Appears in the Proceedings of the 2001 International Conference on Computer Design

Static Energy Reduction Techniques for Microprocessor Caches

Heather Hanson* M.S. Hrishikesh* Vikas Agarwal* Stephen W. Keckler Doug Burger

Computer Architecture and Technology Laboratory
Department of Computer Sciences
*Department of Electrical and Computer Engineering
The University of Texas at Austin
cart@cs.utexas.edu — www.cs.utexas.edu/users/cart

Abstract

Microprocessor performance has been improved by increasing the capacity of on-chip caches. However, the performance gain comes at the price of increased static energy consumption due to subthreshold leakage current. This paper compares three techniques for reducing static energy consumption in on-chip level-1 and level-2 caches. One technique employs low-leakage transistors in the memory cell. Another technique, power supply switching, can be used to turn off memory cells and discard their contents. A third alternative is dynamic threshold modulation, which places memory cells in a standby state that preserves cell contents. In our experiments, we explore the energy/performance trade-offs of these techniques and find that dynamic threshold modulation achieves the best results for level-1 caches, improving the energy-delay product by 2% in a level-1 instruction cache and 7% in a level-1 data cache. Low-leakage transistors perform best for the level-2 cache as they reduce static energy by up to 98% and improve the energy-delay product by more than a factor of 50.

1 Introduction

Continued improvements in integrated circuit fabrication technology have enabled the number of transistors in microprocessors to more than double with every generation. A vast majority of transistors in modern microprocessors are used for on-chip storage, including level-1 and level-2 caches, and meta-state such as renaming registers, numerous predictor structures, and trace caches. As leakage current increases with each technology generation, the energy consumption of memory structures will increase dramatically with future process technologies. In this paper, we explore the energy/performance trade-offs of three leakage-reduction techniques for on-chip level-1 and level-2 caches.

One method, *dual-V_t*, involves fabricating the SRAM array transistors to have a high threshold voltage. Transistors in the remainder of the SRAM circuit have a lower threshold voltage for faster switching speed. This dual-V_t method decreases subthreshold leakage currents but increases the cell access time compared with an SRAM composed of fast, leaky transistors [9, 13]. Another method dynamically adjusts the effective size of the array by employing a circuit technique dubbed *gated-V_{dd}*. In this scheme, a low-leakage transistor is used to selectively shut off the power supply to a subset of SRAM cells [11]. Thus, the capacity of the array adjusts dynamically as the amount of active information in the cache changes throughout the duration of the program.

A third technique, *MTCMOS*, dynamically changes the threshold voltage by modulating the backgate bias voltage [8, 10]. With this technique, memory cells can be placed into a low-leakage "sleep" mode yet still retain their state. Cells in the active mode are accessed at full speed while accesses to cells in the sleep mode must wait until the cell has been awakened by adjusting the bias voltage. The MTCMOS technique has been implemented for an entire SRAM [10]; we examine this idea using a fine-grain control of each cache line.

While the fundamental circuits for leakage reduction have been introduced by other researchers, our contributions in this paper are to examine the energy/performance tradeoffs of these techniques applied to the memory hierarchy of a modern microprocessor. The paper is organized as follows. Section 2 introduces leakage current and its effects on cache energy. Section 3 describes three methods for reducing leakage current in memory cells; Section 4 explains our experimental methodology. Results of the experiments and a comparison of these techniques are presented in Section 5. Section 6 highlights relevant related work, and is followed by concluding remarks in Section 7.

2 Leakage Current

Power consumption in a digital integrated circuit is governed by the equation:

$$P = \alpha C V^2 f + I_{off} V \quad (1)$$

where α is the average switching activity factor of the transistors, C is capacitance, V is the power supply voltage, f is the clock frequency, and I_{off} is the leakage current. The first term of the equation is dynamic power and the second term is static power. Smaller feature sizes in each generation of silicon process technologies have been accompanied by reduced power supply voltages that have helped to mitigate the impact of increased transistor counts and higher clock frequencies on dynamic power. However, as the power supply voltage decreases, threshold voltages of the transistors must also decrease to achieve fast switching speeds and sufficient noise margins. Subthreshold leakage current I_{off} is dominated by temperature T and transistor threshold voltage V_t in the following equation:

$$I_{off} \propto e^{(\frac{-V_t}{T})} \quad (2)$$

Thus, lower threshold voltages lead to increased subthreshold leakage current and increased static power [3]. Most previous efforts at power reduction have focused on dynamic power sources because static power due to leakage current has been a small fraction of the total power dissipated by a chip. However, as transistor threshold voltages are reduced, subthreshold leakage current increases dramatically. If left unchecked, subthreshold leakage current will become a major contributor to total power dissipation.

Figure 1 shows estimated static power consumption due to leakage current in large secondary caches through five technology generations. In this chart, cache capacities are scaled from 1MB to 16MB, reflecting high-performance microprocessor cache sizes projected by [6]. Supply voltages are scaled from 1.6V to 0.6V. Leakage data are based on a circuit temperature of 110° C, a high temperature achieved during chip operation. Three leakage-current scaling models are charted: a linear projection from [1] for 180nm through 100nm that is scaled for 70nm and 50nm, and two experimental leakage models for high V_t (low leakage) and low V_t (high performance) devices. Estimates of leakage current vary due to expectations of circuit parameters and material properties, though the trend of exponential increase in static power as transistor sizes decrease is evident in each curve. Thus, static energy also increases with each process generation.

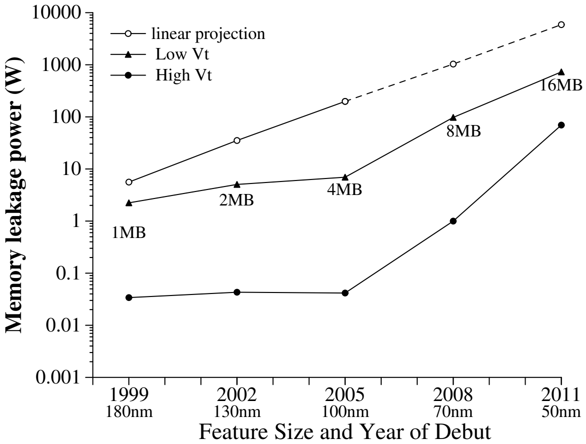

Figure 1. Projected Leakage Power of Level-2 Caches Through Technology Generations.

3 Leakage Reduction Techniques

This section describes our implementation of each leakage reduction strategy and our experimental methodology to simulate each technique applied to the level-1 instruction cache (IL1), level-1 data cache (DL1), and level-2 cache (L2).

3.1 Static Threshold Selection: Dual-V_t

The dual-V_t technique employs transistors with higher threshold voltages in memory cells and faster, leakier transistors elsewhere within the SRAM. This technique requires no additional control circuitry and can substantially reduce the leakage current when compared to low V_t devices. The amount of leakage current is engineered at design time, rather than controlled dynamically during operation. No data are discarded and no additional cache misses are incurred. However, high-V_t transistors have slower switching speeds and lower current drive. In our experiments, we consider an additional cycle of access time for SRAMs with these high-threshold devices.

3.2 Power Supply Switching: Gated-V_{dd}

The gated-V_{dd} technique interposes a high-threshold transistor between the circuit and one of the power supply rails [11]. The left circuit in Figure 2 shows the schematic of a gated-V_{dd} SRAM cell with an NFET selectively connecting the cell to the ground rail. When the `active` signal is asserted, the SRAM cell operates normally, but when

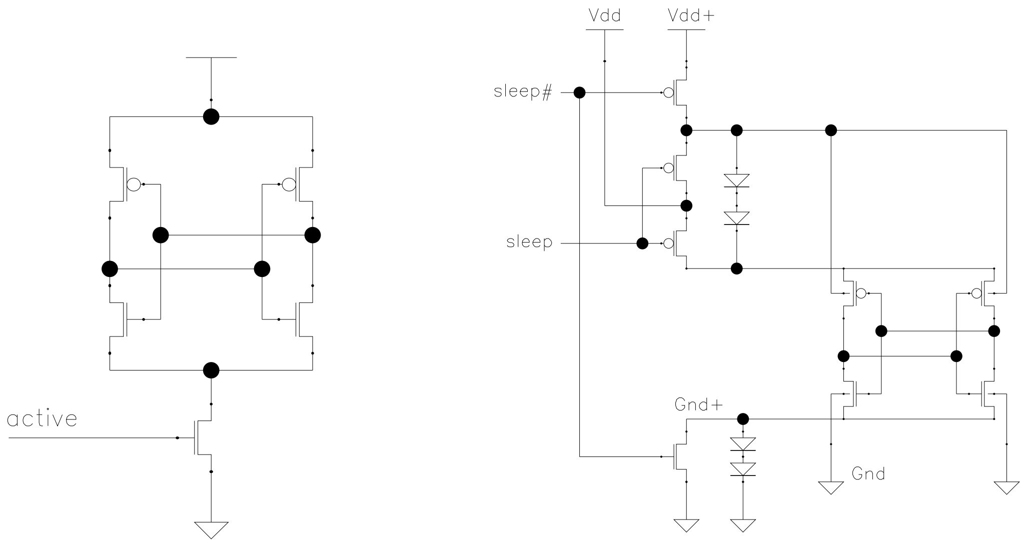

Figure 2. Gated-V_{dd} and MTCMOS SRAM cell schematics

Table 1. Summary of Leakage Reduction Techniques

Technique	Benefit	Detriment
Dual-V_t	no additional circuitry	each read access is slower
Gated-V_{dd}	simple circuit	additional cache misses
MTCMOS	no additional cache misses	complex circuitry with diodes

`active` is deasserted, the cell is disconnected from ground and the state contained within the cell is lost. The activation transistor and the control mechanism for `active` can be shared by all cells within a cache line to minimize the extra area needed by the control transistor. We assume that this power supply gating transistor is sized so that the increase in memory array access time is negligible.

3.3 Dynamic Threshold Modulation: MTCMOS

Leakage current may also be reduced by dynamically raising the transistor threshold voltage, typically by modulating the back-gate bias voltage. A technique amenable to fine-grain control is Auto-Backgate-Controlled Multi-threshold-CMOS (which we will refer to as MTCMOS), as shown in the right circuit of Figure 2 [8, 10]. During normal operation, when `sleep` is deasserted, the SRAM is connected to V_{dd} and ground and back-gate voltages are set to the appropriate power rails. When `sleep` is activated, the PFET wells are biased using an alternative power supply voltage, V_{dd+}, at a higher voltage level than the source terminals. Diodes allow the voltage levels of source terminals of the NFETs to increase by two diode drop voltages while the NFET well remains at Gnd. Thus all transistors experience higher threshold voltages and a corresponding drop in leakage current. As with gated-V_{dd}, we assume that any increase in memory array access time is negligible while `sleep` is not asserted.

The advantage of adjusting the threshold voltage dynamically, rather than gating the power supply, is that memory cell values are preserved during sleep mode, so there are no additional cache misses caused by accessing a line in the low-power mode. This technique provides an opportunity to reduce static power consumption without incurring the cost to retrieve data from another level of the hierarchy. The disadvantages of MTCMOS include an additional power supply voltage that must be distributed throughout the array and larger electric fields placed across the transistor gates during sleep mode that may adversely affect reliability. Table 1 summarizes the primary advantages and disadvantages of the three techniques for reducing leakage energy.

3.4 Decay Intervals

Energy-saving techniques such as gated-V_{dd} and MTCMOS that disable cache line rely on two properties of the data stored in caches. First, only a small fraction of the information in the cache is *live*, meaning that it will be referenced again before being replaced or over-written. In our experiments, we found that only 1–30% of a 2MB level-2 cache holds live data, depending on the application. Even in level-1 caches, less than half of the cache contains useful data across our benchmark suite. Second, most lines that will be reused are accessed within a relatively short time interval.

Cache lines containing information that is either not useful or will not be accessed for a long time can be put into an idle, low-leakage mode to save energy without a significant effect on processor performance. We determine which lines to place in an idle mode in the gated-V_{dd} and MTCMOS methods by measuring inter-access times, similar to Kaxiras *et al.* [7] who proposed low frequency counters to measure the time since last reference for every cache line. A read or write to a cache line resets its counter; when the counter reaches its maximum value after a duration named the *decay interval*, the line is deactivated.

4 Experimental Methodology

To evaluate the effectiveness of the leakage-reduction techniques, we modified a version of the SimpleScalar simulator [2]. We added the capability to discard cache lines or put them to sleep after a specified decay interval had passed since the last access to the cache line.

4.1 Simulation Methodology

Our benchmark suite for this study consists of five SPEC2000 benchmarks: *gcc*, *eon*, *equake*, *mcf*, and *vpr* compiled for the Alpha instruction. The simulation execution core is configured as a 4-wide superscalar pipeline organization roughly comparable to the Compaq Alpha 21264. The memory hierarchy consists of a 64KB, 2-way set associative level-1 instruction cache with a single-cycle hit latency, a 64KB, 2-way set associative level-1 data cache with a 3-cycle hit latency, and a unified 2MB 4-way level-2 cache with a 12-cycle hit latency. In the gated-V_{dd} and MTCMOS techniques, data bits may be placed into an idle mode and cache tags are kept in the active state to provide fast lookup times.

In each experiment, we applied a leakage reduction technique to one cache and simulated benchmark execution with SimpleScalar. The simulations ran for 1 billion instructions after fast-forwarding through the first 500 million instructions. We measured instructions per cycle (IPC), active and inactive durations for each cache line, the number of hits and misses in each level of the hierarchy, and the number of times any cache line is enabled or disabled. For gated-V_{dd}, disabling a cache line is equivalent to switching off the power supply, while for MTCMOS, it is equivalent to placing the cache line into sleep mode. We calculated the total energy by multiplying these measured quantities by the relevant static and dynamic energy parameters described below and summing the energy consumed by individual components of the system.

4.2 Energy Parameters

Leakage currents and energy values were measured with the HSPICE circuit simulator using anticipated 70nm technology parameters; the clock rate is set to 16 fanout-of-four inverter delays [5]. Table 2 summarizes the experimental parameters used in this study. In this table, I_{max} and I_{min} are projected leakage currents when SRAM cells are active and disabled, respectively. In each experiment, V_t = 0.4V for high-threshold voltage transistors and V_t = 0.2V for low-threshold voltage transistors. E_{switch} approximates the energy required to switch the cell between active and inactive modes. E_{IL1}, E_{DL1}, and E_{L2} represent the energy to read data from the level-1 instruction, level-1 data, and level-2 caches, respectively, based on a modified version of the cache tool CACTI 2.0 [12] and our projected 70nm process parameters. The energy to drive package pins for off-chip memory accesses to service L2 misses is represented by E_{pins} [4]. We account only for the pin energy that is expended in driving the address to the pins of the CPU, and not energy expended to receive data.

The total dynamic energy is calculated as the number of cache accesses multiplied by the appropriate energy per access parameter, plus the number of transitions into idle mode multiplied by the energy per transition (where applicable). To compute the dynamic energy expended in cache accesses, we make the following approximations: (1) level-1 cache miss energy is equal to two cache hit accesses (one to detect the miss and one to load new data); (2) level-2 cache miss energy is equal to two cache hit accesses plus the energy to drive an address to 32 address pins for off-chip memory; and (3) any power consumed outside the CPU chip is not included in this study.

Static energy is computed as the product of static power per cycle and the number of cycles of program execution. In this paper, we focus only on the leakage in the cache memory arrays; this approximation neglects the leakage current due to the small fraction of transistors in the peripheral circuitry. The total energy is the sum of dynamic and static energy calculations.

Energy consumption and performance of the leakage-reduction techniques are compared to a baseline case to evaluate the experimental techniques' effectiveness in static energy reduction and performance. Implementation details specific to this baseline and the experimental techniques are outlined below.

Baseline: The baseline case in this study is a high-performance cache without leakage current control. Each transistor in the SRAM cell has a threshold voltage of 0.2V,

Table 2. Experimental Parameters for Energy Calculations.

	70nm Technology		Per-Bit Leakage Current (110 C)		Per-Bit Transition Energy	Dynamic Energy Per Cache Access			
Technique	Clock Rate (GHz)	V_{dd} (Volts)	I_{max} (nA)	I_{min} (nA)	E_{switch} (fJ)	E_{IL1} (nJ)	E_{DL1} (nJ)	E_{L2} (nJ)	E_{pins} (nJ)
Baseline	2.5	0.75	1941	-	-	0.07	0.07	4.5	0.9
Dual-V_t	2.5	0.75	-	26	-	0.07	0.07	4.5	0.9
Gated-V_{DD}	2.5	0.75	1939	9.7	0.35	0.07	0.07	4.5	0.9
MTCMOS	2.5	0.75	1941	12	50	0.07	0.07	4.5	0.9

with a high leakage current of I_{max} at all times. The baseline case has the maximum performance and maximum energy consumption for the set of experiments.

Dual-V_t: Though the dual-V_t technique has low-leakage transistors in memory cells and high-leakage transistors elsewhere, we account for static energy only in the memory array, and thus only use the reduced-leakage current, I_{min}. The dual-V_t technique does not transition between idle and active states and thus does not incur extra cache misses.

Gated-V_{dd}: For the gated-V_{dd} technique, I_{max} is the leakage current when the memory cell is in the active state, and I_{min} is the leakage current when the memory cell is disconnected from the power supplies. The gating transistor has a high threshold voltage of 0.4V, and the other SRAM cell transistors' threshold voltages are the low-V_t value of 0.2V. The value of E_{switch} is based on the gate capacitance of the activation transistor and the wire capacitance to reach all of the cells in the cache line. Only "clean" lines that do not require a write back to the memory hierarchy are disabled; "dirty" lines that are not accessed before the decay interval expires are kept in the active state.

MTCMOS: The leakage current for MTCMOS SRAM arrays is controlled on the granularity of a cache line. Transistors in the SRAM cells have a V_t of 0.2V. I_{max} is the leakage current when the memory cell is awake, and I_{min} is the leakage current when the cells have transitioned into sleep mode. The time and energy to enter and exit sleep mode depend directly on the effective capacitance of the well that contains the PFETs in the SRAM cell; in this study, we use a single cycle of delay to awaken a sleeping cache line prior to accessing it. E_{switch} is the energy required to charge the cache line's well plus the energy consumed to discharge the source terminals of the NFETs.

5 Results

This section presents our experimental results and compares trade-offs between performance and energy reduction for three leakage-reduction techniques. We use a metric of the energy-delay product to balance the benefits of lower leakage with the potential penalty of reduced performance. We calculate the energy-delay product as the total energy divided by IPC, which is equivalent to the product of energy and a measure of time (cycles per instruction).

To evaluate the gated-V_{dd} and MTCMOS strategies, we observed the techniques' performance throughout a range of decay intervals, and chose intervals that resulted in the minimum energy-delay product. The best-case decay interval depends upon program cache access patterns and circuit parameters unique to each leakage-reduction technique [4]. In our study, the best decay interval for the gated-V_{dd} technique is 64K cycles for each cache. For the MTCMOS technique, the best decay interval is 8K cycles for the level-1 instruction cache, 1K cycles for the level-1 data cache, and immediate sleep mode for the level-2 cache. Table 3 summarizes the experimental results, reported as the harmonic mean of IPC, energy, and energy-delay product for simulated program execution across the benchmark suite.

For the parameters in this study, the MTCMOS technique has the best combination of energy reduction and performance for level-1 caches. The dual-V_t technique produces the lowest energy-delay product for level-2 caches. Figure 3 shows the total energy required for program execution for each leakage-reduction technique applied independently to one cache. The charts present data from the best decay interval in the gated-V_{dd} and MTCMOS techniques. In the figures in the left column, stacked bar charts illustrate the contribution of static and dynamic energy for each benchmark. Note that in the level-1 caches, the majority of energy consumption is due to dynamic energy, whereas in level-2 caches, static energy dominates the total energy. Charts in the right column of Figure 3 show the energy-delay product for each benchmark and highlight the variation between techniques. Each of the three leakage-reduction methods in this study achieves lower leakage energy compared to the baseline case with high-performance SRAM cells but sacrifices performance to do so, whether by slowing cache accesses or causing delays to re-fetch data.

Table 3. Summary of Experimental Results: Harmonic Mean Across Benchmark Suite

Level-1 Instruction Cache						
Technique	Decay Interval	IPC	Total Energy(J)	Dynamic Energy (J)	Leakage Energy (J)	Energy-Delay (E/IPC)
Baseline	-	1.645	4.688	4.539	0.141	2.663
Dual-V_t	-	0.680	4.525	4.520	0.005	6.181
Gated-V_{dd}	64K	1.641	4.584	4.539	0.039	2.613
MTCMOS	8K	1.644	4.580	4.539	0.035	2.607

Level-1 Data Cache						
Technique	Decay Interval	IPC	Total Energy (J)	Dynamic Energy (J)	Leakage Energy (J)	Energy-Delay (E/IPC)
Baseline	-	1.645	1.679	1.530	0.141	0.942
Dual-V_t	-	1.540	1.520	1.518	0.002	0.898
Gated-V_{dd}	64K	1.643	1.571	1.531	0.030	0.885
MTCMOS	1K	1.639	1.547	1.530	0.017	0.874

Level-2 Cache						
Technique	Decay Interval	IPC	Total Energy(J)	Dynamic Energy (J)	Leakage Energy (J)	Energy-Delay (E/IPC)
Baseline	-	1.645	4.540	0.004	4.513	2.424
Dual-V_t	-	1.625	0.084	0.004	0.061	0.042
Gated-V_{dd}	64K	1.386	0.239	0.005	0.225	0.112
MTCMOS	0	1.626	0.140	0.004	0.115	0.072

5.1 Dual-V_t

The dual-V_t cache is effective at reducing leakage; however, with an extra cycle of delay, the technique has a negative effect on performance for level-1 caches. The dual-V_t technique reduces the static energy consumed by the IL1 cache by 96%, at the expense of reducing the IPC by over half. The energy-delay product of the dual-V_t technique is more than twice that of the IL1 baseline case. Although the leakage current and therefore static energy is reduced, the performance penalty may be unacceptable for a dual-V_t method applied to an instruction cache, or other structures that rely on fast access times. The dual-V_t DL1 cache reduces static energy by 98%, with an energy-delay product that is 4% better than the baseline case. In the level-2 cache experiment, the dual-V_t technique improves both static energy and energy-delay product. Static energy decreases by 98% with negligible performance degradation and the energy-delay product improves by over a factor of 50.

5.2 Gated-V_{dd}

With gated-V_{dd}, static energy savings are offset by the dynamic energy and time required to service additional misses to prematurely disabled cache lines. The total energy of the frequently accessed primary caches is dominated by dynamic energy of read accesses, and despite substantial static energy savings, the energy-delay product is only slightly better than the baseline case. The gated-V_{dd} technique applied to an IL1 with a 64K decay interval produces a 72% static energy savings, with a 2% improvement in energy-delay compared with the baseline. In the DL1 cache, the technique had similar results: 79% reduction in static energy, with a 6% improvement in the energy-delay product. In the level-2 cache, the penalty for additional execution time creates a noticeable drop in IPC. However, the energy savings with the gated-V_{dd} technique is 95%, for an overall effect of improving the energy-delay by a factor of 20.

5.3 MTCMOS

The MTCMOS technique has the best energy-delay performance for level-1 caches. The MTCMOS IL1 cache with an 8K decay interval reduces static energy by 75%, an improvement in energy-delay of 2%. In the DL1 cache, the MTCMOS technique and a 1K decay interval decreases static energy by 88%, while improving the energy-delay product by 8%. For the level-2 cache and an aggressive sleep policy, leakage current is dramatically reduced at the expense of a slightly lower IPC. The level-2 cache with MTCMOS circuitry and an immediate sleep mode reduces static energy by 97% the energy-delay product by a factor of approximately 34.

6 Related Work

Leakage-reducing circuit techniques can be incorporated into architectural solutions that rely on programs' use of system resources to reduce static energy. One example employs a gated-V_{dd} circuit to selectively disable cache lines based on miss rates, dynamically resizing the instruction cache (DRI I-cache) to a size appropriate for the currently executing program. Yang *et al*. found that a 64K DRI I-cache reduced the energy-delay product by 62% with a 4% increase in execution time with SPEC95 benchmarks, com-

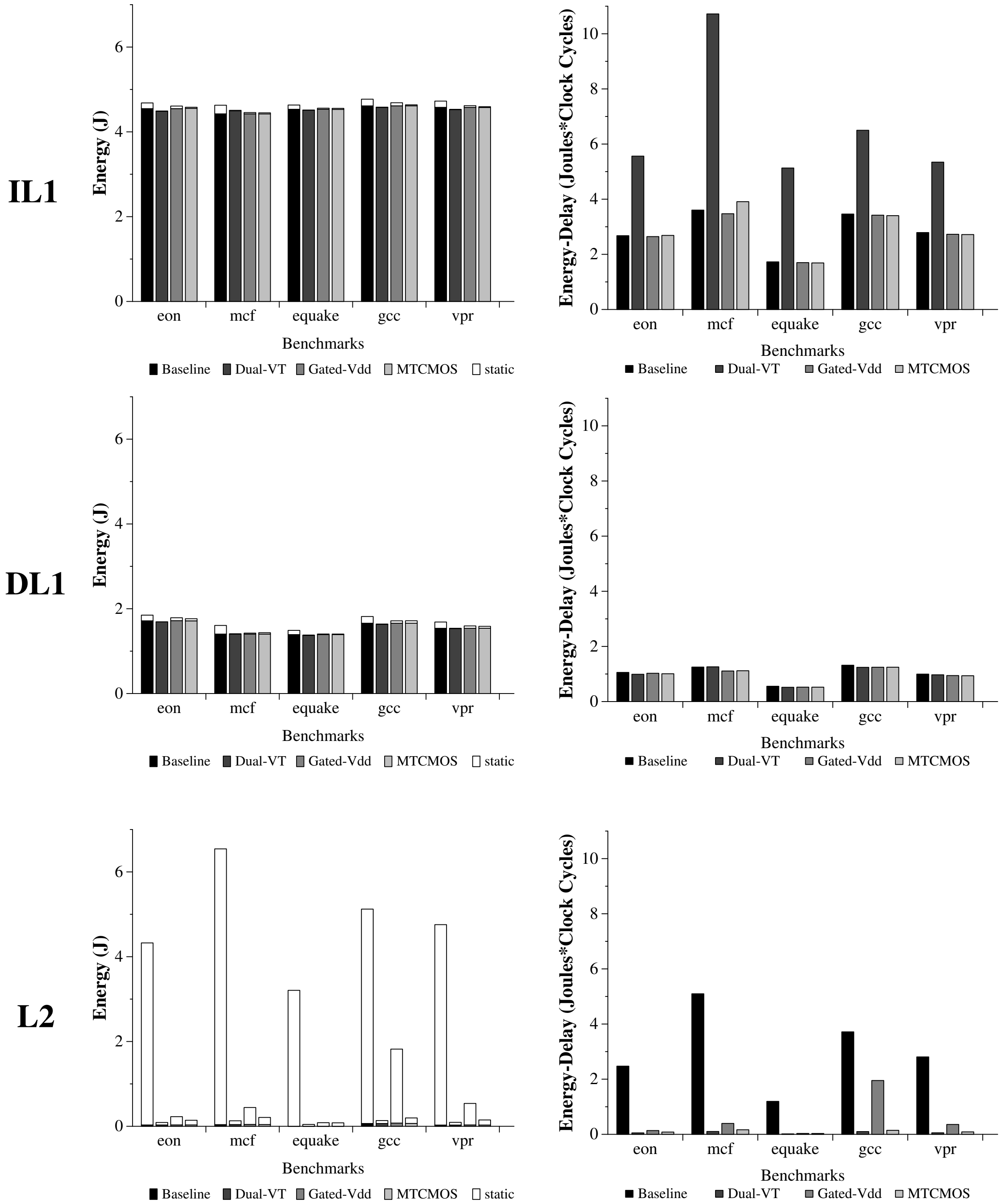

Figure 3. Energy and Energy-Delay Product for L1 and L2 Caches.

pared to a standard cache [14]. Kaxiras *et al.* are continuing development of the gated-V_{dd} technique with an adaptive control on the gating transistor, and have shown that their technique can reduce leakage energy in level-1 caches by a factor of five [7]. Zhou *et al.* have proposed a low-leakage cache design named Adaptive Mode Control that dynamically adjusts the number of cache lines turned off by the gated-V_{dd} method throughout program execution to keep the number of standard cache misses proportional to extra misses caused by disabling cache lines. With adaptive mode control, a level-1 instruction cache with an average of 74% of the cache lines disabled and a level-1 data cache with an average of 50% disabled cache lines results in an IPC drop of less than 1.6% [15].

7 Conclusion

In this paper we have explored energy/performance trade-offs associated with three techniques for reducing static energy consumption in on-chip caches: high-V_t transistors in memory arrays, power supply switching, and dynamic transistor threshold modulation. With our assumptions, the MTCMOS technique yields the best energy-delay product for level-1 caches, improving by 2% in the IL1 cache and 7% in the DL1 cache compared to the experimental baseline. Each technique is more effective in a level-2 cache than a level-1 cache. The dual-V_t technique applied to the level-2 cache resulted in a 50-fold improvement of energy-delay, while the gated-V_{dd} and MTCMOS techniques resulted in overall reductions of factors of 20 and 34, respectively.

In our models, MTCMOS achieves the best results for energy-delay in the level-1 caches because it does not experience higher miss rates like gated-V_{dd}, and incurs a single-cycle wakeup penalty for standby data. However, the results are sensitive to additional latency and energy penalties contributed by the leakage reduction strategy. In our experiments, increasing the wake-up latency to 5 cycles in the level-1 caches noticeably reduces performance of MTCMOS [4].

While this paper has emphasized reducing static energy in cache memories, the same principles may be applied to other on-chip structures as well. For example, the static energy required to maintain the state of branch predictor table entries may be balanced against the dynamic energy required to execute with fewer correct predictions. Future work will include an analysis of redundant and unneeded data in modern microprocessors and its implications for energy and performance.

Acknowledgments

We thank the anonymous reviewers for their helpful suggestions. This work was supported by Intel and IBM in the form of equipment grants and fellowships for Heather Hanson and Vikas Agarwal, by the NSF CADRE program, grant no. EIA-9975286, and by a grant from the Intel Research Council.

References

[1] S. Borkar. Design challenges of technology scaling. *IEEE Micro*, 19(4):23–29, July-August 1999.

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

[3] J. A. Butts and G. S. Sohi. A static power model for architects. In *Proceedings of 33rd Annual International Symposium on Microarchitecture*, December 2000.

[4] H. Hanson. Comparison of leakage energy reduction techniques. Technical Report TR-01-18, Computer Sciences Department, University of Texas at Austin, June 2001.

[5] M. Horowitz, R. Ho, and K. Mai. The future of wires. In *Semiconductor Research Corporation Workshop on Interconnects for Systems on a Chip*, May 1999.

[6] International technology roadmap for semiconductors, 2000 update, overall technology roadmap characteristics, 2000. http://public.itrs.net/Files/2000UpdateFinal/ORTC2000final.pdf.

[7] S. Kaxiras, Z. Hu, and M. Martonosi. Cache-line decay: Exploiting generational behavior to reduce leakage power. In *The 28th Annual International Symposium on Computer Architecture*, pages 240–251, July 2001.

[8] H. Makino, Y. Tujihashi, K. Nii, C. Morishima, Y. Hayakawa, T. Shimizu, and T. Arakawa. An auto-backgate-controlled MT-CMOS circuit. In *Symposium on VLSI Circuits*, pages 42–43, 1998.

[9] T. McPherson, R. Averill, D. Balazich, K. Barkley, S. Carey, Y. Chan, Y. Chan, R. Crea, A. Dansky, R. Dwyer, A. Haen, D. Hoffman, A. Jatkowski, M. Mayo, D. Merrill, T. McNamara, G. Northrop, J. Rawlins, L. Sigal, T. Slegel, and D. Webber. 760 mhz g6 s/390 microprocessor exploiting multiple V_t and copper interconnects. In *International Solid-State Circuits Conference*, pages 96–97, 2000.

[10] K. Nii, H. Makino, Y. Tujihashi, C. Morishima, Y. Hayakawa, H. Nunogami, T. Arakawa, and H. Hamano. A low power SRAM using auto-backgate-controlled MT-CMOS. In *International Symposium on Low Power Electronics and Design*, pages 293–298, 1998.

[11] M. Powell, S.-H. Yang, B. Falsafi, K. Roy, and T. Vijaykumar. Gated-V_{dd}: A circuit technique to reduce leakage in deep-submicron cache memories. In *International Symposium on Low Power Electronics and Design*, pages 90–95, 2000.

[12] G. Reinman and N. Jouppi. An integrated cache timing and power model, 1999. Unpublished document.

[13] K. Roy. Leakage power reduction in low-voltage CMOS designs. In *International Conference on Electronics, Circuits and Systems*, pages 167–73, 1998.

[14] S.-H. Yang, M. D. Powell, B. Falsafi, K. Roy, and T. Vijaykumar. An integrated circuit/architecture approach to reducing leakage in deep-submicron high-performance caches. In *International Symposium on High-Performance Computer Architecture*, pages 147–157, 2001.

[15] H. Zhou, M. C. Toburen, E. Rotenberg, and T. M. Conte. Adaptive mode-control: A static-power-efficient cache design. In *International Conference on Parallel Architectures and Compilation Techniques*, 2001.