Dynamic Vt SRAM : A Leakage Tolerant Cache Memory for Low Voltage Microprocessors

Chris H. Kim
hyungil@ecn.purdue.edu

Kaushik Roy
kaushik@ecn.purdue.edu

Department of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907, USA

ABSTRACT

This paper presents a Dynamic Vt SRAM (DTSRAM) architecture to reduce the subthreshold leakage in cache memories. The Vt of each cache line is controlled separately by means of body biasing. In order to minimize the energy and delay overhead, a cache line is switched to high Vt only when it is not likely to be accessed anymore. Simulation results from SimpleScalar framework show that even after considering the energy overhead, the DTSRAM can save 72% of the cache leakage with a performance loss less than 1%. Layout of the DTSRAM shows that the area penalty is minimal.

1. INTRODUCTION

Increasing on-chip integration and the large fraction of chip area devoted to memory sturctures has resulted in an unacceptably large leakage power dissipation for state-of-the-art microprocessor designs [1, 2]. Recent energy estimates for 0.13 μm processes indicate that leakage energy accounts for 30% of L1 cache energy and as much as 80% of L2 cache energy [3].

This paper presents a Dynamic Vt SRAM (DTSRAM) architecture to reduce the large leakage energy dissipation in memory structures. Body biasing was used to reduce the subthreshold leakage without sacrificing data stability [4]. A time-based dynamic Vt scheme is devised for the DTSRAM which only assigns a high Vt to the cache lines which are not accessed for a certain time period ($30\mu s \sim 100\mu s$). A low Vt is assigned to the cache lines which are in frequent use to maintain high performance. A Vt control circuit is designed which implements this time-based leakage reduction strategy. The analog implementation enabled us to reduce the leakage energy using a very simple hardware. Optimal design parameters for the DTSRAM are found by exploring their impact on total leakage energy savings. This paper also evaluates in detail the energy, performance, and area tradeoffs of the capacitor-discharging circuit scheme using architectural and circuit-level simulations.

2. DYNAMIC VT SRAM

2.1 Leakage in SRAM

ISLPED'02, August 12-14, 2002, Monterey, California, USA.

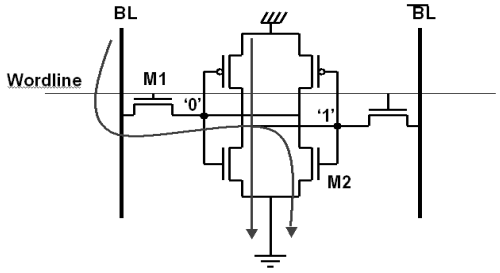


Figure 1: The two dominant leakage paths (Vdd to ground, bitline to ground) for a 6T SRAM cell. Leakage through the two paths consist 93% of the total leakage.

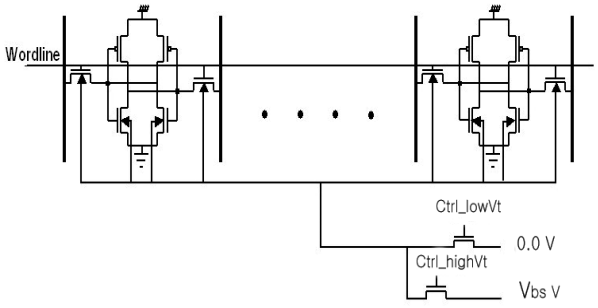


Figure 2: Schematic of a dynamic Vt SRAM set.

Fig. 1 depicts the two dominant leakage paths for a conventional 6T SRAM cell, the i) Vdd to ground and ii) bitline to ground leakage paths [6]. Together, they make up 93% of the total leakage. Substantial amount of leakage savings can be achieved by biasing the NFETs only since most of the leakage pass through the turned off NFETs in Fig. 1. Of course, reverse body biasing both the PFETs and NFETs can give the maximum leakage savings. However, the additional leakage savings gained by biasing the PMOS substrate is minute. Transition energy consumed while charging (or discharging) the substrate and the extra area required to separate the substrate contacts, isolate the substrates, and globally route the body bias network can be halved by not implementing the PMOS substrate biasing. Due to these considerations, PMOS substrate biasing is not implemented in our DTSRAM design. Fig. 2 shows the schematic of a DTSRAM cache line. The NFET substrate can be switched to 0V for high performance. In times when the cache line is not in use, the substrate can be switched to a negative voltage Vbs to reduce leakage. The following section describes an efficient strategy to turn on and turn off the cache lines.

2.2 A Time-Based Dynamic Vt Approach

SPICE simulations using TSMC 0.18μm show that the energy required for 1 transition is larger than the leakage energy saved during one clock cycle, by more than 4 orders of magnitude. Hence, making a V_t transition every cycle is disastrous in terms of energy savings. Speed of the NFETs on the discharging path also decreases as a negative body bias is applied. Apparently, the energy overhead and performance loss due to reverse body biasing is considerable.

To tackle the above-mentioned energy and delay overheads, a time-based approach is devised which intelligently turns off the cache lines. The strategy is based on the general access pattern of a cache line. When data is first brought in, it sees a burst of accesses. Then there is a dead period between the last access and the point when the data is replaced [7]. Leakage can be saved by turning off the cache line during the dead period. While the cache line is experiencing a burst of accesses, it is remained "on" to maintain the performance. Namely, rather than turning off a cache line right after its access, we leave it "on" for a certain time period ($30\mu s \sim 100\mu s$) so that the upcoming accesses within the time period will not impose energy or delay penalties. Energy and delay overhead is imposed when there is an access to a cache line which is in high V_t state. However this happens very rarely since most of the cache accesses are limited to the turned on portion of the cache due to the locality of reference.

3. CAPACITOR-DISCHARGING SCHEME FOR THE DYNAMIC VT SRAM

3.1 Overview

Schematic and waveforms of the V_t control circuit are shown in Fig. 3 and Fig. 4, respectively. The circuit consists of an RC decay circuit, a level converter to adjust the logic levels, and V_{sub} switches which drive the body terminals. When a cache line is accessed, V_{cap} is charged, immediately switching V_{sub} to 0V and making the corresponding cache line low V_t. V_{cap} starts discharging slowly at a decay time ($30\mu s \sim 100\mu s$) determined by the RC values. After this certain amount of time elapses without having any accesses, the SRAM cache line is switched to high V_t. In case there are accesses to the cache line before it is switched to high V_t, V_{cap} will be charged and the cache line will continue to remain low V_t until there is an idle period so long as the time for V_{cap} to completely discharge.

Performance of the cache is not affected by using the capacitor-discharging scheme, since most of the accesses will be on the fast, low V_t cache lines. Leakage energy is saved in the cache lines which are in idle mode. Energy and delay overhead is impose only when these high V_t cache lines have to be waken up. However, this happens very rarely, making the capacitor-discharging scheme profitable.

3.2 Circuit Design

A separate low supply voltage is used for the inverters in the V_t control circuit to reduce short circuit current. The short circuit current in the inverters is induced due to the intermediate voltage level while V_{cap} is decaying. The lower supply voltage for the inverters weakens the gate drive for the level converters. Larger PFETs are used in the level converters to compensate for the weak drive signal. The time constant of the capacitor decay can be changed by an analog control voltage, $V_{discharge}$. Our design shows a decay time of approximately 1ms when $V_{discharge}$ is 0V.

The V_{sub} switches in Fig. 3 are sized so that the time to switch from high V_t to low V_t is squeezed inside 1 clock cycle. An extra

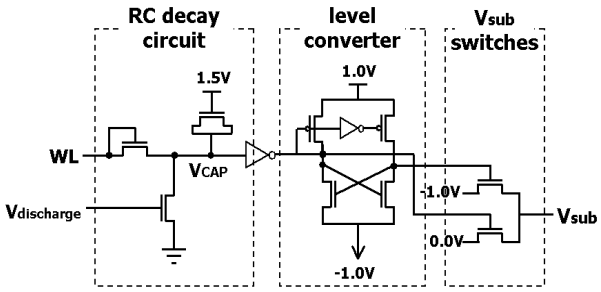


Figure 3: Schematic diagram of the Vt control circuit using capacitor-discharging scheme.

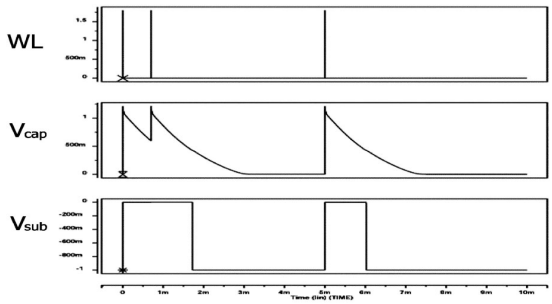


Figure 4: WL, Vcap and NMOS body bias voltage waveforms for the capacitor-discharging scheme.

clock cycle has to be added for the V_t transition whenever a high V_t cache line is accessed. This extra cycle becomes the delay penalty for the DTSRAM.

3.3 DTSRAM Layout

Fig. 5 shows the layout of 4 cache lines, each having 96 SRAM cells. The layout of the DTSRAM was done using TSMC 0.18μm technology. Due to area considerations, consecutive cache lines are flipped so that their substrates can be shared. Area overhead of the Vt control circuit is also reduced since only one V_t control circuit is required for 2 cache lines. A deep N-well layer was used to isolate the P-substrates between each cache line. TSMC 0.18μm design rules require a comparably large area margin on the edge of each deep N-well layer and this leads to an increase in the SRAM cell area [8]. Table 1 shows the layout and area of a conventional SRAM cell and a DTSRAM cell. The DTSRAM cell turns out to be 15.5% larger than the conventional SRAM cell. An additional area penalty of 64.94μm^2 is imposed due to the V_t control circuits in each cache line.

For technologies which are optimized for body biasing, the area penalty has been examined to be much less than our results [4, 9]. More leakage power can be saved with less increase in area for these dedicated technologies. Even though it is shown (section 4.3) that isolating the substrate of each cache line is advantageous in terms of leakage reduction, the area overhead becomes unacceptably large (up to 44%), restricting separate body biasing of each cache line. Technology issues for the DTSRAM can be a future research work.

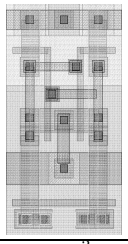
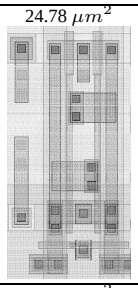
4. SIMULATION RESULTS

4.1 Simulation Setup

Architectural behavior of the DTSRAM cache was examined based on a SimpleScalar-3.0 framework. The DTSRAM param-

Figure 5: Layout of 4 cache lines with each having 96 SRAM cells. 2 cache lines are controlled by a single Vt control circuit to save area.

Table 1: Area of conventional SRAM and DTSRAM.

	Conventional SRAM	DTSRAM
6T cell area	21.45 μm^2 	24.78 μm^2 
Vt control circuit area	0 μm^2	64.94 μm^2

eters such as energy overhead, leakage savings and total leakage energy were extracted from the DTSRAM layout. These parameters are fed into the SimpleScalar simulator, which emulates the capacitor-discharging circuit scheme. The total energy overhead and leakage savings were evaluated for a 64KB L1 instruction cache. The simulation results were acquired after running 100 million instructions from various SPEC2000, SPEC95 benchmark applications. The out-of-order model processor has the memory hierarchy shown in table 2.

4.2 Leakage Energy Savings

Simulation results on energy penalty, absolute leakage savings and net leakage savings are shown in Fig. 6. "Absolute leakage savings" denote the pure leakage savings without considering the energy penalty. "Net leakage savings" is derived by subtracting the energy penalty from the absolute leakage savings in order to take the energy overhead into consideration. Energy penalty in Fig. 6 sharply decreases as the time constant increases due to the less number of V_t transitions. The V_t transitions occur less frequently since the interval between accesses to a same cache line become shorter than the time constant. The absolute leakage savings in Fig. 6 slightly decreases since it takes a longer time for a cache line to be turned off and save leakage power. It turns out that the reduction in transition energy is much larger than the decrease in absolute leakage savings and hence, the net leakage savings improves when a larger time constant is used. The net leakage savings curve remains at its maximum value for time constants larger than 16K instruction cycles.

4.3 Optimal Bank Size

The optimal bank size (i.e. number of cache lines grouped together, sharing the same substrate) has to be determined for maximum leakage savings. Having a larger bank size, reduces the number of V_t transitions since some cache lines are turned on in advance to their usage due to accesses on other locations in the same bank. However, energy required for one V_t transition becomes larger because a larger substrate capacitance must be charged (or discharged) for larger banks. Trade off between the number of transitions and the energy per transitions dictate the total transition energy overhead for the DTSRAM. From the leakage savings point of view, it turns out to be less efficient to have a larger bank size because some of the unused cache lines will be turned on, even though they can be turned off to save further leakage power for smaller bank sizes. The increase in total transition energy is greater than the decrease in leakage savings which makes the DTSRAM less efficient in saving leakage for larger bank sizes. (Fig. 7) Therefore, it is optimal to have a bank size of 1 where the V_t of each cache is controlled separately.

Table 2: Memory hierarchy used for SimpleScalar simulations.

L1 Icache	64KB, 64B block, direct mapped, write back
L1 Dcache	64KB, 64B block, direct mapped, write back
L2 I,Dcache	Unified, 1MB, 64B block, 8-way, write back
Main memory	100 cycle

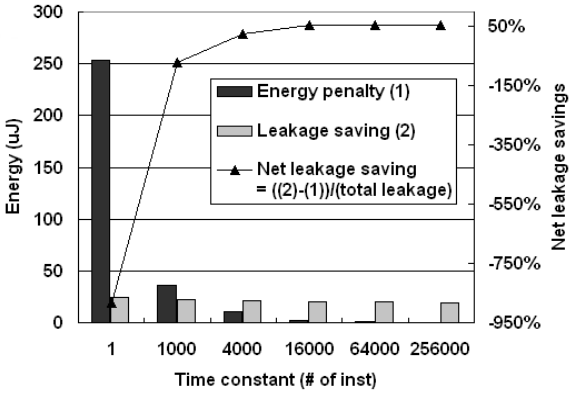


Figure 6: Energy penalty, absolute leakage savings and net leakage savings of a DTSRAM.

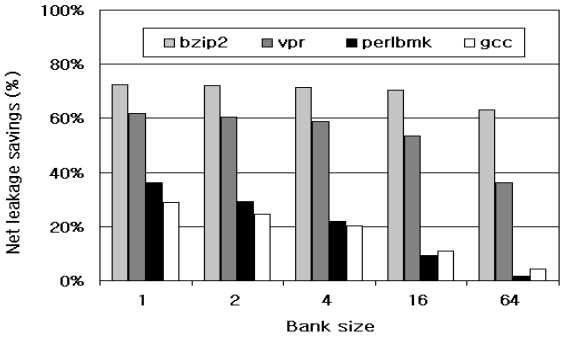


Figure 7: Net leakage savings vs. bank size.

4.4 Optimal Time Constant

The net leakage savings for SPEC2000 benchmark programs are shown in Fig. 8. The optimal time constant giving maximum leakage savings differ from benchmark to benchmark. Even though we mentioned in Fig. 3 that the time constant can be adjusted by changing the $V_{discharge}$, the high sensitivity of analog circuits make it difficult to obtain a precise time constant. Fortunately, any time constant between 64K and 256K instruction cycles gives a near-optimal leakage savings due to the flat optimal region in Fig. 8.

Table 3: Delay penalty vs. time constant for SPEC2000 benchmark applications.

time constant	gcc	gzip	vpr	vortex	mcf	twolf	bzip2	perlbmk	gap
1	19.2%	3.6%	22.4 %	19.5%	22.2%	21.2%	20.2%	24.5%	18.2%
1000	2.7%	0.0%	4.4 %	3.9%	5.7%	3.1%	0.0%	5.7%	0.4%
4000	1.0%	0.0%	3.0 %	0.0%	2.6%	0.9%	0.0%	2.3%	0.2%
16000	0.0%	0.0%	0.0 %	0.0%	0.7%	0.2%	0.0%	0.8%	0.0%

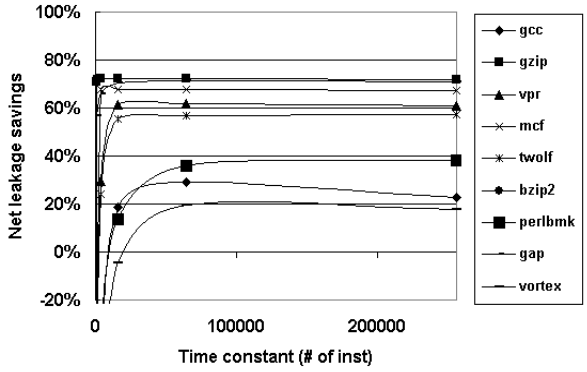


Figure 8: Net leakage savings vs. time constant for SPEC2000 benchmark applications.

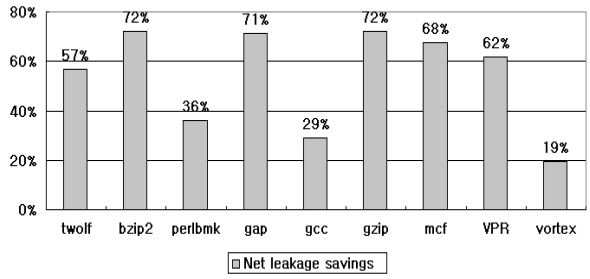


Figure 9: Net leakage savings using a time constant of 64K instruction cycles.

The performance loss of the DTSRAM turns out to be minimal for the optimal time constants. Performance degradation occurs when a high V_t cache line is accessed. The cache data can be immediately read, but an extra clock cycle has to be inserted for the V_t transition. Delay penalty can be simply derived from the ratio of cache accesses which cause a V_t transition. Table 3 shows the simulation results. For time constants greater than 16K instruction cycles, the delay penalty becomes less than 1% for all benchmark programs.

Leakage savings, performance degradation and functionality are not sensitive to the choice of time constant as long as it is in the wide optimal region in Fig. 8. This nature implys robust operation of the DTSRAM against temperature and process variations. Finally, the net leakage savings for various benchmark programs are shown in Fig. 9. A near-optimal time constant of 64K instruction cycles (equivalent to $32\mu s$ for a 2 GHz microprocessor) was used for the capacitor-discharging circuit scheme. The results show that the DTSRAM can save 54% of the leakage power on average, with a performance degradation less than 1%. A bank size of 1 gives the maximum leakage reduction where each cache line has its dedicated V_t control circuit.

5. CONCLUSIONS

This paper deals with a dynamic V_t SRAM (DTSRAM) which can reduce the leakage energy dissipation in deep-submicron cache memories. Body biasing is used to separately control the V_t of each cache line. To minimize the energy and delay overhead, a capacitor-discharging circuit scheme is devised which selectively "turns off" the cache lines, which are not likely to be used anymore. Only the cache lines in frequent use are "turned on" to maintain the performance. Layout of the DTSRAM shows 15.5% increase in cell area compared to the conventional SRAM cell for TSMC $0.18\mu m$.

Architecture simulation results of the DTSRAM are presented by running SPEC2000 benchmarks on a SimpleScalar framework. The results indicate that even after considering the energy overhead, the DTSRAM can save up to 72% of the leakage power, with a performance degradation less than 1%. We have also discussed the optimal time constant and bank size which gives the maximum leakage savings. Results show that a large ($>$64K instruction cycles) time constant and a bank size of 1 gives the maximum leakage savings.

6. REFERENCES

[1] S. Borkar, "Design Challenges of Technology Scaling", *IEEE Micro*, 19(4):23-29, July 1999.

[2] M. Powell, S. Yang, B. Falsafi, et al, "Gated-Vdd: A Circuit Technique to Reduce Leakage in Cache Memories", *ISLPED*, pp.90-95, July 2000.

[3] V. De, Private communication

[4] H. Mizuno, K. Ishibashi, T. Shimura, et al, "An 18-μA Standby Current 1.8-V, 200-MHz Microprocessor with Self-Substrate-Biased Data-Retention Mode", *IEEE JSSC*, vol. 34, no. 11, nov 1999.

[5] A. Keshavarzi, S. Ma, S. Narendra, et al, "Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual V_t CMOS ICs", *ISLPED*, pp. 207-212, 2001.

[6] A. Agarwal, H. Li, and K. Roy, "DRG-Cache: A Data Retention Gated-Ground Cache for Low Power", *DAC*, to be published, 2002.

[7] S. Kaxiras, Z. Hu and M. Martonosi, "Cache Decay: Exploiting Generational Behavior to Reduce Cache Leakage Power", *ISCA*, pp. 240-251, 2001.

[8] *TSMC 0.18UM Mixed Signal/RF 1P6M+ Salicide 1.8V/3.3V Design Rule*, Taiwan Semiconductor Manufacturing Co., LTD

[9] T. Kuroda, T. Fujita, S. Mita, et al, "A 0.9-V, 150-MHz, 10-mW, 4mm^2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage Scheme", *IEEE JSSC*, vol.31, no. 11, nov 1996.