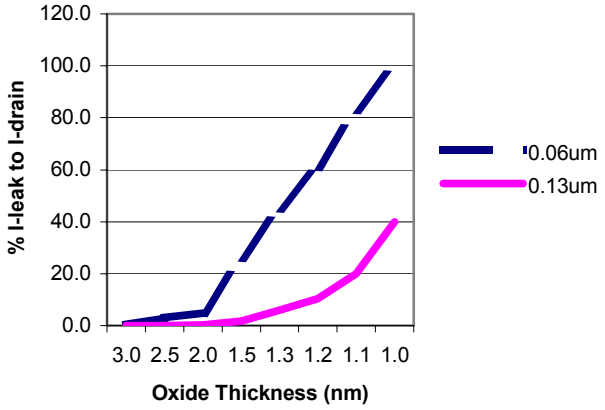

Figure 7. % I_{leak} to I_{drain}, 0.06µm and 0.13µm processes

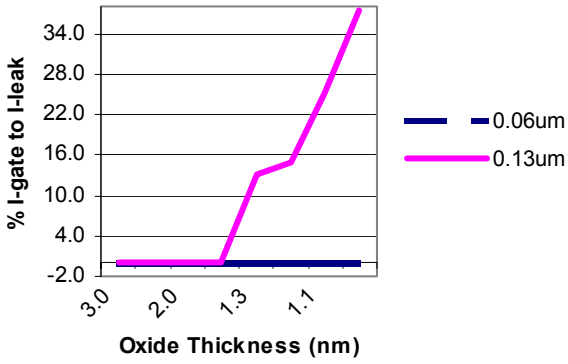

Figure 8. % I_{gate} to I_{leak}, 0.06µm and 0.13µm processes

Table 1. Key simulation parameters

Process	DLCIG	X_j	V_{offset}	V_{cc}
0.06µm	0.015µm	70nm	50mV	1.0V
0.13µm	0.03µm	70nm	50mV	1.4V

As the graphs indicate above, leakage parasitics become extreme for oxide thicknesses in the range of 1.2 – 1.1nm. Figure 6 indicates an initial improvement in performance as I_{drain} increases, but this trend quickly deteriorates due to increased leakage current, I_{leak}. And in Figure 8 we see that I_{gate} only becomes a substantial percentage of I_{leak} for very thin oxide thicknesses, and then only for the 0.13µm process and not for the 0.06µm process. This is largely due to I_{gate}'s heavy dependence on the transistor's channel length. In our 0.06µm process, a V_{offset} of 50mV could not be created between the bitlines of the sense-amplifier at gate oxide thicknesses of 1.2nm and below. This is directly due to short-channel effects and tunneling parasitics that overwhelm the current-sinking capabilities of the enabled SRAM cell.

In our simulations the device parameters in Table 1 gave us our best performance, and are considered to be reasonable for high-performance, on-chip caches. Varying device parameters in our

0.06µm process in search of improved circuit performance, we observed that the leakage currents are most sensitive to the substrate doping profile, in addition to gate oxide thickness, supply voltage, and channel length. In the analysis above, both our 0.06µm and 0.13µm processes used a substrate dopant concentration of $6 \cdot 10^{16}/cm^{-3}$.

Varying this dopant concentration for our 0.06µm process, then, we analyzed the impact of dopant variations on 32- and 64-bit SRAM circuits comprised of gate oxide thicknesses between 1.1 and 2.5nm. Figure 9 below shows the impact of these variations, where the Pass/Fail criterion is simply the ability to develop a 50mV V_{offset} at the sense-amplifier.

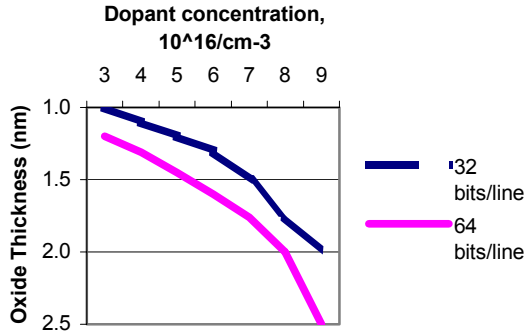

Figure 9. Dopant concentration vs. oxide thickness

The lines in the figure above represent the Pass/Fail boundaries for bitlines with 32- and 64-bit SRAM cells. The "Pass" criteria for the data sets exist below each line; and the combinations of dopant concentrations and oxide thicknesses at each point above the lines represent the cases where the sense-amplifiers are unable to develop a V_{offset} of 50mV or more.

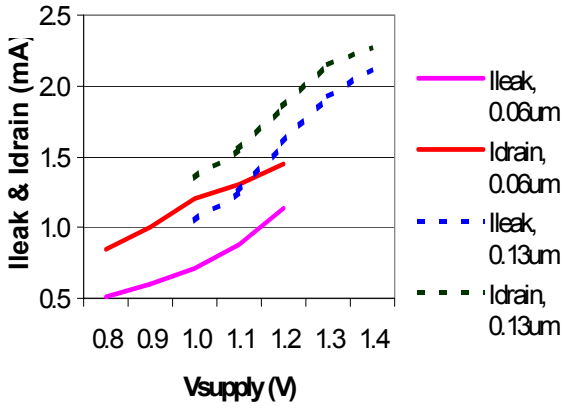

Figure 10. I_{leak} & I_{drain} vs. supply voltage

Analyzing the relative impact of the supply voltage on both I_{leak} and I_{drain}, we varied V_{supply} from 1.2V to 0.8V; and from 1.4V to 1.0V for our 0.06µm and 0.13µm processes, respectively. Though total power consumption may impose a scaling limit on the use of ultra-thin oxides, it is nonetheless application dependent. Therefore, we analyzed these particular voltage ranges so that our data would be fairly representative of a variety of applications at these process generations. Figure 10 below shows our simulation results.
These results were obtained for a substrate dopant concentration of $6 \cdot 10^{16}/cm^{-3}$, and T_{oxe} parameters at the extreme end of our Pass/Fail criterion of 50mV V_{offset} for each speculative process. Here, the I_{leak}/I_{drain} ratio remains high across the supply voltage range for each process; and Fig. 10 indicates the relative limitations on oxide scaling in terms of power consumption and acceptable circuit performance, even with the advantage of relaxed supply voltage constraints.

4. CONCLUSIONS

We have performed an analysis of leakage sensitivities on the *Read* operation of typical CMOS SRAM circuits for high-performance, on-chip caches in speculative 0.06µm and 0.13µm processes. By carefully analyzing the worst-case behavior of 32- and 64-bit SRAM arrays, we have identified functional failures at several design corners in a speculative 0.06µm process with gate oxide thicknesses of 1.2nm and lower. Due to the severe leakage parasitics present in this technology generation, it is likely that 16-bits or even fewer SRAM cells per bitline will be required to alleviate leakage problems in future high-performance, on-chip caches. Since the amount of on-chip cache is expected to increase dramatically over the next decade for most high-performance applications [11], this requirement will have a pronounced impact on SRAM cache architectures in the near future unless novel dielectrics such as those found in the literature [12] [13] can be successfully incorporated into low-cost manufacturing processes.

5. ACKNOWLEDGMENTS

This research has been supported in part by the Semiconductor Research Corporation (Contract No. 2001-HJ-891). We would like to thank our colleagues at Synopsys Corporation and Intel Corporation for their timely advice and insight into the issues discussed in this paper: Vaishali Khedekar, B.W. Khedekar, Changhong Dai, Gloria Leong, and Stefan Rusu.

6. REFERENCES

[1] S. Borkar, *IEEE Micro*, **19**, pp. 23 – 29, July 1999.
[2] D.H. Albonesi, Proc. 32nd IEEE/ACM Int. Symp. Micro., 1999.
[3] N. Bellas, *et al.*, Proc. ISLPED '99, pp. 64 – 69, Aug. 1999.
[4] J. Kin, *et al.*, Proc. 30th IEEE/ACM Int. Symp. Micro., 1997.
[5] V. De, S. Borkar, Proc. ISLPED'99, pp. 163 – 168.
[6] C. Wann *et al.*, IEEE Trans. Electron Dev., Oct. 1996.
[7] C.-H. Choi, *et al.*, "Modeling of MOS Scaling with Emphasis on Gate Tunneling and Source/Drain Resistance", CIS, Stanford University, Stanford, CA.
[8] F. Hamzaoglu, *et al.*, ISLPED '00, pp. 15-19.
[9] H. Tran, 1996 Symp. VLSI Ckt., pp. 68-69.
[10] N. Yang, *et al.*, IEEE Trans. On Electron Dev., Aug. 2000.
[11] M. Powell, *et al.*, IEEE Trans. VLSI, pp. 77 – 89, Feb. 2001.
[12] J. P. Chang, *et al.*, J. Vac. Sci. Technol. B 19, 5 (2001).
[13] J. P. Chang and Y. -S. Lin, Appl. Phys. Lett. 23, 3824 (2001).