A Noise Tolerant Cache Design to Reduce Gate and Sub-threshold Leakage in the Nanometer Regime

Amit Agarwal, and Kaushik Roy
School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47906, USA
<amita, kaushik> @ ecn.purdue.edu

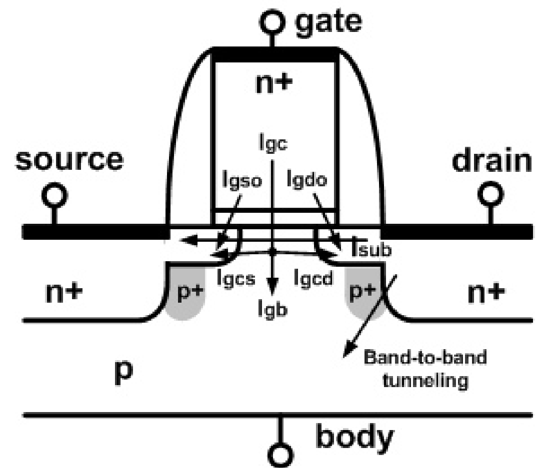
ABSTRACT
Scaling devices while maintaining reasonable short channel immunity requires gate oxide thickness of less than 20A^o for CMOS devices beyond the 70nm technology node. Low oxide thickness gives rise to considerable direct tunneling current (gate leakage). Power dissipation in large caches is dominated by the gate and sub-threshold leakage current. This paper proposes a novel cache that has high noise immunity with improved leakage power. For every bank of SRAM cells, this technique requires an extra diode in parallel with a gated-ground transistor connected between the source of NMOS transistors and ground in SRAM cells. The row decoder itself can be used to control the extra gated-Ground transistor. Our simulation results on 70nm process (Berkeley Predictive Technology Model augmented with our gate leakage model) show 39.2% reduction in consumed energy (leakage + dynamic) in L1 cache and 59.4% reduction in L2 cache energy with less than 2.5% impact on execution time. The technique is applicable to data and instruction caches as well as different levels of cache hierarchy such as the L1, L2, or L3 caches.

Categories and Subject Descripters
B.3.2 [**Memory Structure**]: Design Styles --- *Cache memories;* B.3.1 [**Memory Structure**]: Semiconductor Memories --- *Static memory (SRAM)*; B.7.1 [**Integrated Circuits**]: Types and Design Styles --- *Memory technology.*

General Terms: Design, Performance and Experimentation.

Keywords: SRAM, gate leakage, diode, low leakage cache.

1. INTRODUCTION
To achieve high integration density and high performance semiconductor devices are aggressively scaled in each technology generation. This requires use of ultra-thin gate oxide to maintain reasonable short-channel effect [1]. The International Technology Roadmap for Semiconductors (ITRS) predicts the gate oxide thickness to be 1.1–1.6nm for sub-70nm CMOS [2]. This low oxide thickness, coupled with high electric field results in considerable direct tunneling current [3]. For CMOS devices with larger oxide thickness, major leakage mechanism is sub-threshold current, which increases due to short channel effect. However, in the ultra-thin gate oxide regime, gate tunneling current becomes appreciable and dominates the total "off" state leakage current of the transistor along with sub-threshold leakage [4]. Hence, circuit techniques used to control sub-threshold leakage needs to be reinvestigated to evaluate their effectiveness in improving overall leakage current.



Many design techniques [5-7] have been proposed to reduce sub-threshold leakage in cache. These design techniques suffer from either low noise immunity or do not deal with gate leakage. The contributions of the paper are as follows:

- A novel design is investigated to improve both sub-threshold and gate leakage in nanometer regime.
- This design has high noise margin.

To observe the effect of gate direct tunneling in the behavior of the transistor, NFET and PFET were designed by modeling gate leakage as current sources in BPTM 70nm technology file. Gate leakage was modeled using BSIM4 equation [8]. To verify the above model, an NFET and PFET devices were designed using the doping profiles given in [9] and the design guidelines given in 2001 ITRS Roadmap [2]. These devices were then simulated using TAURUS device simulator [10] and compared with the above model.

2. LEAKAGE COMPONENTS IN CACHE
There are two dominant components of leakage in MOSFETs in the nanometer regime, 1) Sub-threshold leakage, which is the leakage current from drain to source (I_{sub}, Figure 1) and 2) Gate direct tunneling current (gate leakage), which is due to the tunneling of electron (or hole) from the bulk silicon through the gate oxide potential barrier into the gate. A third component, band-to-band tunneling at the source and drain junctions is also important, when channel engineering such as "halo" doping are used in scaled MOSFETs [3]. However, this component can be reduced if SOI technology or other doping profiles are used.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ISLPED'03, August 25–27, 2003, Seoul, Korea.
Copyright 2003 ACM 1-58113-682-X/03/0008…$5.00.

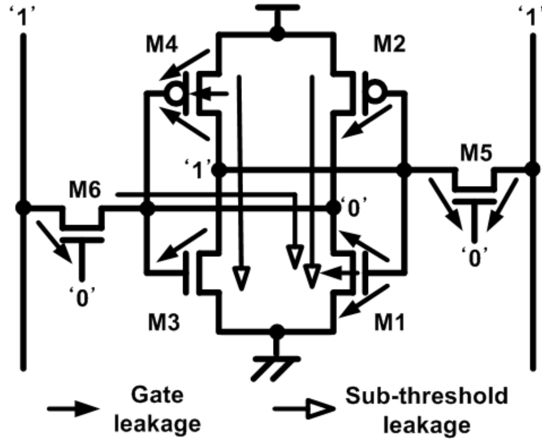


Figure 2. Dominant leakage components in a 6T SRAM.

The direct tunneling current increases exponentially with decrease in the oxide thickness and increase in voltage across oxide. Figure 1 describes the various components of gate tunneling in a scaled NMOSFET [11]. The gate current (I_g) can be divided into three major components:

(1) Gate-to-Source ($I_{gs} = I_{gso} + I_{gcs}$) ;
(2) Gate-to-Drain ($I_{gd} = I_{gdo} + I_{gcd}$);
(3) Gate-to-Substrate (I_{gb}).

Depending on the biasing condition, I_{gs} (if $|V_{gs}|$ is high) or I_{gd} (if $|V_{gd}|$ is high) dominates the total gate leakage current. Figure 2 shows the dominant leakage components in cache while it is in standby mode (not accessed). Gate leakage components of a transistor depend on the voltage across its terminal. The leakage through M5 and M6 depends on the voltage at which bit-lines are pre-charged. It is shown that gate leakage through PMOS is smaller than the NMOS [12]. The dominant component of gate leakage is through M1 and that includes all the three types of gate leakage current as described above. The sub-threshold leakage depends on the number of "off" transistors in the leakage path [13]. Higher the number of "off" transistor in a path, lower the sub-threshold leakage is through that path. In an SRAM cell, there are three sub-threshold leakage paths that have only one "off" transistor. The leakages through these paths govern the sub-threshold leakage of a cache.

3. A NOISE TOLERANT LOW LEAKAGE CACHE

In DRG-Cache [7] an extra NMOS transistor (gated-ground transistor) is introduced between source of NMOS transistors and ground to improve sub-threshold leakage. However this extra transistor increases the resistance of pull-down path. Furthermore, when the gated-ground transistor is turned off in the standby mode, the node storing '0' and the virtual ground are floating and get strapped to a positive voltage by the weak leakage currents. This makes DRG-cache more susceptible to noise sources (e.g. from adjacent lines) and reduces the soft error immunity. To mitigate the coupling noise and soft error problem in gated-ground cache [7], we utilize the idea of putting a diode in parallel with the gated-ground transistor (Figure 3). Putting a diode in parallel with gated-ground transistor straps the virtual ground (and hence node storing '0') to a fixed small voltage V_d when the gated-ground transistor is off. This voltage depends on the size and V_{th} of the MOS diode.

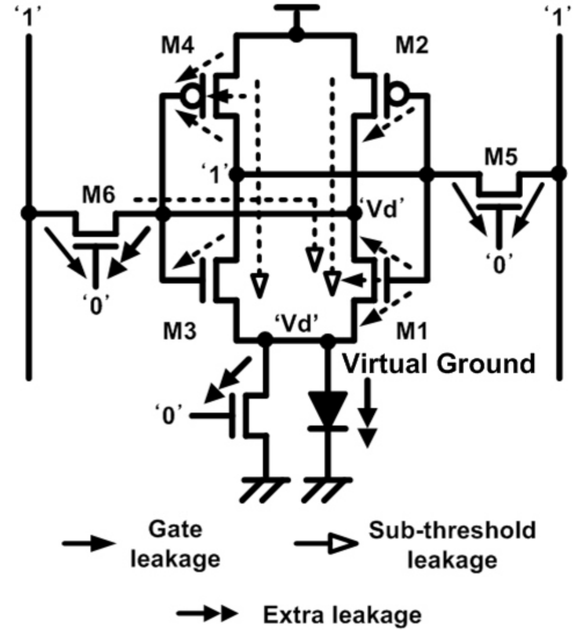


Figure 3. Sub-threshold leakage and gate leakage reduction using diode footed cache. Doted lines represent the saved leakage component.

Turning on the gated-ground transistor restore the data at the node storing '0' back to zero.

Conventional SRAM stores the data as long as power supply is on. This is because the cell storage nodes at '0' and '1' are firmly strapped to power rails through conducting devices (by a pull-down NFET in one inverter and a pull-up PFET in the other inverter). Turning 'off' the gated-ground transistor in diode-footed cache nicely cuts-off the leakage paths to the ground. However, it also straps the voltage at virtual ground (and hence voltage at node storing '0') to a fixed voltage. Since virtual ground and node storing '0' is strapped to a fixed voltage by strong diode 'ON' current, data is retained. If, due to some noise source, voltage at the virtual ground node increases, the diode turns on strongly and keeps the virtual ground strapped to the intrinsic diode voltage (V_d). This voltage should not exceed the trip point of cell to retain the data. Hence, the diode size and V_{th} needs to be carefully designed to get desirable V_d for maximum savings without affecting the data retention.

3.1 Architecture and Layout

To amortize the area overhead the gated-ground transistor and diode are shared by a bank of SRAM cells. They are placed along the length of the SRAM cells with proper pitch matching. The metal line, which is used as a ground line in a conventional cache, is connected to the drain of the gated-ground transistor and acts as a virtual ground line. Substrate is still connected to ground by a separate metal line. The gated-ground transistor is controlled by the gated-ground control signal, which is generated by the row decoder logic when the bank address gets decoded. By the time word-line signal reaches the pass transistors of the row, which is being read or written, gated ground transistor pulls the virtual ground back to '0'. It is important to note that the diode-footed cache core is fully compatible with current cache design. The area overhead for present design is around 7-8%.

Because the size of the diode plays a major role in the data retention capability of the diode-footed cache, and also affects the power and performance (section 4), the diode must be carefully sized with respect to the SRAM cell transistors. Also the size of gated-ground transistor must be made large enough to sink the current flowing through the SRAM cells during a read/write operation in the active mode. However, too large a diode and a gated-ground transistor may reduce the virtual ground voltage considerably, thereby diminishing the energy savings. Moreover, large transistors also increase the area overhead. To get a desirable V_d for maximum leakage savings without affecting the data retention, the diode is designed using a high V_{th} transistor. The size of the diode is equal to the gated-ground transistor, which is equal to the size of M1 or M3, (Figure 3) with a V_{th} = 0.35V (cell transistors V_{th} = 0.2V).

3.2 Static Noise Margin

Static noise margin (SNM) of a CMOS SRAM cell is defined in [14] as the minimum dc noise voltage necessary to change the state of a cell. The conventional and diode-footed cache static transfer characteristics during a read access are simulated for 70nm technology. Figure 4 shows the superimposed static transfer characteristics of two inverters in a single cell for both conventional and diode-footed cache. The SNM is the noise voltage equal to the maximum width of the enclosed square in the superimposed voltage transfer curves of V(Q) and V(Qbar). Figure 4 shows that the SNM of diode-footed cache is comparable to the conventional cache. This result shows that diode-footed cache does not suffer from SNM problem.

3.3 Leakage Savings (Sub-threshold and Gate Leakage)

Since the diode-footed cache raises the virtual ground voltage to V_d, it produces an effect similar to stacking effect (reverse biasing the source). This positive potential at the intermediate nodes has three effects: 1) Gate to source voltage becomes negative. 2) Negative body to source potential causes more body effect resulting in increased threshold voltage. 3) Drain to source potential decreases, resulting in less Drain Induced Barrier Lowering (DIBL). These three effects combined together results in low sub-threshold leakage [7,13].

Analyzing for gate leakage shows that this design is efficient in reducing gate leakage too. Previous section shows that turning off the gated-Ground transistor makes the node storing ‘0’ to get strapped at a small voltage, V_d. It acts as an increased ground potential (NMOS substrate is still connected to ground). The gate leakage through any transistor depends on the voltage difference across gate-drain, gate-source and gate-body. Increasing the ground potential of internal nodes reduces most of the above voltage differences and hence gate leakage across transistors M1, M2, M3 and M4, as shown by dotted lines in Figure 3. It also introduces some extra leakage due to the gated-ground transistor,

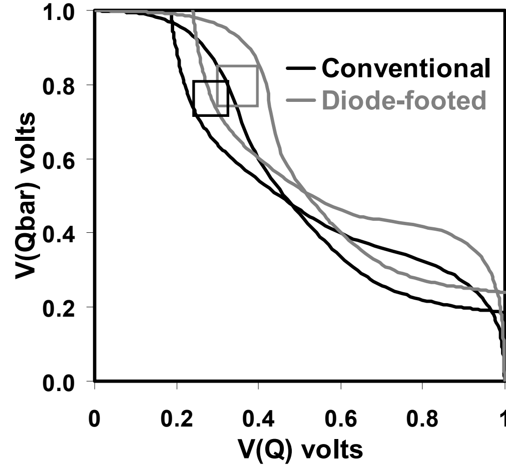


Figure 4. Static noise margin of diode-footed cache.

the diode and the increased virtual ground voltage. These components are negligible compared to the leakage savings in the other components. Extra leakage introduced by the change in virtual ground voltage (0 to V_d) is much smaller than the reduction in gate leakage by the lowered voltage difference (V_{DD} to V_{DD}- V_d) in the dominant leakage terminals.

4. RESULTS

The diode-footed cache looses performance against conventional cache due to the increase in bit-line/sense amplifier delays, causing the overall performance degradation to be 2.5 % compared to the low V_{th} conventional cache.

Table 1 shows gate (I_{GATE}) and sub-threshold (I_{SUB}) leakage currents flowing in the diode-footed SRAM cell compared to conventional cache in the standby mode. It also shows the extra leakage component introduced in the proposed design. The proposed diode-footed cache reduces almost every leakage component present in conventional cache. The leakage overhead due to extra transistor and raised ground voltage is negligible. Here bit-lines are at V_{DD}. Lowering the bit-line voltage to a smaller potential reduces the gate leakage components (I_{gd5}, I_{gd6}) further. It also reduces the sub-threshold leakage through M6. To consider the worst case, bit-lines were assumed to be pre-charged to $V_{DD.}$ Diode-footed cache reduces 65.8% sub-threshold leakage and 44.1% gate leakage with respect to conventional cache. Table 2 shows the tradeoff between leakage improvements and performance of the cell. The choice of high V_{th} decides the intrinsic diode voltage which in turn defines virtual ground voltage and hence, leakage improvement. However, too high a V_{th} may decrease the drive current of diode and hence, the performance of the cell. Decreasing the size of the diode does not have much effect on the diode intrinsic voltage (V_d) (hence, leakage savings). However, it degrades the performance.

Table 1. Dominant leakage components in cache (Didoe V_{th} = 0.35V , Size M =1.0)

Cache	I_{SUB} (nA)	I_{GATE} (nA)													
		$\lvert I_{gd1}\rvert$	$\lvert I_{gs1}\rvert$	$\lvert I_{gb1}\rvert$	$\lvert I_{gd2}\rvert$	$\lvert I_{gd3}\rvert$	$\lvert I_{gd4}\rvert$	$\lvert I_{gs4}\rvert$	$\lvert I_{gb4}\rvert$	$\lvert I_{gd5}\rvert$	$\lvert I_{gs5}\rvert$	$\lvert I_{gd6}\rvert$	$\lvert I_{gs6}\rvert$	$\lvert I_{gd7}\rvert$	$\lvert I_{gs}\rvert_{diode}$
Conventional	254	10.7	10.7	0.01	0.13	7.03	0.10	0.10	0.01	4.98	4.98	4.98	0	NA	NA
Diode-Footed	86.8	3.36	3.36	0	0.05	2.36	0.04	0.04	0	4.98	4.98	4.98	0.02	0.13	0.13

Table 2. Power vs Performance tradeoff ($Sub\text{-}Leak_{Conv}$ = 254nA, $GateLeak_{Conv}$ = 43.7nA and Normal transistor V_{th} = 0.2V)

Diode Size (M)	Diode V_{th} (V)	Diode V_d (V)	Diode Performance Loss (%)	Sub Leak (nA)	Gate Leak (nA)
1.0	0.35	0.200	2.50	86.8	24.4
0.5	0.35	0.209	3.75	80.8	23.4
1.0	0.45	0.245	4.00	64.6	21.8
0.5	0.45	0.255	4.75	61.4	21.3

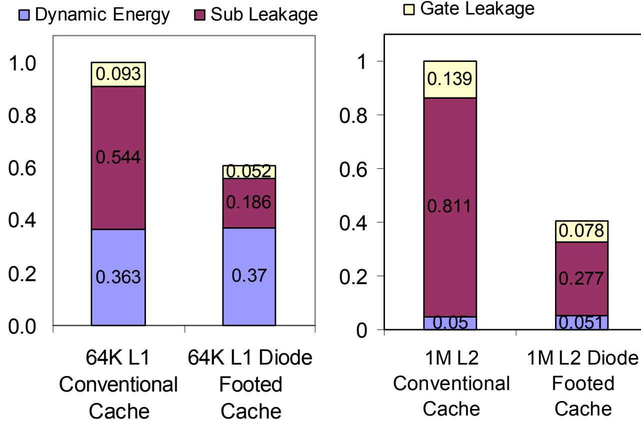


Figure 5. Energy savings in 64K L1 and 1MB L2 cache.

Leakage energy consumption in a cache depends on its utilization. More the cache is in the idle mode, more is the ratio of leakage energy over dynamic energy. In the context of aggressive modern out-of-order microprocessors that exploit instruction level parallelism with the help of dynamic scheduling, branch prediction, and speculative execution, it can be assumed that L1 cache is accessed in each and every cycle (conservative assumption). Simplescalar [15] is used for getting the information about the L2 cache utilization factor and on average it was found to be 20%.

Figure 5 shows the contribution of leakage and dynamic energy in L1 and L2 caches based on measured utilization factor. Assuming that total energy of conventional cache is 1 unit, the leakage and dynamic energy consumption of diode-footed cache is normalized to the conventional cache. For example, for 70nm L1 cache, sub-threshold leakage contribution of diode-footed cache is only 18.6% of the total conventional L1 cache energy. The graph indicates that in conventional cache design, the sub-threshold leakage energy accounts for 54.4% in L1cache and 81.1% in L2 cache in 70nm technology. The gate leakage contributes 9.3% in L1 cache and 13.9% in L2 cache of the total cache energy consumption. The diode-footed cache improves sub-threshold leakage by 65.8% and gate leakage by 44.1% compared to conventional cache in 70nm process. The energy overhead associated with decoder results in 1.95% and 2.34% increase in total dynamic energy of L1 and L2 caches, respectively, compared to conventional cache. The overall energy reduction achieved by the diode-footed cache is as much as 39.2% in L1 cache and 59.4% in L2 Cache.

5. CONCLUSIONS

In current deep submicron devices with low threshold voltage and oxide thickness, sub-threshold and gate leakage have become the dominant sources of leakage and are expected to increase with technology scaling. This paper proposed a circuit mechanism that is noise tolerant and effective in improving both gate and sub-threshold leakage in cache memories. The diode-footed cache design is fully compatible with current cache architecture and applicable to each level of cache hierarchy.

Acknowledgement: This research was funded in part by DARPA PACC program, MARCO Giga-scale system research center, semiconductor Research Corporation, Intel, and IBM.

6. REFERENCES:

[1] K. Roy and S. C. Prasad, Low-Power CMOS VLSI Circuit Design, New York, USA: Wiley Interscience Publications, 2000, ch. 5, pp. 224-226.

[2] http://public.itrs.net/Files/2001ITRS/Home.htm

[3] Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, New York, USA: Cambridge University Press, 1998, ch.2, pp. 94-97.

[4] N. Yang, W. K. Henson, and J. Wortman, "A comparative study of gate direct tunneling and drain leakage currents in N-MOSFETs with sub-2100-nm gate oxides," IEEE Transaction on . Electron Devices, vol. 47, pp. 1636–1644, Aug. 2000.

[5] H. Yamauchi 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.

[6] K. Kumagai et al., "A novel powering-down scheme for low Vt CMOS circuits", Symp on VLSI Circuits, pp. 44-45, 1998.

[7] A. Agarwal, H. Li, and K. Roy, "A Single-Vt Low-Leakage Gated-Ground Cache for Deep Submicron" IEEE Journal of Solid-State Circuits, 2003.

[8] http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html

[9] http://www-mtl.mit.edu/Well/

[10] TAURUS: Three-Dimensional Semiconductor Device Simulation Program, AVANT! Corp., Fremont, CA, 2000.

[11] K. Cao, W.-C.Lee, W.Liu, X.Jin, P.Su, S. Fung, J. An, B.Yu, and C. Hu, "BSIM4 Gate Leakage Model Including Source Drain Partiotion", in IEDM Technical Digest, pp. 815-818, 2000.

[12] F. Hamzaoglu and M. Stan, "Circuit-Level Techniques To Control Gate Leakage For Sub-100 Nm CMOS," to be presented in International Symposium on Low Power Design, August 2002.

[13] Z. Chen, L. Wei, M. Johnson, K. Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," IEEE Int. Conf, on Comp. Aided Design, 1998.

[14] J. Lohstroh, E. Seevinck, and J. Groot, "Worst-case noise margin criteria for logic circuits and their mathematical equivalence," IEEE J. Solid State Circuits, vol. SC-18, pp. 803–806, Dec. 1983.

[15] D. Burger and T. M. Austin. The SimpleScalar tool set, version 2.0. Technical Report 1342, Computer Sciences Department, University of Wisconsin–Madison, June 1997.