AFRL-IF-WP-TR-2005-1586

LEAKAGE REDUCTION FOR ON-DIE CACHES

Kaushik Roy

Purdue University
Dept. of Electrical and Computer Engineering
465 Northwestern Avenue
West Lafayette, IN 47907

OCTOBER 2005

Final Report for 01 July 2002 – 31 December 2004

Approved for public release; distribution is unlimited.

STINFO FINAL REPORT

INFORMATION DIRECTORATE
AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7334

NOTICE

Using government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them

This report was cleared for public release by the Air Force Research Laboratory Wright Site (AFRL/WS) Public Affairs Office (PAO) and is releasable to the National Technical Information service (NTIS). It will be available to the general public, including foreign nationals.

PAO Case Number: AFRL/WS 05-2471, 08 November 2005.

THIS TECHNICAL REPORT HAS BEEN APPROVED FOR PUBLICATION

//S//

RONALD W. BROWER, Ph.D
Project Engineer
Embedded Information Systems Branch
Advanced Computing Division

//S//

AL SCARPELLI
Team Leader
Embedded Information Systems Branch
Advanced Computing Division

//S//

JAMES S. WILLIAMSON, Chief
Embedded Information Systems Branch
Advanced Computing Division
Information Directorate

This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
October 2005	Final	07/01/2002 – 12/31/2004

4. TITLE AND SUBTITLE
LEAKAGE REDUCTION FOR ON-DIE CACHES

5a. CONTRACT NUMBER
F33615-02-1-4003

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62301E

6. AUTHOR(S)
Kaushik Roy

5d. PROJECT NUMBER
M765

5e. TASK NUMBER
40

5f. WORK UNIT NUMBER
03

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Purdue University
Dept. of Electrical and Computer Engineering
465 Northwestern Avenue
West Lafayette, IN 47907

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Information Directorate
Air Force Research Laboratory
Air Force Materiel Command
Wright-Patterson AFB, OH 45433-7334

Defense Advanced Research Projects Agency (DARPA)
3701 N. Fairfax Drive
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL-IF-WP

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-IF-WP-TR-2005-1586

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Report contains color.

14. ABSTRACT

Technology scaling is associated with exponential increase in leakage for every subsequent technology generation. Since cache memories take up a significant portion of the die in modern processors, leakage contributes largely to power dissipation in caches. Our effort was to reduce leakage in large caches. We investigated several leakage-tolerant co-design techniques at the circuit and architecture levels. We also considered the effectiveness of the proposed techniques with predictive scaled devices. In particular, we determined that source biased and body biased caches can be effective in reducing leakage significantly. A 0.18 μm, 1.8 V, 16 KB source-biased static random access memory (SRAM) test chip shows 94.2 percent reduction in SRAM cell leakage at a performance penalty of less than 2 percent. Measured results also indicate that our proposed memory cell improves SRAM static noise margin by 25 percent. It should be noted that our techniques do consider different components of leakage current and process parameter variations.

15. SUBJECT TERMS
Leakage, power-aware computing, cache, memory, SRAM

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ronald Brower
Unclassified	Unclassified	Unclassified	SAR	38	19b. TELEPHONE NUMBER *(Include Area Code)* (937) 255-6548 x3590

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

Table of Contents

Section Page

List of Figures

List of Tables

ACKNOWLEDGEMENTS

This research has been funded in part by the DARPA PAC/C program and the Semiconductor Research Corporation. The author also wants to thank A. Keshavarzi and S. Narendra for technical discussion; M. Johnson for the CAD tool support; B. Graybill for the assistance with the solid-state lab equipments; and the Intel Ph.D. fellowship program.

1. INTRODUCTION

Scaling of CMOS technology has enabled a phenomenal growth in computing capability throughout the last four decades. With the number of transistors on a chip rapidly approaching 1 billion and the integrated cache memory dominating the chip area, leakage power management has become indispensable in high-end microprocessors for cost-effective packaging and cooling solutions. Leakage power is also a concern in low-end mobile system-on-chips where the low standby power feature is crucial. Recent energy estimates for 0.13 μm process indicate that leakage energy accounts for 30 percent of L1 cache energy and as much as 80 percent of L2 cache energy. With 3X increase in I_{OFF} every technology generation, caches will continue to account for a large component of leakage power dissipation in a microprocessor. From a researcher's viewpoint, SRAM caches are an interesting target for leakage reduction because of their inherent features of (1) iterative array structure, (2) low activity factor, and (3) unique architectural behaviors.

According to the International Technology Roadmap for Semiconductors [22], the number of devices will increase from about 1 billion/chip today to approximately 10 billion/chip in a decade. Majority of the devices will be used in on-die cache memories since micro-architectural performance can be improved without incurring large increase in dynamic power consumption. As a result, more than 50 percent of the chip area is already occupied by on-die caches in recent designs and the cache area will continue to grow with technology scaling. Adverse effect of having larger caches is the large leakage power which dominates the total chip power consumption. Thus, leakage power management in caches is indispensable in cutting-edge microprocessor designs for cost effective packaging and cooling solutions. Cache leakage control is also essential in low-end mobile system-on-chips (SoC) that have stringent standby power requirements for extended battery life. Recent announcements by industry experts reveal that active leakage power accounts for 40 percent of the total power consumption in today's high-performance microprocessor designs [23]. With 2-3X increase in device leakage every new technology generation, circuit techniques for controlling cache leakage are necessary for both high performance and low power consumption in nanoscale LSI systems.

There has been a spectrum of research activities to deal with the leakage power crisis at different levels of abstraction (device, circuit, architecture, and software). At the device level, novel transistor structures such as the double-gate or ultra-thin body MOSFET's are being developed where short-channel-effect is controlled by geometry of the device rather than the high impurity doping. This provides high on-current at a low off-current (better sub-threshold slope) and relieves the short-channel-effect potentially down to the ultimate limit of sub-20 nm channel length [24]-[26]. Dual Vt CMOS [27] have been suggested to reduce the overall leakage power without impacting circuit performance. Here, low Vt transistors are used in critical paths for high performance, and high Vt transistors are employed in non critical paths for low leakage power. Circuit level techniques that change the bias condition of transistors to achieve a low leakage state have been well studied. Power gating, dynamic Vdd, input vector control, body biasing are some examples [28]-[35]. Each of these techniques gives the opportunity to reduce microprocessor leakage in standby mode. A processor typically stays in standby mode for a considerable amount of time so the overhead delay and energy for activation or deactivation is relatively small. Architecture level techniques such as the dynamically-resizable instruction cache [36] has been proposed where only the memory space required to hold the working set of the current application is in active mode. The rest of the cache is in sleep mode for leakage reduction. To determine the adequate memory space, the miss rate is monitored using an adaptive hardware algorithm. Cache decay scheme [37] exploits the locality of reference in caches to turn-off portions of the cache which are not likely to be accessed. Zhang [38] investigated complier level techniques to reduce run-time leakage current in a VLIW processor using sleep vectors and sleep transistors.

Leakage control during circuit operation is more challenging than in standby mode due to the short time to deactivate blocks, large overhead energy, and run-time leakage variations. Fine grain leakage reduction techniques are necessary that can activate or deactivate small portions of a microprocessor so that unused functional blocks can be put into a low leakage state while the rest of the processor is still running. For an active leakage reduction scheme to be profitable, the overhead energy for activation or deactivation must be considerably smaller than the amount of leakage saved by the FBSRAM architecture.

2. FORWARD BODY-BIASED LOW-LEAKAGE SRAM CACHE

2.1 Introduction

Forward body-biasing (FBB) has proven to effectively improve performance, suppress short channel effects, and reduce Vt variations [1,2]. Vt rolloff and drain induced barrier lowering (DIBL) which limits scalability of channel length can be relieved by forward body-biasing devices during normal operation [1b]. In this section, a dynamic FBB scheme for low-leakage SRAM caches is presented where leakage power is significantly reduced by using super high Vt transistors [3]. FBB is dynamically applied to only the active portion of the cache for fast read and write operation. The idea of using a high Vt device and applying a FBB in active mode has already been discussed in previous literature [1,4]. They have also mentioned that withdrawing the FBB in standby mode can significantly reduce leakage power consumption. However, it is not clear how one can optimize the high Vt device for the FBB scheme in scaled technologies where different leakage mechanisms make it non-trivial to obtain a device with a desirable Vt. Moreover, it has not been reported whether a dynamic FBB scheme is profitable enough so that it can be applied in fine grain to reduce the cache leakage even in active mode. The goal of this section is to explore if such a fine grain dynamic FBB scheme can be useful in reducing active cache leakage and to devise a set of solutions to achieve best performance and cell stability under given leakage power constraints. For this, we developed combined device-circuit-architecture level techniques and compared the proposed forward body-biased SRAM (FBSRAM) with prior state-of-the-art techniques. This section makes the following contributions:

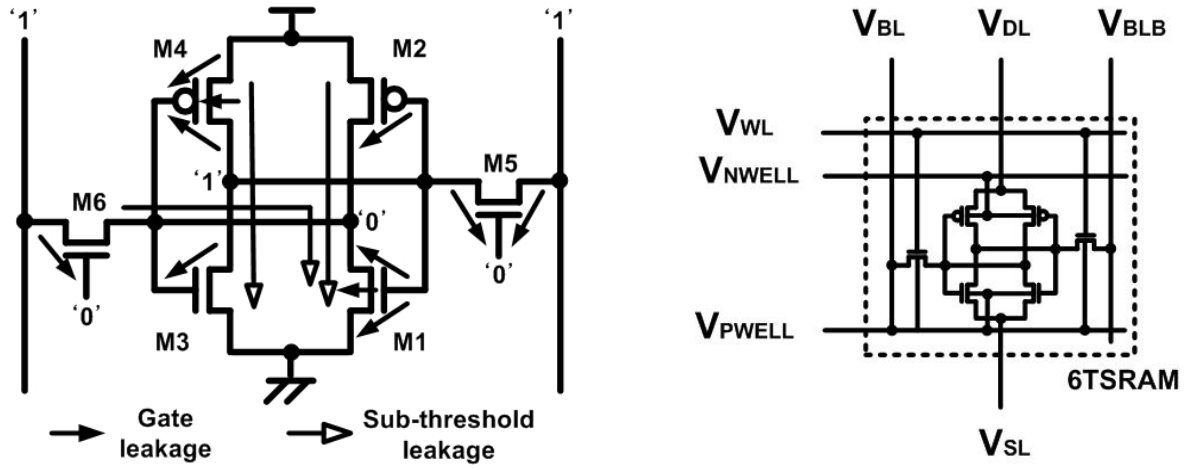
- For the first time, we apply the concept of using a super high Vt device and FBB to dynamically reduce the active leakage in cache memories.
- We show optimization of the 2-D halo doping profile of a super high Vt device to achieve total leakage reduction while suppressing the gate leakage and junction band-to-band tunneling (JBTBT) leakage.
- Circuit techniques and architectural behavior of caches are exploited to reduce FBB transition latency and energy overhead.
- We present results showing that FBSRAM achieves higher performance compared to a prior state-of-the-art low-leakage SRAM under iso-leakage conditions.

This section is organized as follows. In section 2.2, previous low-leakage SRAM cell techniques are introduced and evaluated. Section 2.3 deals with device optimization for the FBSRAM considering sub-threshold, gate and JBTBT leakage. Circuit and architecture techniques to reduce the FBB overhead are discussed in section 2.4. Simulation results of overall leakage savings, performance, and cell stability are presented in section 2.5. Finally, we conclude the forward body-biased low-leakage SRAM cache part of the report in section 2.6.

2.2 Low Leakage SRAM Cells

Figure 1(a) shows the dominant leakage components in a conventional six transistor SRAM cell during standby mode. When the cell is not accessed, the word line signal is low and the two bit lines are precharged to high. One of the subthreshold leakage components comes from the bit line and contains the access transistor M6, which is in off state. The other two sub-threshold leakage components are the cell leakage through the off state devices M2 and M3. Gate leakage components of a transistor depend on the voltage across its terminals. Gate tunneling leakage of a PMOS device is substantially lower than that of an NMOS device. This is due to the barrier height for HVB (hole tunneling from the valence band, 4.5 eV) being significantly larger than that for ECB (electron tunneling from the conduction band, 3.1 eV) in the SiO_2/Si system [5]. Out of the various sources, the most dominant gate leakage comes from the inversion mode device M1 having tunneling components through the (1) gate-channel, (2) gate-source overlap and (3) gate-drain overlap regions.

Figure 1(b) shows the seven available terminals in a conventional 6T SRAM cell; V_{SL}, V_{PWELL}, V_{NWELL}, V_{DL}, V_{WL}, V_{BL}, and V_{BLB}. Various SRAM cell architectures have been proposed in the past where one or more of the seven terminal voltages are controlled during standby mode for reducing the leakage components shown in Figure 1(a). Each technique exploits the fact that the active portion of a cache is very small, which gives the opportunity to put the large idle portion in a low-leakage sleep mode. Effectiveness and overhead of each technique are evaluated based on the discussions, described in the following paragraphs.



(a) (b)

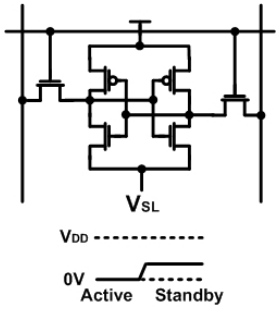
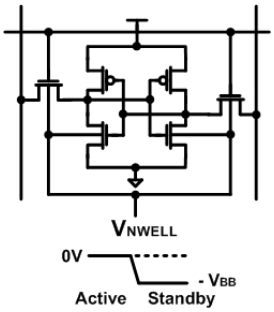
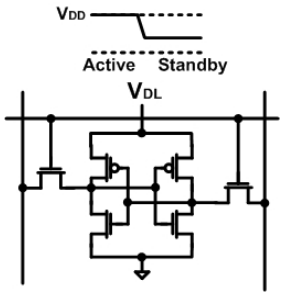
Figure. 1. (a) Dominant leakage components in a 6T SRAM cell. (b) Seven terminals of a 6T SRAM cell.

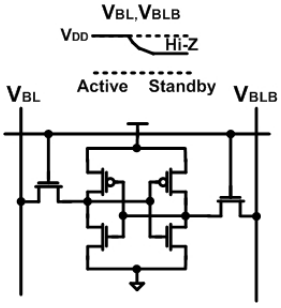
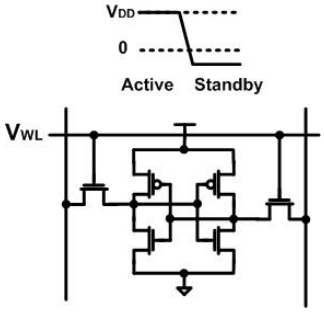
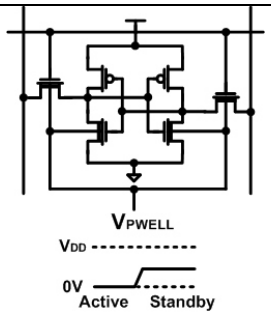
First, the impact of the technique on various leakage components must be considered. Although sub-threshold leakage still continues to dominate the I_{OFF} at high temperatures, ultra-thin oxides and high doping concentrations have led to a rapid increase in direct tunneling gate leakage and JBTBT leakage in the nanometer regime. Each leakage reduction technique needs reevaluation in scaled technologies where sub-threshold conduction is not the only leakage mechanism. Second, the impact of the leakage reduction technique on SRAM read/write delay should be considered. Third, the transition latency and energy overhead should be taken into account because of the limited time and energy budget for the mode transition. Last, the leakage reduction technique should not have a noticeable impact on SRAM cell stability or soft error rate (SER). Previously proposed low-leakage SRAM cells are summarized in Table 1 based on which of the seven terminal voltages in Figure 1(b) are controlled during standby mode. Advantages and disadvantages of the FBSRAM proposed in this work are also described for comparison.

Source biasing scheme raises the source line voltage in sleep mode to generate a negative Vgs in the access transistor and reduce the bit line leakage [6,7,8,9]. The reduced signal rail (V_{DD}-V_{SL}) and the body effect in the NMOS transistors also lowers the sub-threshold leakage in the SRAM cell. Since raising the source line voltage has the similar effect as reducing the supply voltage, the gate leakage is also reduced due to the lower voltage stress across the device terminals. An extra NMOS device has to be series connected in the pulldown path to cut off the source line from the actual ground during sleep mode, and this imposes a delay penalty. The reduced signal charge in sleep mode also causes the SER to rise, which requires additional error correction coding circuits [9].

Reverse body-biasing (RBB) the NMOS (or PMOS) devices can reduce sub-threshold leakage in sleep mode via body effect. The access time is unaffected by having a zero body-bias (ZBB) in active mode [8,10,11]. A large latency and energy overhead is imposed for the body-bias transition due to the large body-bias swing and the parasitic RC components in the substrate. This scheme becomes less effective in nanoscale dimensions as JBTBT leakage gets enhanced by RBB [1].

TABLE 1
PREVIOUSLY PROPOSED LOW-LEAKAGE SRAM CELL TECHNIQUES. (* MAIN LIMITATION)

Scheme	**Source Biasing (V_{SL})**	**Reverse Body-biasing (V_{PWELL}, V_{NWELL})**	**Dynamic V_{DD} (V_{DL})**
			
Ref.	[6], [7],[8],[9]	[8], [10], [11]	[8], [12]
Leakage	Sub-threshold, gate: ↓↓	sub-threshold: ↓↓ *JBTBT: ↑	sub-threshold, gate: ↓ *bit line leakage: -
Perform.	*Delay increase	No delay increase	No delay increase
Overhead	Medium transition overhead	Large transition overhead	Large transition overhead
Stability	Impact on SER	No impact on SER	*Impact on SER

Scheme	**Floating Bit Lines (V_{BL}, V_{BLB})**	**Negative Word Line (V_{WL})**	**Forward Body-biasing + Super high Vt (V_{PWELL})**
			
Ref.	[13]	[14]	[3]
Leakage	sub-threshold, gate: ↓	sub-threshold: ↓ *gate: ↑	sub-threshold: ↓↓ gate: -
Perform.	No delay increase	No delay increase	Minimal delay increase
Overhead	*Precharge latency overhead	*Low charge pump efficiency	*Area overhead, process complexity
Stability	No impact on SER	No impact on SER, high voltage stress	No impact on SER

Supply voltage is lowered in a dynamic V_{DD} SRAM (DVSRAM) [8,12] to reduce the sub-threshold, gate, and JBTBT leakage. The bit line leakage however, cannot be reduced using this scheme since the bias condition in the access transistors does not change. This scheme requires a smaller signal rail (V_{DL}-V_{GND}) compared to the SBSRAM for equivalent leakage savings since unlike the SBSRAM scheme, it doesn't have the negative Vgs effect. Although there is no impact on delay during active mode, the large V_{DD} swing between sleep and active mode imposes a larger transition overhead compared to the SBSRAM. Moreover, the greatest drawback of the DVSRAM is the substantial increase in SER with voltage scaling.

A technique that lets the bit lines float during standby mode has been proposed to reduce the bit line leakage via DIBL [13]. Since only the access transistors benefit from this technique, the overall leakage savings is marginal. Unlike the three previously mentioned techniques, this scheme cannot be applied to individual cache lines since the

bit line is shared across different cache lines. Normally, bit lines have to be precharged and ready for the word line access. Since the bit lines are floating for this technique, an extra precharge cycle is required whenever a new subarray is accessed. This would mean that modification in the pipeline is inevitable for handling the multiple hit times [12,20].

The negative word line scheme [14] has been proposed to cut off the sub-threshold leakage through the access transistors. However, the designer has to resolve issues such as the increase in gate leakage and higher voltage stress in the access transistors by using a lower voltage in cells and bit lines. Although this technique has no impact on performance or SER, there is a dynamic power overhead for generating the negative voltage [7].

2.3 Device Optimization For Forward Body-Biasing

Previous low-leakage SRAM cell techniques start with a given device and try to utilize circuit and architectural means to minimize the performance loss associated with the additional circuitry. The main difference of our proposed scheme compared to others is that we also consider the device level optimization into the low-leakage SRAM cache design. The FBSRAM technique that we propose lowers the active leakage using super high Vt devices, i.e., utilizing channel doping techniques or work function engineering. To obtain fast read and write operation, the NMOS transistors in the selected subarray are dynamically switched from ZBB to FBB. Next, we will explain the various leakage components in a nanometer regime MOSFET and discuss how collaboration between device engineering and FBB can effectively lower active leakage and achieve high drive current.

2.3.1 Nanoscale Leakage Mechanisms

We first describe the five short channel leakage mechanisms illustrated in Figure 2. Currents I_1, I_2, I_3, and I_4 constitute the drain leakage while I_5 is the gate leakage component. I_1 is the sub-threshold leakage, which occurs when gate voltage is below Vt. Weak inversion conduction typically dominates modern device I_{OFF} due to the aggressively scaled Vt. I_2 is due to the high field effect in the drain-gate overlap region, referred to as the gate induced drain leakage (GIDL). The narrowing of depletion layer at or near the surface causes increase in the local electric field, and thereby enhancing the high field effects [15,16]. I_3 is the reverse biased p-n junction leakage due to (1) the minority carrier diffusion/drift near the edge of the depletion region and (2) the electron-hole pair generation in the depletion region. High electric field ($> 10^6$ V/cm) across the p-n junction also causes significant JBTBT of electrons from the valence band of the p-region to the conduction band of the n-region. The higher doping levels due to well doping or halo implant near the channel edges causes larger JBTBT, which makes it more challenging to suppress the short channel effect as technology scales [16]. In short channel devices, due to the proximity of the drain and the source, the depletion regions at the drain-substrate and source-substrate junctions extend into the channel. An increase in the reverse bias across the junctions (with increase in Vds) leads to the merging of the depletion regions, causing punchthrough (I_4). Gate direct tunneling current I_5 is due to the tunneling of electron (or hole) from the bulk silicon through the oxide potential barrier into the gate.

In this report, we will divide the various leakage components into three categories; (1) sub-threshold leakage I_1, (2) direct tunneling gate leakage I_5, and (3) JBTBT currents I_2+I_3 through the drain-well junction.

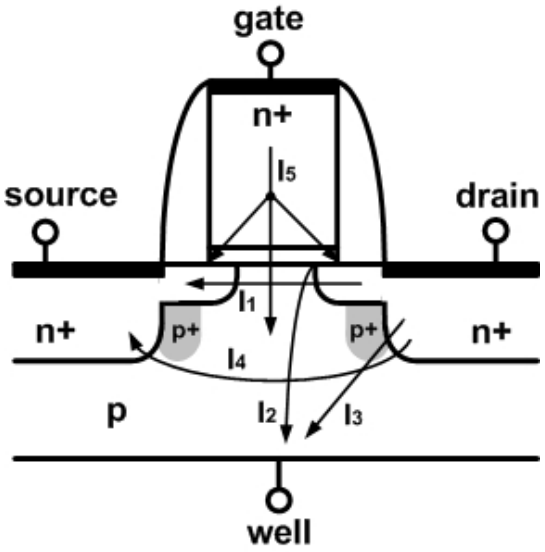


Figure 2. Leakage current mechanisms of a nanometer regime MOSFET (Vgate=Vsource=Vwell=0, Vdrain= V_{DD}). The shaded p+ region indicates the area with halo doping.

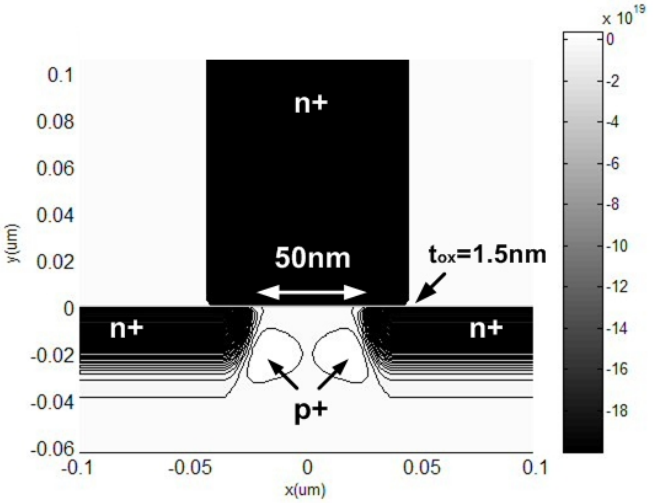


Figure 3. Dimensions and doping profile of the 50nm L_{EFF} device. (V_{DD} =1.0V, T=110˚C)

2.3.2 Device Optimization

2.3.2.1 Super-halo 2-D Doping Profile

A 50 nm effective channel length (L_{EFF}) device based upon the super-halo discussion by Y. Taur et al., has been incorporated for the MEDICI simulations [16,17]. The dimensions and doping profile of the device are shown in Figure 3. The nominal NMOS device has a physical oxide thickness (t_{ox}) of 1.5 nm and a Vt of 270 mV. Super-halo uses a non-uniform p+ doping in the source-body and drain-body boundaries to reduce the source-drain depletion width, and effectively suppresses the body punchthrough [16]. Vt roll off and DIBL are also controlled by the 2-D halo doping profile. Both the p+ halo and the n+ source/drain doping regions are modeled as 2-D Gaussian functions. The device profile is mirror symmetric about the vertical line x = 0 in Figure 3.

Vt of a super-halo device can be effectively changed by adjusting the halo doping concentration or varying the halo implant location or angle [18b]. Changing the background channel doping is less effective because (1) Vt is less sensitive to channel doping, (2) DIBL and punchthrough cannot be suppressed as effectively, and (3) the impact on drive current is more severe [19]. Other general means to raise the device Vt is to have a thicker physical t_{ox} or a longer channel length. However, the former will worsen the short channel effect and the latter will increase the load capacitance and area of the FBSRAM.

2.3.2.2 Super High Vt Device Design

A new super high Vt (Vt = 350 mV) doping profile for the FBSRAM is generated by raising the peak halo doping concentration of the nominal Vt device. One of the concerns with increasing the peak halo doping is the exponential increase in JBTBT current. As the depletion width gets narrower with higher doping levels, the field across the p-n junction becomes higher, causing more number of electrons to tunnel through the drain-body junction. We were able to generate a super high Vt device that gives equivalent leakage savings as prior art [6,7,8,9] while keeping the JBTBT current to be less than 3 percent of the total leakage consumption.

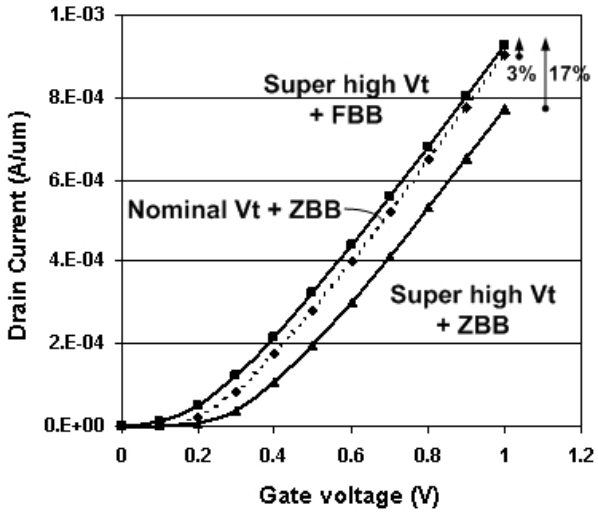


Figure 4. NMOS drain current with and without forward body-biasing for a nominal Vt device (Vt_nom=270mV) and a super high Vt device (Vt_high=350mV).

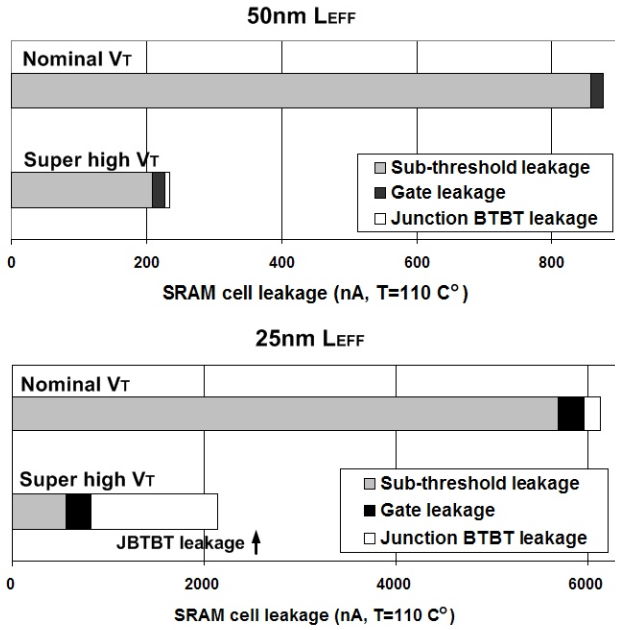


Figure 5. SRAM cell leakage of nominal Vt device and super high Vt device. The super high Vt device is generated using channel engineering.

Figure 4 shows the drain current of the super high Vt device compared to the nominal Vt device. Under ZBB, device I_{OFF} of the super high Vt device is reduced by 4X offering a low standby leakage. By applying a 500 mV FBB to the super high Vt device, I_{ON} is improved by 17 percent, offering a three percent higher drive current compared to the nominal Vt device. The total leakage of a 6TSRAM cell shown in Figure 5 indicates that the reduction is mainly due to the improvement in sub-threshold leakage dominating the I_{OFF} in high temperatures.

2.3.2.3 Scaling Trends

As mentioned in section 2.3.2.2, the super high Vt device accomplished by raising the halo doping concentration has a higher JBTBT leakage (Figure 5, top). This will become more evident in future process generations as the halo concentration needs to increase to suppress the worsening Vt roll off and DIBL. Raising the halo on top of the increased doping levels causes the JBTBT leakage to become unacceptable in sub-25 nm (Figure 5, bottom). In technologies where one cannot afford a higher halo doping, a super high Vt device can be built using a gate material with a higher flat band voltage [19]. Figure 6 shows that a super high Vt device can be built without impacting the JBTBT leakage by engineering the work function of the gate material.

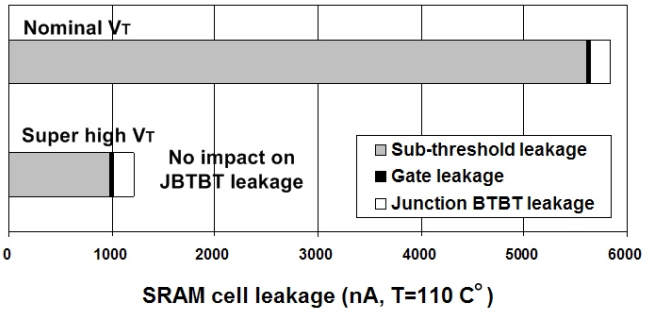


Figure 6. SRAM cell leakage of nominal Vt device and super high Vt device. The super high Vt device is generated using work function engineering.

2.4 FBSRAM Instruction Cache Design

2.4.1 Subarray-by-subarray Leakage Control Scheme

Figure 7 shows the circuit diagram of a 32bx32b FBSRAM subarray with body-bias drivers M1-M3. Each subarray has a subarray select signal (SUBSL) which is generated by the row decoder circuit. When a subarray is selected for access, the SUBSL signal goes high and M1 and M2 are turned on. This switches the NMOS body-bias (V_{PWELL}) to a 0.5V FBB and increases the drive current for fast read/write operation. On the other hand, if a subarray is not accessed, the SUBSL signal stays low and a ZBB is applied to the super high Vt devices via M3. This substantially reduces the total I_{OFF} during the inactive periods. The 0.5V FBB voltage can be generated by a high-efficiency DC-DC converter and is routed throughout the SRAM array using a mesh to reduce the voltage fluctuation. Triple well technology is required for n-well processes to isolate the NMOS body of a particular subarray from its neighbors. The V_{PWELL} line can be routed using an upper layer metal, so the only area overhead comes from the boundary of each subarray where design rule requires an area margin for well isolation. This area overhead, however, is significantly less than previous row-by-row body-biasing techniques where each cache line is isolated from the adjacent ones [10b,11b]. Triple well is not a requirement in case a p-well process is available. Noise through the substrate junctions while switching the body bias must be carefully considered when implementing the FBSRAM.

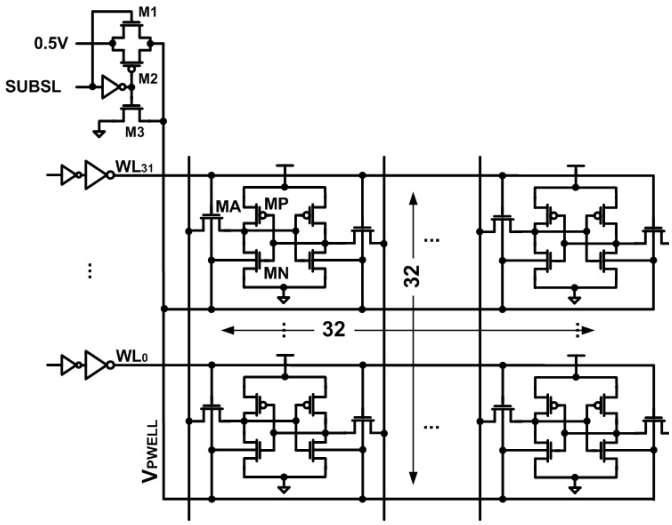


Figure 7. 32b x 32b FBSRAM subarray with body-bias drivers.

Switching between FBB and ZBB requires additional transition delay and energy. Four NMOS transistors residing on the p-well in each SRAM cell yield six drain-body capacitances and four gate-body capacitances to charge and discharge in every body transition event. The p-well resistance (R_{SUB}) and the bottom capacitance (C_{TUB}) between the p-well and the underlying deep n-well (or n-substrate) also have impact on the body-bias transition time. To estimate the body transition delay considering the parasitics, we ran HSPICE simulations assuming an R_{SUB} of 13 ohm·cm and using a C_{TUB} value based on the following equation for deriving the junction capacitance of a p-n diode.

$$C_{TUB} = \sqrt{\frac{q\varepsilon_s}{2(V_{bi} - V_{PWELL})} \frac{N_a N_d}{N_a + N_d}} \tag{1}$$

Since the source and drain junction capacitances are already included in BSIM4 device models, we only need to attach external components of R_{SUB} and C_{TUB}. q is the charge of an electron, ε_s is the permittivity of the semiconductor, V_{bi} is the built-in voltage, V_{PWELL} is the NMOS body voltage, N_a is the p-doping, and N_d is the n-doping. Figure 8 shows the waveforms of SUBSL and V_{PWELL} for estimated C_{TUB} values. Note that results are also shown under a pessimistic condition with 10X higher C_{TUB}. As shown in the figure, there is a large delay for the V_{PWELL} to reach a 500mV FBB after the SUBSL goes high. It is unacceptable to pay this delay overhead every time

a ZBB to FBB transition is required. The following discussions in section 2.4.2 and 2.4.3 show how the transition latency and energy overhead can be alleviated in the proposed FBSRAM architecture.

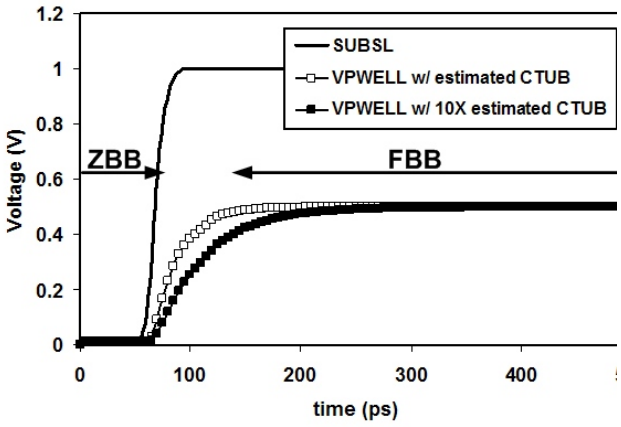


Figure 8. ZBB to FBB transition of NMOS body voltage (V_{PWELL}) for estimated C_{TUB} values.

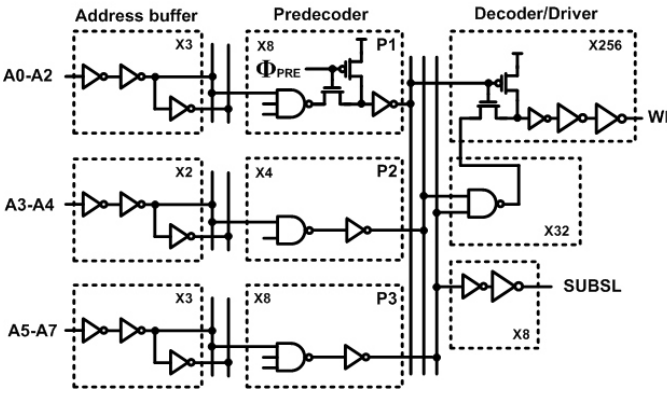


Figure 9. Row decoder circuit generates SUBSL signal which activates the selected subarray ahead of time.

2.4.2 Transition Latency Hiding

SUBSL signal in Figure 7 triggers the ZBB to FBB transition. Figure 9 shows the row decoder circuit that generates the SUBSL signal even before the word line signal arrives at the cells. The decoder has three predecoder blocks P1, P2, P3 for the 3 groups of address inputs A0-A2, A3-A4, and A5-A7. The output of predecoder block P1 is gated by the predecoder clock, Φ_{PRE}. This clock signal makes the word line signals to operate in a precharge-evaluate fashion so that all the word lines will stay at low while the bit lines are in the precharge cycle. Output of predecoder block P3 is used to determine which 32b x 32b FBSRAM subarray is selected. For any given address input, only one of the outputs from P3 is high.

The timing diagram in Figure 10 illustrates the operation of the FBSRAM. V_{PWELL} transition is triggered by the SUBSL signal. The SUBSL signal is generated as soon as the address bits arrive at the row decoder block. Φ_{PRE} is asserted after the address bits arrive so that the previous cycle data will not ripple through the decoders and result in a false evaluation. We exploit the difference in arrival time between the address bits A0-A7 and Φ_{PRE} for the body-bias transition to complete. We also complete the body transition operation and decoding operation in parallel, so that there is a minimal increase in access time due to the body-bias transition.

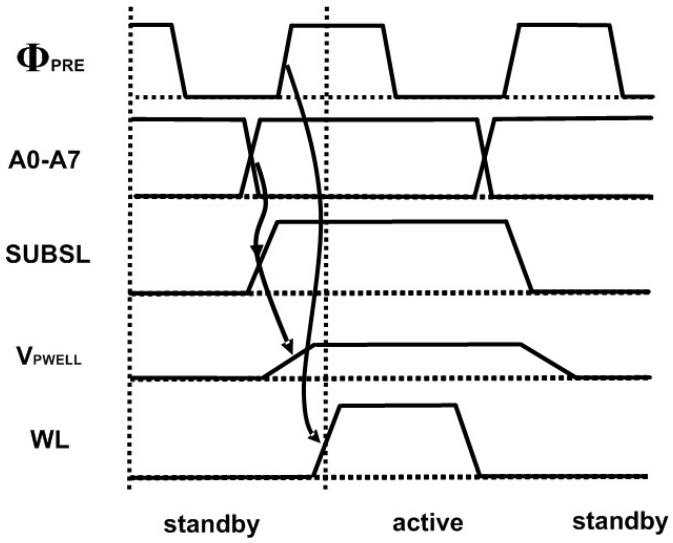


Figure. 10. V_{PWELL} is switched from ZBB to FBB before the word line (WL) signal arrives at the cells. This hides the transition latency overhead associated with body-bias transition.

2.4.3 Transition Energy Reduction

The SUBSL signal in Figure 9 is generated from the most significant bits (MSB) A5-A7, and thus it will not toggle unless these three bits change. Although hiding the transition latency as explained in section 2.4.2 enables an early body-bias transition, the transition energy remains unchanged. However, observation of the cache access pattern reveals that the number of body-bias transitions is significantly less than the worst case. When data is first brought into a cache, it experiences a burst of accesses. After the flurry of accesses, there is a considerably long period of time between the last access and the point when the data is replaced, referred to as the "dead period" [20]. This implies that there is (1) a high probability that a subarray in access will be accessed again in the next cycle and that (2) a subarray in sleep mode is more likely to stay in sleep mode throughout the dead period. This behavior is often referred to as the locality of reference in a cache and can be visualized in Figure 11. We used the SimpleScalar-3.0 toolset to extract the architectural behavior of an out-of-order processor with a 32 KB, 4-way L1 instruction cache. We ran 500 million instructions after skipping the initial 500 million to extract the architectural behavior of the actual benchmark application that we are interested in. Results show that the average percentage of accesses hitting the same subarray in the consecutive cycle is 93 percent for SPEC2000 benchmark applications. The ZBB to FBB transition happens in only seven percent (100 percent minus 93 percent) of the total accesses, and thus the energy overhead of the proposed FBSRAM cache is considerably low.

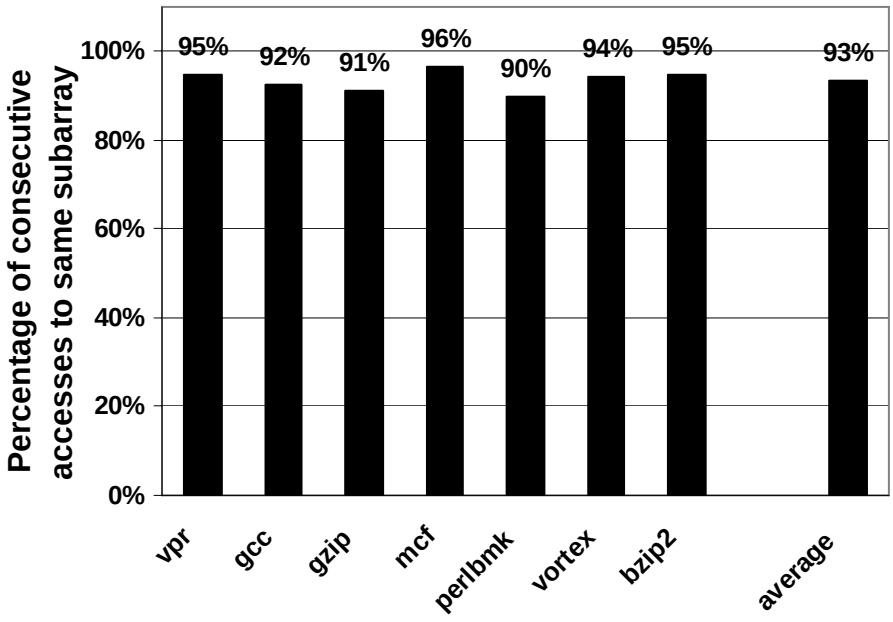


Figure 11. Percentage of hitting the same subarray in consecutive cycles for SPEC2000 benchmark applications (32 KB, 4 way, L1 instruction cache). Access to a different subarray happens in only seven percent of total accesses.

2.5 Simulation Results

Total leakage power savings, performance impact and static noise margin (SNM) of the proposed FBSRAM was derived through extensive MEDICI and HSPICE simulations. Despite the long simulation time and limited number of devices that can be included, device simulators such as MEDICI provide detailed information on the various leakage components. This helps understand the leakage power components in deep submicron circuits. For SNM calculations, we have used HSPICE instead of MEDICI due to convergence problems in the device simulator, and because of the fact that SNM is governed by the device I_{ON}. All MEDICI simulations use an L_{EFF} of 50nm, a supply voltage of 1.0V and threshold voltages of 270mV (nominal device) and 350 mV (super high Vt device). BPTM 70 nm technology with the same supply and threshold voltage was used for the HSPICE simulations. The L1 instruction cache for comparison has the same geometry as the one given in section 2.4.3 (32 KB, 4-way, 32b x 32b subarray).

We have compared the proposed FBSRAM technique with both conventional and SBSRAM. The reason why we selected the SBSRAM for comparison is because it has already been proven through silicon measurements to effectively reduce leakage with minimal impact on performance [9]. Figure 12 compares the leakage power dissipation of the three SRAM configurations. Devices for the conventional and SBSRAM have a nominal Vt of 270 mV. We used a source biasing voltage of 0.2 V during inactive mode, which gives equivalent leakage savings as the FBSRAM utilizing super high Vt devices. By raising the source line voltage from 0 V to 0.2 V, the SBSRAM was able to reduce the total leakage power by 64 percent including the dynamic power overhead. The overhead dynamic power and the leakage from the selected SRAM cells account for eight percent of the total leakage power. The FBSRAM achieves iso-leakage savings as the SBSRAM by applying ZBB to the unselected portion of the cache.

Under iso-leakage conditions, bit line delays of the SBSRAM and FBSRAM are compared, together with the conventional scheme. We define the bit line delay to be the time for the differential voltage between the bit lines to reach 100 mV. Figure 13 shows that the bit line delays of both the FBSRAM and SBSRAM are larger than the conventional SRAM cell. The reason for the longer delay in the SBSRAM is the extra series NMOS device in the pull down path which aggravates the drive current. The larger bit line capacitance compared to conventional due to the increased junction capacitance is the main reason for the performance loss in FBSRAM. We have shown in

Figure 3 that the drive current of a super high Vt device with FBB is three percent higher than a nominal Vt device with ZBB. However, the junction capacitance of the super high Vt device is larger due to the FBB and increased halo doping. This results in a larger bit line capacitance and hence the bit line delay of the FBSRAM is 10.1 percent (138 ps→152 ps) slower than the conventional cell. However, if we compare the bit line delays of the two low-leakage SRAM schemes, FBSRAM (152 ps) turned out to be faster than the SBSRAM (164 ps) by 7.3 percent. The bit line delay of the SBSRAM can be improved by increasing the width of the sleep transistor. However, this will impact the leakage savings, increase the transition energy overhead, and worsen the impact of the sleep transistor on SRAM cache area [6].

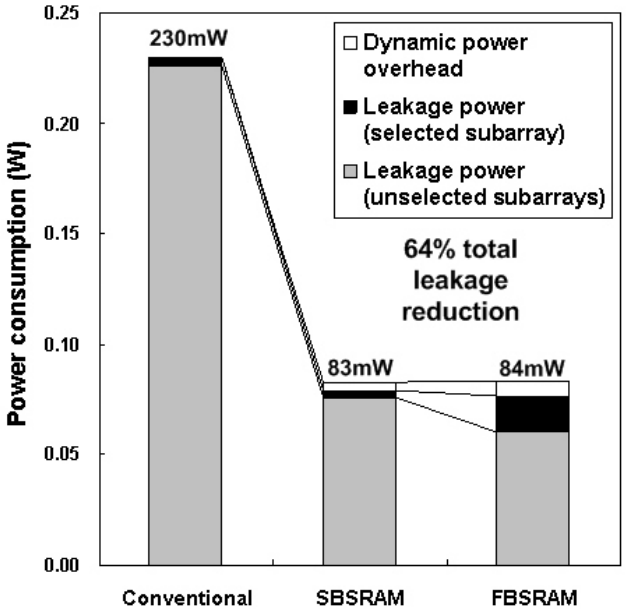


Figure. 12. Leakage power and dynamic power overhead of 3 SRAM schemes (50nm L_{EFF}, V_{DD} =1.0V, T=110˚C).

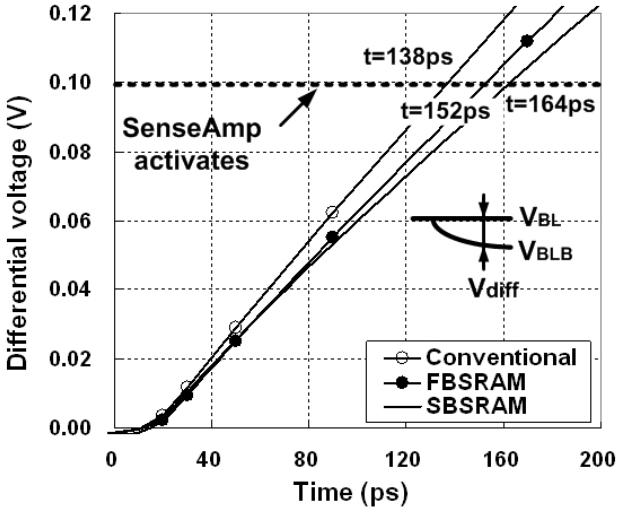


Fig.ure 13. Differential voltage (=V_{BL}-V_{BLB}) of FBSRAM compared to conventional and SBSRAM (50nm L_{EFF}, V_{DD} =1.0V, T=110˚C).

SNM of an SRAM cell is defined by Seevinck [21] as the minimum dc noise voltage necessary to change the state of a cell. Seevinck et al. shows that the SNM is dependent on the magnitude of Vt, V_{DD}, and the ratios of β's (transconductance factor) of MA, MN and MP in Figure 7. Static transfer characteristic of the proposed FBSRAM is simulated using HSPICE and is shown in Figure 14. The SNM can be visualized as the maximum width of the enclosed square in the superimposed voltage transfer curves of V(Q) and V(QB).

SNM of the three different SRAM schemes during read operation and standby mode are shown in Table 2. We have developed a software that automatically simulates a given SRAM cell circuit and extracts the SNM value from the voltage transfer curves. Results show that the SBSRAM has 52 percent higher SNM during read mode than a conventional cell. This is because the stacking effect due to the bottom transistor causes the effective Vt of the cell transistors to become higher in the SBSRAM. The higher Vt in the SRAM cell makes it more difficult for the access transistors to destroy the data, which translates into a higher SNM. For the FBSRAM, having a super high Vt device to start with, gives 17 percent increase in SNM compared to conventional during standby mode. As we forward body-bias these super high Vt device, the Vt is effectively lowered and the SNM improvement drops to six percent. Overall, SNM of the proposed FBSRAM is comparable to conventional.

TABLE 2
STATIC NOISE MARGIN OF DIFFERENT SRAM SCHEMES (50NM, VDD=1.0V)

	Conventional	SBSRAM	FBSRAM
SNM (read)	104mV	158mV	110mV
SNM (standby)	277mV	266mV	324mV

As device scaling continues into the sub-90nm regime, the ultra-thin oxide to control the short channel effect has caused the gate direct tunneling leakage to become a major leakage component. Although the proposed forward body-biasing technique does not reduce the gate leakage component, optimizing the super high Vt device using a thicker t_{ox} can reduce both sub-threshold and gate leakage. Joint optimization between device, circuit, and architecture for gate leakage dominated technologies can be a future research topic.

2.6 Conclusion

Previous low-leakage SRAM architectures have inherent limitations due to either delay overhead, limited effectiveness in reducing leakage, multiple cache hit times, or impact on cell stability. Even state-of-the-art low-leakage SRAM cell techniques such as the source biasing scheme has issues such as large performance penalty and degradation in SER.

This section starts from a simple initiative of utilizing super high Vt devices for low-leakage and forward body-biasing them for high performance. We have looked at different levels of the design (device, circuit, and architecture) to effectively reduce leakage power and still achieve high performance. At the device level, the super high Vt doping profile was optimized to improve the DIBL and Vt roll-off while suppressing the JBTBT leakage component. Transition latency associated with FBB was hidden by modifying the decoder circuit to give an early notice to the subarray that is to be accessed. At the architectural level, the general cache access pattern is exploited to lower the body-bias transition energy. As a result, the combined device-circuit-architecture level techniques achieve 64 percent total leakage reduction (overhead included) and 7.3 percent improved bit line delay over prior art with no impact on cell stability.

3. LOW LEAKAGE SOURCE-BIASED CACHE

3.1 Introduction

For an active leakage reduction scheme to be profitable, the overhead energy for activation or deactivation must be considerably smaller than the amount of leakage that can be saved by the technique. As the leakage distribution gets wider with technology scaling, circuit designers need to guarantee that the overhead energy is still smaller than the amount of leakage that can be reduced. Aggravating parameter fluctuations with device scaling has resulted in a 100X or more variation in device leakage. This is shown in Figure 14 for a 180 nm CMOS technology where channel length or doping density variations result in a considerable dispersion in I_{on} (transistor on current) and I_{off} (transistor off current). Effectiveness of an active leakage reduction technique will change significantly due to the large leakage variation and relatively constant dynamic overhead energy.

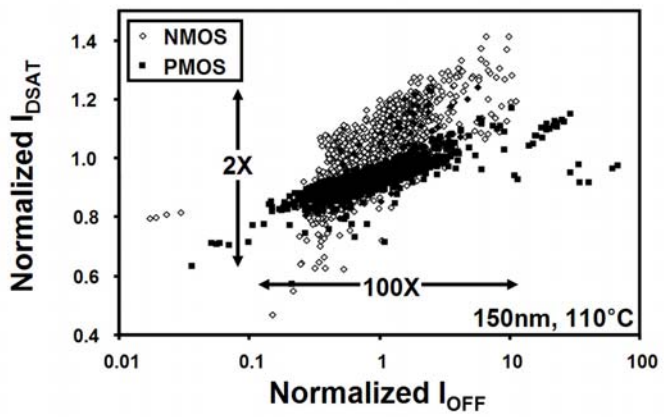


Figure 14. Device I_{on}, I_{off} measurements in 150 nm CMOS technology. More than 100X difference in device leakage is observed due to worsening intrinsic parameter fluctuations.

This section describes a low-leakage source-biased SRAM and an efficient way to utilize this technique under severe leakage variations by exploiting the architectural behavior of a cache. Unlike prior techniques where SRAM blocks are unconditionally turned off after the access, the proposed technique has a self-decay based sleep pulse generator that tracks the Process, Voltage and Temperature (PVT) fluctuations and determines how often the SRAM blocks should enter a sleep mode for optimal power savings. Using the proposed scheme, SRAM blocks are more aggressively put into a low leakage mode under high leakage conditions and less aggressively under low leakage conditions. Simulation results show that the proposed scheme offers 27 percent lower leakage power (overhead included) compared to the prior art. A 16 kB SRAM test chip has been fabricated and tested in a 0.18 μm 6M CMOS technology to validate the effectiveness of the proposed technique. 94.2 percent cell leakage reduction with a 2 percent performance penalty was achieved for the 3.2mmx2.9mm die. Experimental results also show 25 percent improvement in read static noise margin (SNM) for our proposed SRAM cell.

3.2 Self-Decay Based Active Leakage Reduction For SRAM Caches

Figure 15(a) shows the source-biased gated-ground SRAM cell where the virtual ground voltage (VGND) is raised during sleep mode to generate a negative gate-to-source voltage in the access transistors, reducing the bit line leakage by orders of magnitude [39]-[7]. Reduced Drain Induced Barrier Lowering (DIBL) and the body effect raise the V_t of the cell transistors and lower the sub-threshold leakage. Raising VGND has the similar effect as reducing the supply voltage so the gate tunneling leakage also reduces because of the lower voltage stress across the device terminals [40]. An extra NMOS sleep transistor has to be inserted in the pulldown path in order to cutoff the source line from the actual ground during sleep mode. The source-biased SRAM technique requires additional error correction coding circuits to handle the increased soft error rate due to the smaller signal charge [9]. Charging/discharging the VGND of an SRAM block and switching in the extra control circuits requires additional overhead energy. To achieve positive leakage savings, the dynamic overhead energy must be kept low. Architectural access pattern of caches reveals that a simple circuit technique can significantly lower the overhead energy by reducing the number of VGND transitions [37]. When data is first brought into an L1 cache, it experiences a burst of accesses. After this live period, follows a considerably long dead period where the particular cache line is not accessed and the data is replaced. Conventional techniques that immediately turn off cache blocks after they are accessed cause the number of sleep and wake-up procedures to increase during the flurry of accesses to the “hot” cache blocks [39]-[7]. Our proposed technique in Figure 2(b) activates the SRAM block immediately upon access, which is

similar to conventional techniques. However, when it comes to turning off the cache blocks, our proposed technique periodically sends the cache blocks to sleep so that a once activated cache block is kept in active mode until the next sleep pulse arrives. Since the VGND of a cache block stays at 0 V after the access, the flurry of soon-after accesses will not require a VGND transition. In this way, unnecessary wake-up transitions can be avoided during the live period, while the majority of the cache blocks in the dead period can be kept in sleep mode. The periodic sleep pulse is generated from a self-timed sleep pulse generator circuit.

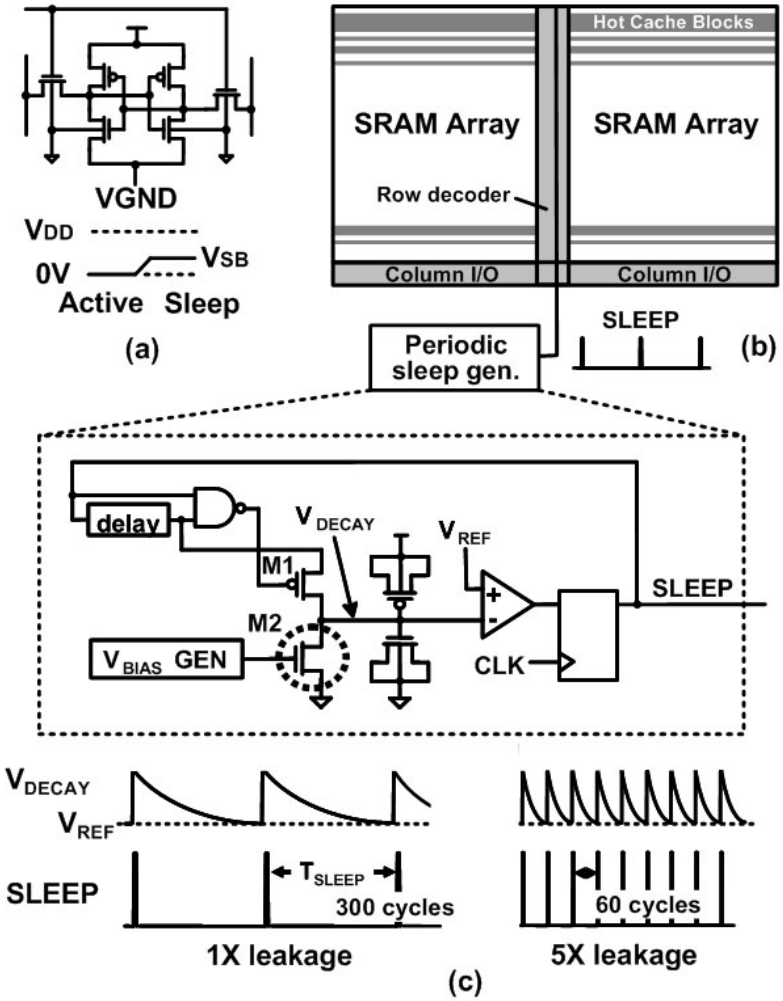


Figure 15. (a) Source-biased gated-ground SRAM cell for leakage reduction. (b) Proposed active leakage reduction scheme which periodically shuts off SRAM using a sleep pulse. (c) Proposed self-decay circuit that adaptively changes the interval between sleep pulses for optimal leakage savings under varying leakage conditions.

For best tradeoff between leakage savings and overhead energy under severe leakage variations, T_{DECAY} (interval between the sleep pulses) needs to be determined adaptively. At low leakage conditions (slow process, low temperature), it is optimal to have a long T_{DECAY} that will seldom turn off the SRAM. This reduces the dynamic energy overhead which is large compared to the relatively small leakage savings. On the other hand, the block must enter the sleep mode more frequent at high leakage conditions (fast process, high temperature) since the amount of leakage the can be saved becomes larger than the overhead energy. We propose a sleep signal generator based on a self-decay circuit that is capable of tracking the process and temperature fluctuations during run-time, providing a near-optimal T_{DECAY} under varying leakage conditions. The self-decay circuit in Figure 15(c) operates as follows. When V_{DECAY} drops to the reference level V_{REF}, the comparator output is sampled by the system clock (CLK). The SLEEP signal synchronized with the clock, is fed back to charge V_{DECAY} and the same procedure is repeated. Since the decay rate is determined by the leakage of M2 in Figure 15(c), T_{DECAY} becomes shorter at high leakage conditions and vice versa. The SRAM block is activated (VGND = 0) immediately upon access and is deactivated (VGND > 0) by the periodic SLEEP pulse. This action significantly reduces the number of unnecessary VGND transitions without sacrificing the leakage savings. Transistor M2 is biased in sub-threshold region by having V_{BIAS} smaller than device V_t. The precharged V_{DECAY} drops due to the M2 leakage, and once it reaches the reference level V_{REF}, the comparator output is flagged and sampled by the system clock (CLK). The output of the comparator is the

SLEEP pulse signal that sends the entire SRAM array to a sleep mode. Once the clock-synchronized SLEEP pulse is fired, it is fed back to precharge V_{DECAY} via M1 and the same procedure is repeated. Since the decay rate is determined by the leakage of M2 in Figure 2(c), T_{DECAY} becomes shorter at high leakage conditions and vice versa.

Figure 16 shows organization of a 16 KB SRAM for active leakage reduction based on the self-decay scheme. The SRAM is divided into 64 blocks, each having 4 by 512 cells. VGND of each SRAM block is separated for independent control and the sleep transistors are shared among the SRAM cells for dense layout. Issues of current crowding and ground bounce are resolved by having the sleep transistors distributed across the SRAM block. When the block enters sleep mode, VGND is raised to V_{SB} via device MP1 (Figure 16). Since entering sleep mode is not time critical, MP1 is sized to be 6 percent of the total sleep transistor width and is physically placed inside the row decoder block. This ensures that the peak current for turning off the entire SRAM is low. The crossed-coupled inverters in Figure 3 preserve the previous state for $\overline{BLOCK\_SEL}$=V_{dd} and SLEEP=0, maintaining a once activated SRAM block in active mode. For entering sleep mode, a self-decay circuit depicted in Figure 15(c) generates a sleep pulse which periodically deactivates the entire SRAM except for the block that is being accessed. G1 in Figure 16, masks SLEEP when $\overline{BLOCK\_SEL}$=0 so the accessed block remains in active mode regardless of the sleep pulse. The SRAM block is activated ahead of time using the predecoder signals, and thus the only delay penalty comes from the bounce in VGND (<50mV) during the read operation which affects the read access time by less than 2 percent (Figure 20).

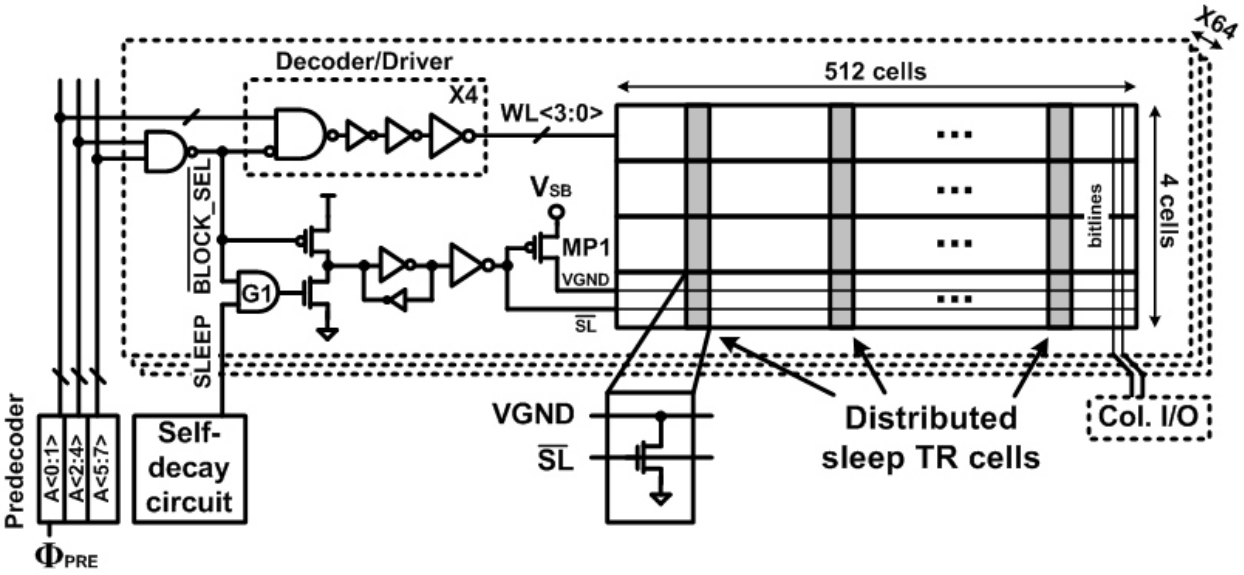
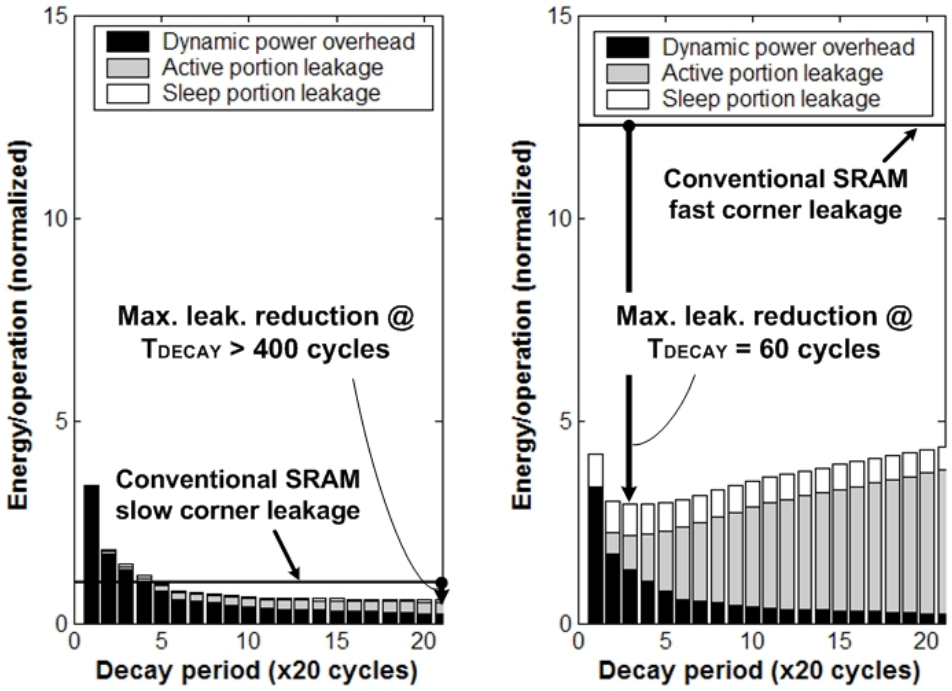


Figure 16. Organization of 16KB SRAM with self-decay circuit. Sleep transistors are distributed across the SRAM block to avoid current crowding or ground bounce issues.

Figure 17 shows decay period (T_{DECAY}) versus total leakage savings (overhead included) for a 64 KB L1 instruction cache at slow (+3σ V_t) and fast (-3σ V_t) process corners. SimpleScalar-3.0 [41] was used to run SPEC2000 benchmark applications for the cache access pattern and a predictive 70 nm technology [42] was used for the circuit parameters. Since the slow corner leakage is less than 1/10 of fast corner leakage, it is optimal to have a long T_{DECAY} (>400 cycles) that will seldom turn off the SRAM. The relatively large dynamic overhead energy will not be wasted for the small leakage savings. For maximum leakage savings in the fast corner dies however, the block must enter the sleep mode more frequently by having a shorter T_{DECAY} (~60 cycles) since the overhead power is relatively small compared to the leakage power. Depending on the amount of leakage with respect to the dynamic power overhead, there is a large difference in T_{DECAY} for maximum leakage savings. The designed self-decay circuit in Figure 15(c) is capable of tracking the process and temperature fluctuations during run-time to provide a near-optimal T_{DECAY} under varying conditions. To compare the effectiveness of the proposed self-decay scheme with prior-art, statistical studies were carried out based on measured leakage data, i.e., the spread of the original leakage distribution in Figure 18 is from experiment data. Again, the cache access pattern was acquired from SPEC2000 benchmark applications and the overhead energy and leakage in the active portion of the cache is also considered. A fixed decay scheme [37] whose decay period is optimized for high leakage conditions shows 52 percent reduction in average leakage. The proposed self-decay circuit that automatically tracks the near-optimal T_{DECAY} offers 27 percent lower leakage power (overhead included) compared to a fixed decay scheme.



(a) slow corner (b) fast corner

Figure 17. Relationship between decay period and leakage energy (overhead included) for (a) slow and (b) fast corner dies. Optimal T_{DECAY} for maximum leakage reduction depends on process corner. SimpleScalar-3.0 [20] was used to run SPEC2000 benchmark applications for the cache access pattern and predictive 70nm technology [21] was used for the circuit parameters.

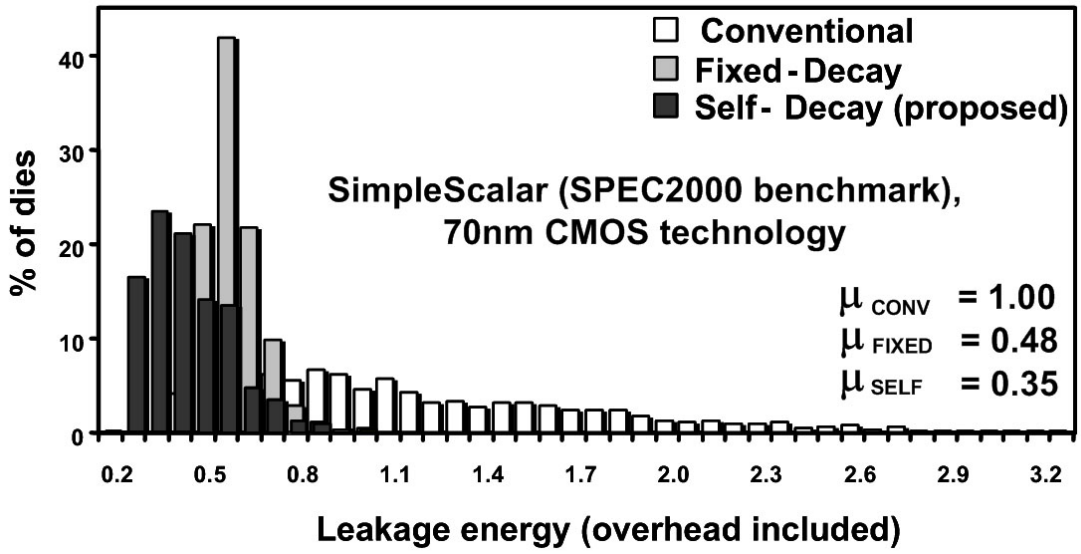
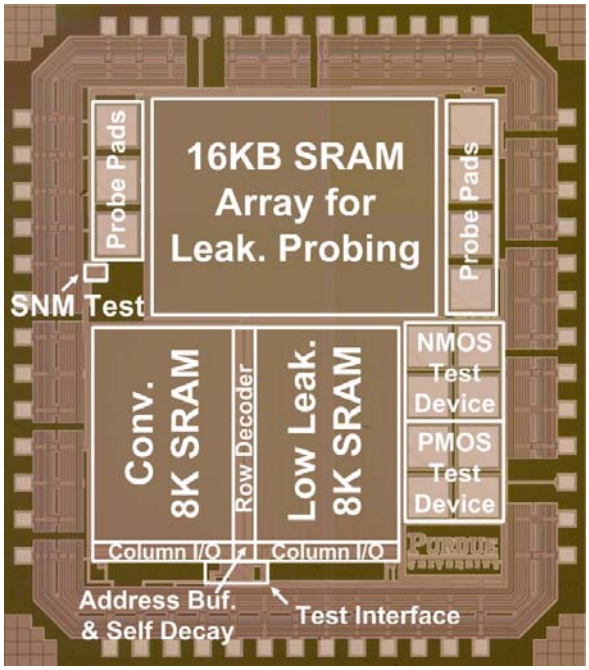


Figure 18. Statistical leakage reduction (overhead included) of proposed self-decay scheme compared to conventional and fixed decay scheme. Proposed self-decay scheme offers 27 percent lower leakage power compared to previous fixed decay scheme.

3.3 Experimental Results

A 16 KB SRAM testchip with the proposed self-decay scheme was fabricated in a 0.18 μm, 1.8 V, 6-metal CMOS technology. The threshold voltages of NMOS and PMOS were 0.53V and -0.53V, respectively. A microphotograph of the 6.94 mm^2 testchip is shown in Figure 19. A conventional 8 KB SRAM array without sleep transistors was also implemented to compare the SRAM access time. The read access cycle of the testchip was 984 MHz at 1.8 V (Figure 20) and the performance penalty of the proposed SRAM with sleep transistors was measured to be less than 2 percent. The active power consumption was 0.14 mW/MHz and the area overhead for the sleep transistors, self-decay circuit and additional peripheral circuits was 6 percent of total SRAM area. In order to accurately measure the reduction in SRAM leakage, a separate 16 KB array with probe pads was built (Figure 21). SRAM leakage components under different bias

conditions were measured from the array using a semiconductor parameter analyzer. Figure 22 shows the measured leakage data at 1.8 V and 45˚ C while sweeping VGND from 0 V to 1.2 V. Bitline leakage is virtually zero for a VGND higher than 0.2 V leaving just the cell leakage and the junction leakage (N-well-substrate, drain-body). Leakage savings becomes moderate for VGND above 0.9 V since the reduction in cell leakage levels off and junction leakage becomes a significant portion. The measured leakage current of the 16 KB SRAM at VGND = 0.9 V was 0.42 µA. This is only 5.8 percent compared to the conventional SRAM leakage. Leakage components illustrated in Figure 23 shows that both bitline and cell leakage are significantly reduced by raising VGND. This technique is also effective in reducing gate tunneling leakage since the voltage stress in each device is lowered. T_{DECAY} was measured from the fabricated self-decay circuit at different temperatures and the results are shown in Figure 24. An external V_{REF} of 0.9 V and V_{BIAS} of 0.2 V were applied (Figure 15(c)). The dynamic range of T_{DECAY} was 7.6X as the temperature is changed from 20˚ C (T_{DECAY} = 8.24 µs) to 75˚ C (T_{DECAY} = 1.25 µs). As leakage gets higher with temperature, T_{DECAY} is reduced which in turn makes the SRAM enter the sleep state more frequently. Vice versa, the longer T_{DECAY} at low temperatures reduces the number of transitions between active and sleep mode. V_{BIAS} and V_{REF} in Figure 15(c) can be changed to get a desired reference T_{DECAY}.



Technology	180nm 6-metal CMOS
Chip Size	3.3X2.9 mm^2 (1.6M transistors)
Supply Voltage	1.8V
Threshold Voltage	NMOS: 0.53V PMOS: -0.53V
Read Access Cycle	984MHz @ 1.8V, RT
Active Current	0.14mW/MHz @ 1.8V
Standby Current	7.27µA (16KB array)

Figure 19. Chip microphotograph and details of the 16KB SRAM testchip with self-decay based leakage reduction scheme.

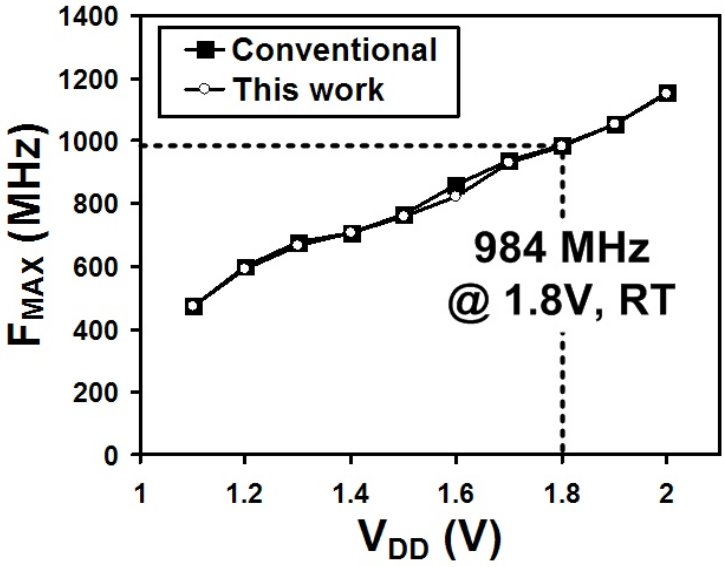


Figure 20. Measured 16 KB SRAM frequency versus V_{dd} with and without sleep transistor. Performance penalty due to the sleep transistor is less than two percent.

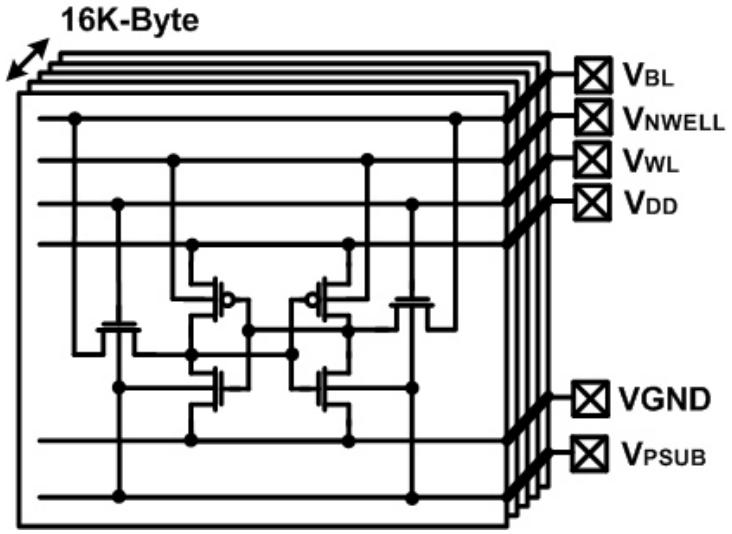


Figure 21. 16 KB SRAM array for leakage measurements. Connecting each terminal of SRAM cells together yields larger leakage current levels for effective probe measurements.

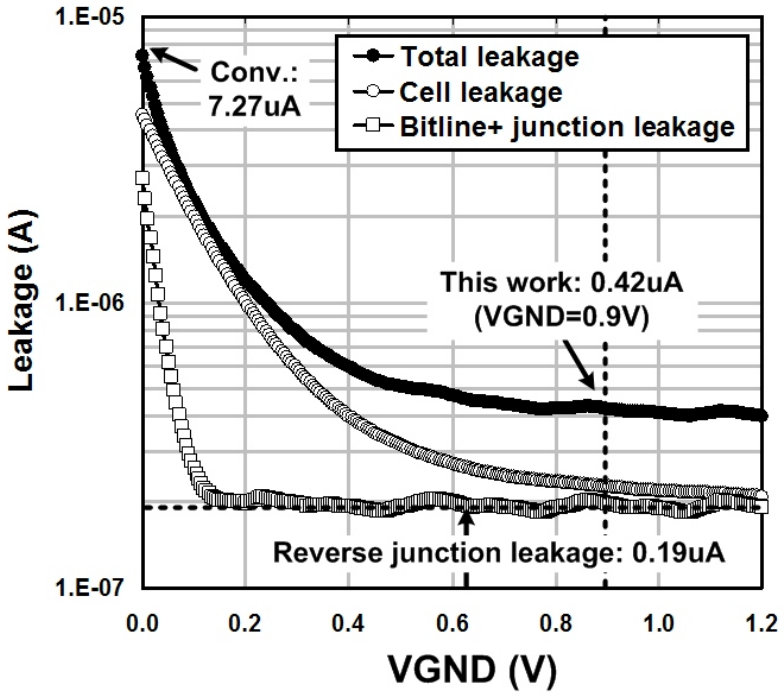


Figure 22. Leakage components versus virtual ground voltage (VGND) of 16 KB SRAM measured at 1.8 V and 45˚ C. Total leakage reaches minimum for VGND > 0.9 V.

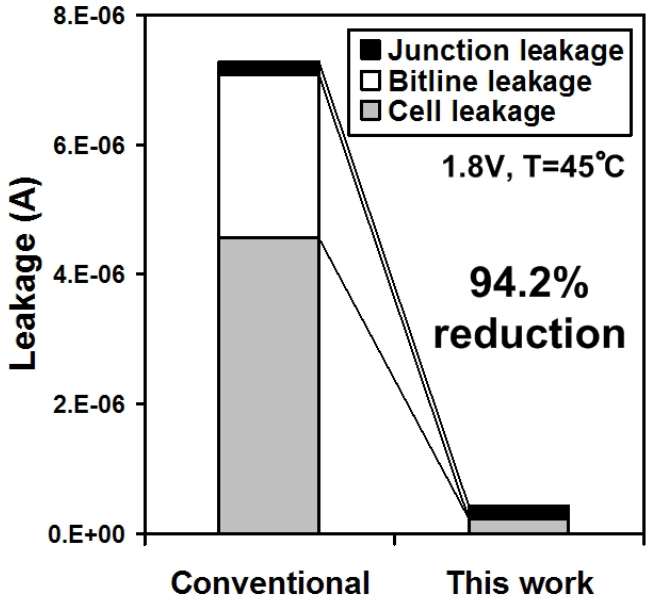


Figure 23. Leakage reduction of 16 KB SRAM measured at 1.8 V, 45˚ C. 94.2 percent SRAM cell leakage reduction is achieved using the proposed source-biased SRAM cell technique.

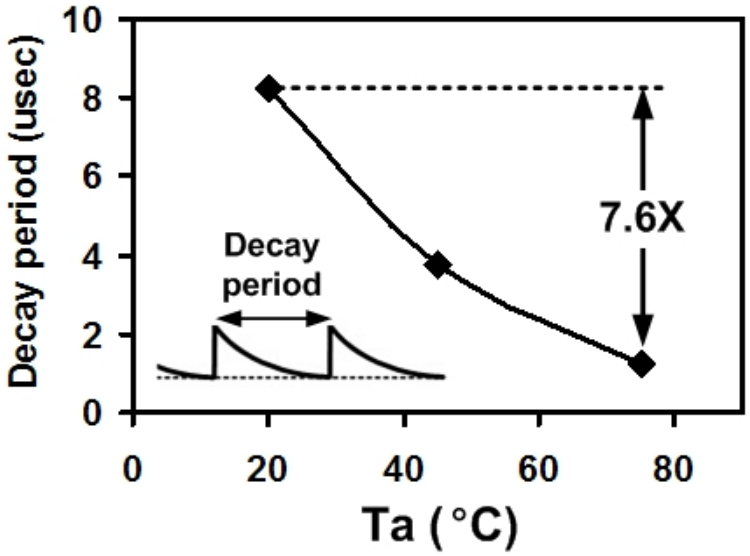


Figure 24. Self-decay period measured at different temperatures. As leakage gets higher with temperature, the decay period becomes shorter, cutting off the SRAM leakage more aggressively.

As supply voltage scales and process variation get worse, maintaining a reasonable SNM during read operation is one of the greatest technological challenges in high-density SRAM designs [43]. SNM of an SRAM cell is defined by Seevinck [21] as the minimum dc noise voltage necessary to change the state of a cell. The test circuit in Figure 25 was implemented in the testchip for studying the impact of sleep transistor on SNM. The sleep transistor width can be changed from $0.05W_N$ to $7.55W_N$ using digital control signals IN0-IN4. The measurement results in Figure 26 show that interestingly, the read SNM improves with a sleep transistor. For G_{SIZE} of 1.0 that gives two percent performance penalty in the designed 16 KB SRAM, the improvement in read SNM was 25 percent. Here, G_{SIZE} indicates the size of the sleep transistor normalized to the size of the NMOS driver in the 6T SRAM cell. To understand this counter-intuitive observation better, voltage transfer curves and VGND of a 6T SRAM cell with and without a sleep transistor is shown in Figure 27. The SNM can be visualized as the maximum width of the enclosed square in the superimposed voltage transfer curves of V_Q and V_{QB}. When a sleep transistor is used, VGND rises at the right end of the figure (V_Q = 1.8 V) causing V_{QB} to rise (denoted in (A) in Figure 27 which has a degrading effect on read SNM. However, the rise in VGND at the center of the figure (V_Q = 0.8 V) makes it harder for the SRAM to flip since the source voltage of the NMOS driver transistor also rises. This results in an increase in DC gain

(denoted in (B) in Figure 27) opening the transfer curve window and improving the overall read SNM by 25 percent. The reason why the effect of (B) prevails over (A) is because VGND always rises higher at the center (V_Q~0.8V, V_{QB}~1.8V) than at the right end (V_Q~1.8V, V_{QB}~0.2V) due to larger pull-up current. Finally, we verified the improvement in SNM for different SRAM cell structures to show that the principle holds general for different SRAM designs. Simulation results in Figure 28 show that for a predictive 70nm process, the improvement in SNM is 18-129mV depending on the beta ratios Q ($\beta_{access}/\beta_{PMOS}$) and R ($\beta_{driver}/\beta_{access}$).

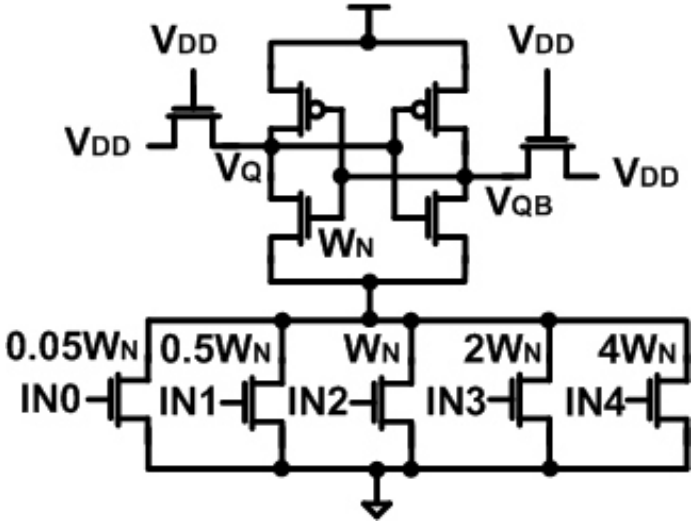


Figure 25. Test circuit with programmable sleep transistors for static noise margin measurements.

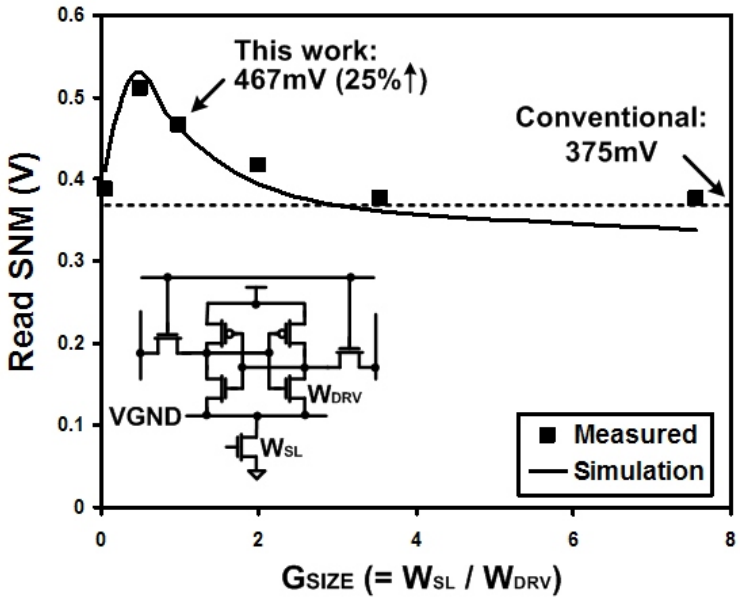


Figure 26. Static noise margin versus sleep transistor size measured at 1.8 V and room temperature. Sleep transistors for leakage reduction offers 25 percent higher SRAM read stability.

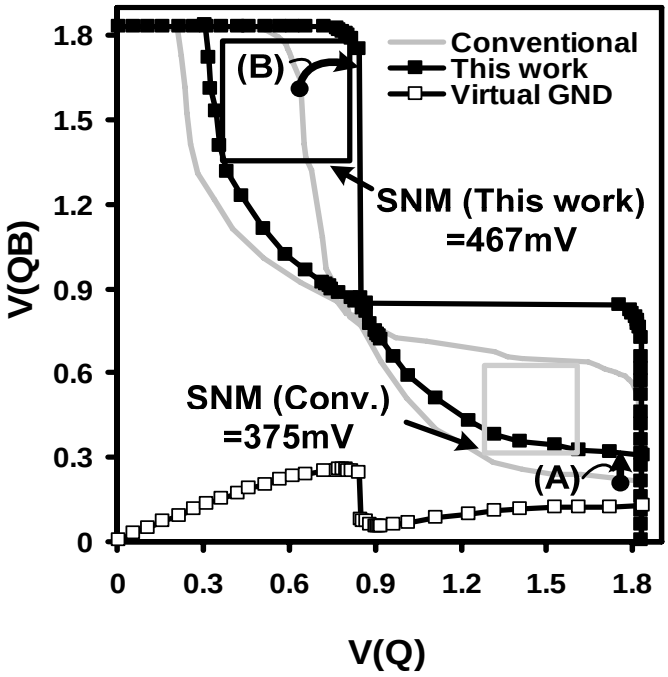


Figure 27. Measured SRAM butterfly curves and virtual ground voltage with and without sleep transistor (G_{SIZE}=1.0). VGND is highest at the center, increasing the DC gain (effect (B)) and improving SRAM cell static noise margin.

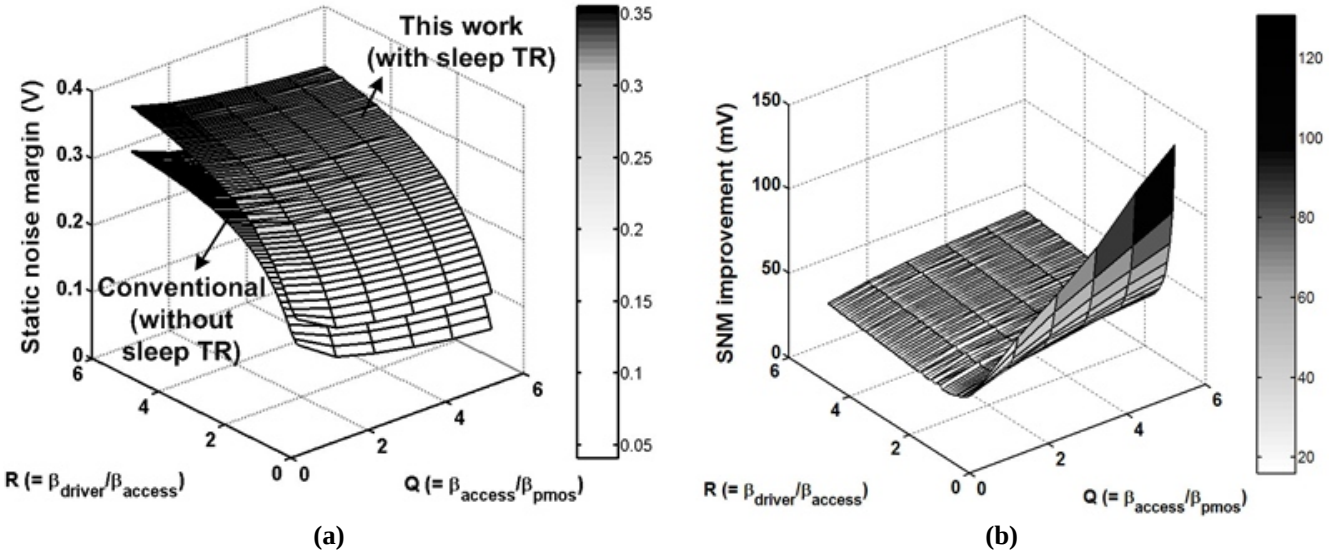
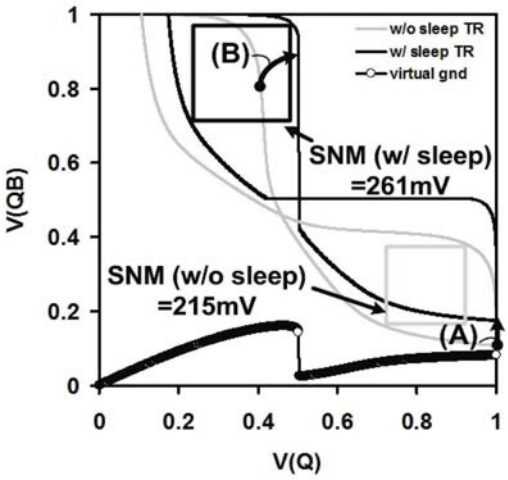
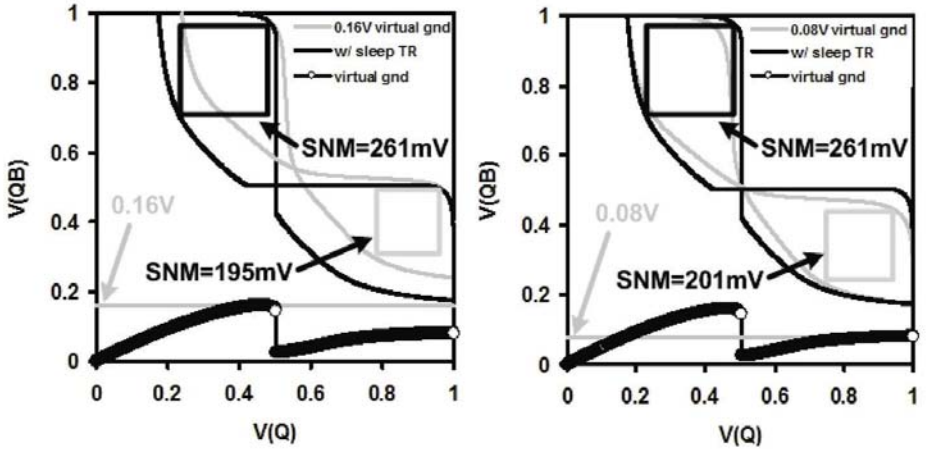


Figure 28. (a) Static noise margin with and without sleep transistor for different SRAM sizing.
(b) Improvement in SNM ranges from 18 to 129 mV (70 nm, V_{dd} = 1.0 V, RT, V_{tn} = 0.29 V, V_{tp} = -0.31 V).

To further investigate how the rise in VGND affects SNM, we compare the proposed SRAM having sleep transistors with the case where the VGND is fixed to a positive voltage (0.16 V and 0.08 V). 0.16 V corresponds to the maximum VGND at the center of Figure 27 and 0.08 V corresponds to the VGND at the right end of Figure 27. SNM of the conventional SRAM cell and proposed gated-ground SRAM cell are 215 mV and 261 mV, respectively. As the results in Figure 29 indicate, a constant VGND of 0.16 V degrades the SNM (215 mV→195 mV) since the effect of (A) becomes larger due to the higher VGND at the right end of the transfer curve plot. SNM is also reduced (215 mV→201 mV) when VGND is 0.08 V because the effect of (B) in Figure 27 is not as prominent as in the proposed SRAM. From these observations, we can conclude that the improvement in SNM attributes to the VGND profile in Figure 27 where (1) a peak value at the center increases effect (B), and (2) lower value at the right end suppresses effect (A).



(a)



(b)

Figure 29. (a) SRAM butterfly curves with and without a sleep transistor. (b) Comparison with constant virtual ground biases of 0.16 V and 0.08 V (70 nm, V_{dd} = 1.0 V, RT, V_{tn} = 0.29 V, V_{tp} = -0.31V). Static noise.

3.4 Conclusion

Reducing cache leakage in active mode is challenging because of the short time to control and the large dynamic power overhead required for the sleep and wake up process. The amount of leakage current that can saved varies by more than 100X due to PVT fluctuations, while the dynamic power overhead for applying the technique does not change much with PVT variations. As leakage variation becomes worse with technology scaling, effectiveness of an active leakage reduction scheme heavily depends on the amount of leakage with respect to the dynamic power overhead. In this report, we have proposed a self-decay scheme that can track the leakage of an SRAM memory and generate a sleep pulse which puts the SRAM into a sleep mode at an optimal rate. For high leakage conditions, the period of the sleep pulse is shortened for aggressive leakage reduction whereas at low leakage conditions, the period is longer so that the large dynamic power overhead is not wasted for marginal savings in leakage. A 16 KB SRAM testchip with the proposed scheme has been fabricated in a 0.18 µm CMOS technology. Measurement results show 94.2 percent cell leakage reduction at a performance penalty and area overhead of two percent and six percent, respectively. Designing robust SRAM cells is another important challenge circuit designers face for low voltage operation. To avoid the SRAM cell data from flipping during read operation, SNM must be maximized while keeping the cell size minimum. In this work we have also discovered that counter-intuitively, the proposed source-biased gated-ground SRAM cell has 25 percent higher SNM compared to conventional SRAM cells.

4. CONCLUSION

Previous low-leakage SRAM architectures faced inherent limitations. Problems they suffered included: (1) delay overhead, (2) reduced cell stability, (3) multiple cache hit times, and (4) insignificant leakage reduction. Even state-of-the-art techniques such as source biasing still pay large performance penalties and see significant soft error rates (SER). The present research addresses these problems using two general approachs: the *forward body-biased low-leakage SRAM cache* and the *low-leakage source-biased cache*.

The forward body-biased low-leakage SRAM cache approach starts from the simple idea of using super high Vt devices for low-leakage. Then we forward body-bias these devices to achieve high performance. We have endeavored to optimize the leakage/performance tradeoff at three different design levels: (1) device, (2) circuit and (3) architecture.

(1) At the device level, the super high Vt doping profile was optimized to improve the DIBL and Vt roll-off while suppressing the JBTBT leakage component.

(2) At the circuit level, transition latency associated with FBB was hidden by modifying the decoder circuit to give an early notice to the subarray that is to be accessed.

(3) At the architectural level, the general cache access pattern is exploited to lower the body-bias transition energy.

These combined device-circuit-architecture techniques achieve 64 percent total leakage reduction (overhead included) and 7.3 percent improved bit line delay over prior art for the forward body-biased low-leakage SRAM cache approach. Cell stability is not impacted.

The low-leakage source-biased cache approach addresses the challenges of reducing cache leakage in active mode. Control times are short, and large dynamic power is required for the sleep and wake up processes. The amount of leakage current that can be saved varies by more than 100X due to fluctuations in Process, Voltage and Temperature (PVT). Dynamic power for applying this technique, however, does not change much with PVT variations. Now with technology scaling, leakage variation becomes worse. Consequently, any successful active leakage reduction scheme must take into consideration the amount of leakage saved versus the dynamic power consumed.

We have proposed a self-decay scheme that can track the leakage of an SRAM memory and generate a sleep pulse which puts the SRAM into a sleep mode at an optimal rate. For high leakage conditions, the period of the sleep pulse is shortened for aggressive leakage reduction. For at low leakage conditions, the period is longer so that the large dynamic power overhead is not wasted for marginal savings in leakage. A 16 KB SRAM testchip with the proposed scheme has been fabricated in a 0.18 μm CMOS technology. Measurement results show 94.2 percent cell leakage reduction at a performance penalty and area overhead of two percent and six percent, respectively.

Designing robust SRAM cells is another important challenge circuit designers face for low voltage operation. To avoid the SRAM cell data from flipping during read operation, SNM must be maximized while keeping the cell size minimum. In this work we have also discovered that counter-intuitively, the proposed source-biased gated-ground SRAM cell has 25 percent higher SNM compared to conventional SRAM cells.

REFERENCES

[1] A. Keshavarzi, S. Narendra, B. Bloechel, et al., "Forward body bias for microprocessors in 130 nm technology generation and beyond", Symposium on VLSI Circuits, pp. 312-315, 2002

[2] M. Miyazaki, G. Ono, T. Hattori, et al., "A 1000-MIPS/W microprocessor using speed-adaptive threshold-voltage CMOS with forward bias", International Solid-State Circuits Conference, pp. 420-421, 2000

[3] C. H. Kim, J. Kim, S. Mukhopadhyay, and K. Roy, "A forward body-biased low-leakage SRAM cache: device and architecture considerations", International Symposium on Low Power Electronics and Design, pp. 6-9, Aug. 2003

[4] S. Narendra, M. Haycock, V. Govindarajulu, et al., "1.1 V, 1 GHz communications router with on-chip body bias in 150 nm CMOS", International Solid-State Circuits Conference, pp. 270-271, 2002

[5] Y. C. Yeo, Q. Lu, W. Lee, et al., "Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric", IEEE Electron Device Letters, vol. 21, no. 11, pp. 540-542, 2000

[6] A. Agarawal, H. Li, and K. Roy, "DRG-cache: a data retention gated-ground cache for low power", Design Automation Conference, pp. 473-478, 2002

[7] H. Yamauchi, T. Iwata, H. Akamatsu, et al., "A 0.8V/100MHz/sub-5mW-operated mega-bit SRAM cell architecture with charge-recycle offset-source driving (OSD) scheme", Symposium on VLSI Circuits, pp. 126-127, 1996

[8] A. J. Bhavnagarwala, A. Kapoor, and J. D. Meindl, "Dynamic-threshold CMOS SRAMs for fast, portable applications", ASIC/SOC Conference, pp. 359-363, 2000

[9] K. Osada, Y. Saitoh, E. Ibe, et al., "16.7 fA/cell tunnel-leakage-suppressed 16 Mb SRAM for handling cosmic-ray-induced multi-errors", International Solid-State Circuits Conference, pp. 302-303, 2003

[10] H. Kawaguchi, Y. Itaka, and T. Sakurai, "Dynamic leakage cut-off scheme for low-voltage SRAM's", Symposium on VLSI Circuits, pp. 140-141, 1998

[11] C. H. Kim and K. Roy, "Dynamic Vt SRAM: a leakage tolerant cache memory for low voltage microprocessors", International Symposium on Low Power Electronics and Design, pp. 251-254, 2002

[12] K. Flautner, N. S. Kim, S. Martin, et al., "Drowsy caches: simple techniques for reducing leakage power", International Symposium on Computer Architecture, pp. 148-157, 2002

[13] S. Heo, K. Barr, M. Hampton, et al., "Dynamic fine-grain leakage reduction using leakage-biased bitlines", International Symposium on Computer Architecture, pp. 137-147, 2002

[14] K. Itoh, A. R. Fridi, A. Bellaouar and M. I. Elmasry, "A deep sub-V, single power-supply SRAM cell with multi-Vt, boosted storage node and dynamic load", Symposium on VLSI Circuits, pp. 132-133, 1996

[15] Y. Taur and E. Nowak, "CMOS devices below 0.1μm: how high will performance go?", International Electron Devices Meeting, pp. 215-218, 1997

[16] Y. Taur, C. H. Wann, and D. J. Frank, "25nm CMOS design considerations", International Electron Devices Meeting, pp. 789-792, 1998

[17] MIT well tempered MOSFET model, Available: http://www-mtl.mit.edu/Well/

[18] B. Yu, H. Wang, O. Milic, et al., "50 nm gate-length CMOS transistor with super-halo: design, process, and reliability", International Electron Devices Meeting, pp. 653-656, 1999

[19] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Nov. 1998

[20] S. Kaxiras, Z. Hu, and M. Martonosi, "Cache decay: exploiting generational behavior to reduce cache leakage power", International Symposium on Computer Architecture, pp. 240-251, 2001

[21] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells", IEEE Journal of Solid-State Circuits, Vol. 22, Issue 5, pp. 748 -754, Oct. 1987.

[22] (2003) International Technology Roadmap for Semiconductors. [Online]. Available: http://pulic.itrs.net

[23] G. Sery, S. Borkar, and V. De, "Life is CMOS: why chase the life after?", Design Automation Conference, pp. 78-83, June, 2002.

[24] H. Wong, D.J. Frank, and P.M. Solomon, "Device design considerations for double-gate, ground-plane, and single-gated ultra-thin SOI MOSFET's at the 25nm channel length generation", International Electron Devices Meeting, pp. 407-410, 1998.

[25] C.T. Chuang, K. Bernstein, R.V. Joshi, R. Puri, K. Kim, E.J. Nowak et al., "Scaling planar silicon devices", IEEE Circuits and Devices Magazine, Vol. 20 , Issue 1, pp. 6–19, Jan.-Feb. 2004.

[26] E.J. Nowak, I. Aller, T. Ludwig, K. Kim, R.V. Joshi, C.T. Chuang, et al., "Turning silicon on its edge [double gate CMOS/FinFET technology]", IEEE Circuits and Devices Magazine, Vol. 20, Issue 1, pp. 20-31, Jan.-Feb. 2004.

[27] L. Wei, Z. Chen, K. Roy, M.C. Johnson, Y. Ye, and V. De, "Design and optimization of dual-threshold circuits for low-voltage low-power applications", Vol. 7, Issue 1, pp. 16-24, Mar. 1999.

[28] M. C. Johnson, D. Somasekhar, and K. Roy, "Leakage control with efficient use of transistor stacks in single threshold CMOS", Design Automation Conference, pp. 442-445, Jun. 1999.

[29] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V power supply high-speed digital circuits technology with multithreshold-voltage CMOS", Vol. 30, Issue 8, pp. 847-854, Aug. 1995.

[30] J. Kao, A. Chandrakasan, and D. Antoniadis, "Transistor sizing issues and tool for multi-threshold CMOS technology", Design Automation Conference, pp. 409-414, Jun. 1997.

[31] H. Kawaguchi, K. Nose, and T. Sakurai, "A CMOS scheme for 0.5 V supply voltage with pico-ampere standby current", International Solid-State Circuits Conference, pp. 192-193, Feb. 1998.

[32] T. Kuroda, T. Fujita, S. Mita, T. Nagamuta, S. Yoshioka, F. Sano, et al., "A 0.9V 150MHz 10mW 4mm^2 2-D discret cosine transform core processor with variable-threshold-voltage scheme", International Solid-State Circuits Conference, pp. 166-167, Feb. 1996.

[33] S. Narendra, M. Haycock, V. Govindarajulu, V. Erraguntla, H. Wilson, S. Vangal et al., "1.1V 1GHz communication router with on-chip body bias in 150nm CMOS", International Solid-State Circuits Conference, pp. 270-271, Feb. 2002.

[34] C.H. Kim, J. Kim, S. Mukhopadhyay, and K. Roy, "A forward body-biased low-leakage SRAM cache: device and architecture considerations", International Symposium on Low Power Electronics and Design, pp. 6-9, Aug. 2003.

[35] H. Mizuno, K. Ishibashi, T. Shimura, T. Hattori, S. Narita, K. Shiozawa et al., “An 18-µA standby current 1.8-V, 200-MHz microprocessor with self-substrate-biased data-retention mode”, IEEE Journal of Solid-State Circuits, Vol. 34, Issue 11, pp. 1492-1500, Nov. 1999.
[36] S. Yang, M. Powell, B. Falsafi, and T.N. Vijaykumar, “Exploiting choice in resizable cache design to optimize deep-submicron processor energy-delay”, International Symposium on High-Performance Computer Architecture, pp. 147-158, Feb. 2002.
[37] S. Kaxiras, Z. Hu, and M. Martonosi, “Cache Decay: Exploiting Generational Behavior to Reduce Cache Leakage Power”, International Symposium on Computer Architecture, pp. 240-251, Jun. 2001.
[38] W. Zhang, J.S. Hu, V. Degalahal, M. Kandemir, N. Vijaykrishnan, M.J. Irwin, “Compiler-directed instruction cache leakage optimization”, International Symposium on Microarchitecture, pp. 208-218, Nov. 2002.
[39] A. Agarwal, H. Li, and K. Roy, “A single V_t low-leakage gated-ground cache for deep submicron”, IEEE Journal of Solid-State Circuits, Vol. 38, Issue 2, pp. 319-328, Feb. 2003.
[40] A. Agarwal and K. Roy, “A noise tolerant cache design to reduce gate and sub-threshold leakage in the nanometer regime”, International Symposium on Low Power Electronics and Design, pp. 18-21, Aug. 2003.
[41] 2000) SimpleScalar-3.0 Architecture Simulator. [Online]. Available: http://www.simplescalar.com
[42] (2002) Berkeley Predictive Technology Model. [Online]. Available: http://www-device.eecs.berkeley.edu/~ptm/
[43] K. Zhang, U. Bhattacharya, Z. Chen, et al., “SRAM design on 65nm CMOS technology with integrated leakage reduction scheme”, VLSI Circuits Symposium, pp. 294-295, Jun. 2004.

LIST OF ACRONYMS AND ABBREVIATIONS

ACRONYM	DESCRIPTION
6M CMOS	Six-Metal Complementary Metal-Oxide-Semiconductor
6T SRAM	Six Transitor Static Random Access Memory
BSIM4	An electronic design simulation tool
CMOS	Complementary Metal-Oxide-Semiconductor
DIBL	Drain Induced Barrier Lowering
DVSRAM	Dynamic V_{DD} SRAM
FBB	Forward Body-Biased
FBSRAM	Forward Body-biased Static Random Access Memory
GIDL	Gate Induced Drain Leakage
HSPICE	An electronic design simulation tool
HVB	Hole tunneling from the Valence Band
JBTBT	Junction Band-To-Band Tunneling
LEFF	Effective channel Length
MEDICI	An electronic design simulation tool
MOSFET	Metal-Oxide-Semiconductor Field Effect Transistor
MSB	Most Significant Bit
NMOS	N-channel Metal-Oxide-Semiconductor
PAC/C	Power-Aware Computing and Communications
PMOS	P-channel Metal-Oxide-Semiconductor
PVT	Process, Voltage and Temperature
RBB	Reverse Body-Bias
RC	Resistor-Capacitor
SBSRAM	Source-Biased Static Random Access Memory
SER	Soft Error Rate
SNM	Static Noise Margin
SoC	System on Chip
SPEC2000	A computer performance test suite
SRAM	Static Random Access Memory
VLIW	Very Long Instruction Word
ZBB	Zero Body-Bias