Yield-driven Near-threshold SRAM Design

Gregory K. Chen, David Blaauw, Trevor Mudge, Dennis Sylvester
Department of EECS
University of Michigan
Ann Arbor, MI 48109
grgkchen@umich.edu, blaauw@umich.edu, tnm@umich.edu, dennis@umich.edu

Nam Sung Kim
Microprocessor Technology Lab
Intel Corporation
Hillsboro, OR 97124
nam.sung.kim@intel.com

***Abstract*—Voltage scaling is desirable in SRAM to reduce energy consumption. However, commercial SRAM is prone to functional failures when V_{dd} is scaled. Several SRAM designs scale V_{dd} to 200-300mV to minimize energy per access, but these designs do not consider SRAM robustness, limiting them to small arrays and sensor type applications. We examine the effects on area and energy for a differential 6T, single-ended 6T with power rail collapsing and an 8T bitcell as V_{dd} is scaled and the bitcells are sized appropriately to maintain robustness. SRAM robustness is examined using importance sampling to reduce simulation runtime. At high voltages, the differential 6T bitcell is the smallest for the same failure rate, but the 8T bitcell is smaller when V_{dd} is scaled below 450mV. For V_{dd} below V_{th}, bitcells must be sized greatly to retain robustness and large arrays become impractical. The differential 6T and 8T designs have the lowest dynamic energy consumption, and the single-ended 6T design has the lowest leakage. The supply voltage for minimum energy operation depends on cache configuration and can be well above V_{th} for large caches with low dynamic activity.**

I. Introduction

Reducing energy consumption is desirable in modern microprocessors to enable longer battery life and adequate heat dissipation. A simple and effective way to reduce energy usage is to scale supply voltage. This delivers a quadratic savings in dynamic energy consumption with a delay degradation [1] [2] [3]. As shown in Figures 1 and 2, as V_{dd} is scaled down to the near-threshold regime, delay degrades gracefully and energy per computation is significantly reduced [1][2]. As V_{dd} is scaled further, delay increases dramatically and total energy per cycle increases because of increased leakage. There exists a supply voltage where total energy per cycle is minimized (V_{min}) that is often near-threshold for SRAM.

CMOS circuitry is quite resilient when subjected to voltage scaling. However, Static Random Access Memory (SRAM) is more prone to functional failures at lower V_{dd}. Based on our tests, typical memory fails at $\frac{2}{3}$ of the nominal supply. Reduction in static noise margin (SNM), shown in Figure 3a, is further evidence of low reliability at lower V_{dd}. A smaller amount of noise can cause the bitcell's state to flip in the reduced voltage case.

Numerous ultra-low energy SRAMs have been implemented to minimize energy usage by reducing V_{dd} to below the threshold voltage (V_{th}) [4][5][6]. V_{th} is 400mV and the nominal V_{dd} is 1.2V for the 0.13μm CMOS technology in our study. A single-ended 6T SRAM has been demonstrated which is functional at 200mV with a 40% area penalty [4].

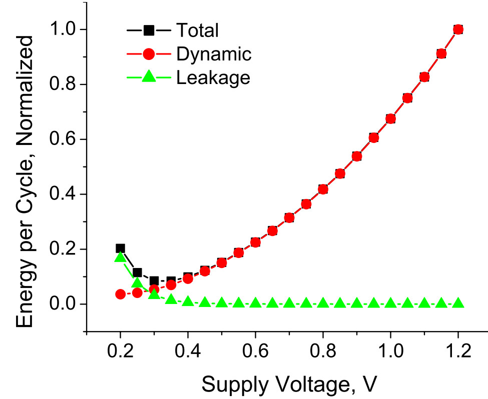


Fig. 1. Energy impact of supply voltage scaling.

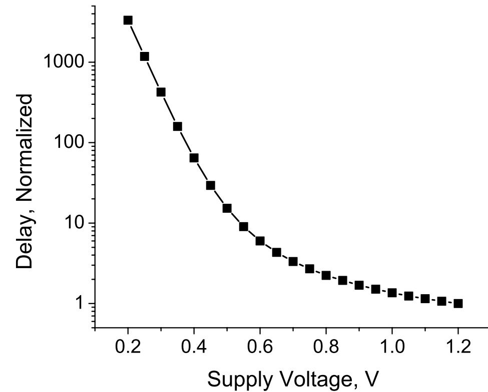
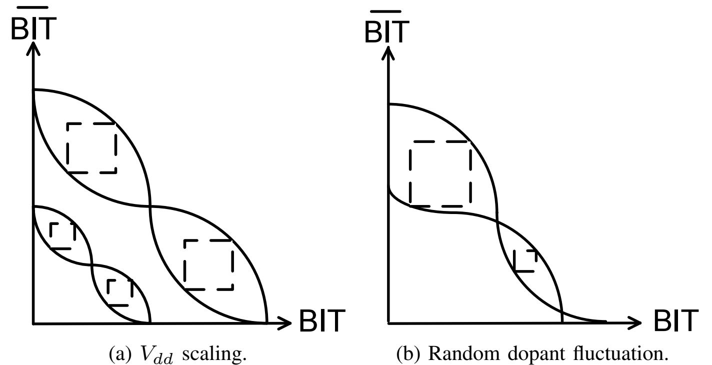
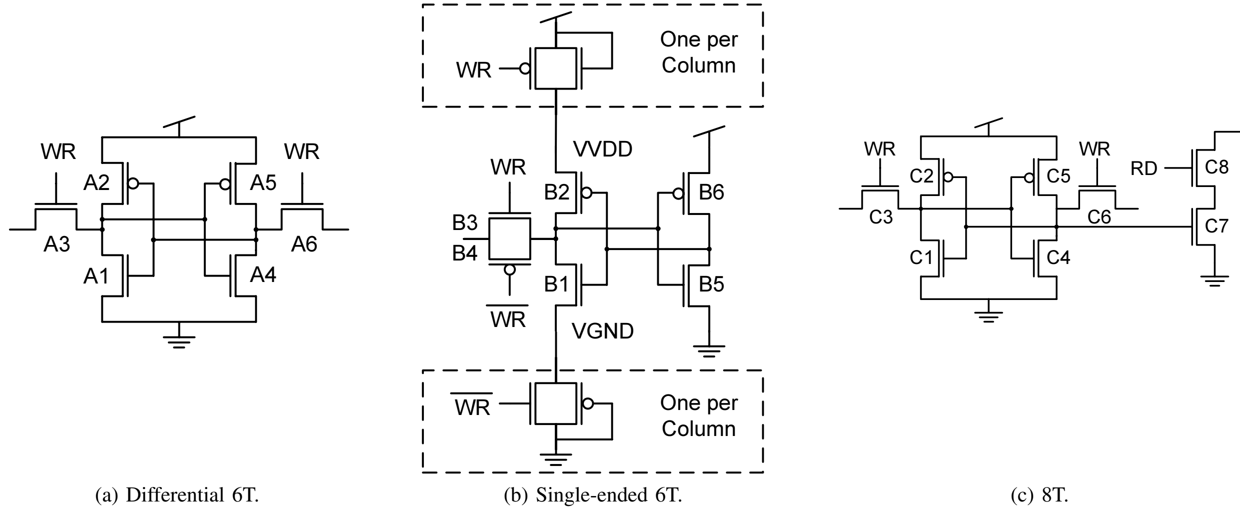


Fig. 2. Delay impact of supply voltage scaling.



(a) V_{dd} scaling. (b) Random dopant fluctuation.

Fig. 3. Static noise margin reduction.



(a) Differential 6T.

(b) Single-ended 6T.

(c) 8T.

Fig. 4. Candidate architectures.

Another proposed low supply voltage solution is an 8T bitcell which isolates the read and write paths. A typical 8T bitcell is more than 33% larger than a differential 6T bitcell [5]. Another SRAM design uses a multiplexer tree to read out data values and thus improve read stability [6]. These ultra-low energy SRAMs are not designed for robustness and would have very low yield for main-stream commercial designs where SRAM sizes reach megabytes. This limits the designs to small arrays and sensor type applications.

We model process variation, especially random dopant fluctuation (RDF), using importance sampling to calculate SRAM robustness. RDF shifts the V_{th} of each transistor independently, producing mismatch within bitcells and greatly reducing SNM, as shown in Figure 3b. The effects of process variation may be tested through either SNM measurement, corner case analysis, Monte Carlo simulation or analytical modeling [7]. However, SNM analysis does not consider the dynamic nature of noise injection. Corner case analysis is pessimistic, resulting in over optimized bitcells and requiring unnecessary area and power. Monte Carlo simulation is computationally expensive when the acceptable failure rate is low, as in the case of SRAM. Alternatively, we calculate SRAM robustness using importance sampling. We sample more heavily near the failure region, reducing the number of necessary samples [8]. The resulting samples are weighted to produce the natural probability of failure [9].

Using importance sampling, we analyze a 6T, a single-ended 6T with power rail drooping and an 8T bitcell at an iso-robustness condition. While the bitcell architectures are known, they have not been thoroughly compared in robustness, area and energy. Each bitcell is simulated across V_{dd} at single and twenty-cycle latency specifications. The single-cycle latency is relevant for L1 caches, and the twenty-cycle latency is relevant for L2 caches. In each simulation, the candidate SRAM bitcell is sized to maintain robustness as V_{dd} is scaled.

We show the areas and energies of the three candidate SRAMs at an iso-robustness and iso-delay condition. The differential 6T is the smallest architecture at high V_{dd}, but the 8T becomes smaller at 450mV for the twenty-cycle case. As V_{dd} approaches V_{th}, bitcells must be sized greatly unless delay is relaxed, making large arrays impractical. For V_{dd} above 400mV, the differential 6T bitcell has the lowest dynamic energy consumption. The 8T dynamic energy is lowest of the architectures below 400mV and is similar to the dynamic energy of the differential 6T throughout the near-threshold range. Leakage per cycle is lowest in the single-ended 6T bitcell and highest in the 8T bitcell. Previously V_{min} for sub-threshold SRAM has been reported as 350mV for the single-ended 6T and 450mV for the mux memory [4][6]. However, at an iso-robustness condition, V_{min} is often substantially higher because bitcells are sized considerably at low voltage, increasing leakage energy.

The rest of this paper is organized as follows: in Section II we discuss the topology and operation of the candidate architectures; in Section III we examine the simulation setup and importance sampling methodology; in Section IV we present our results, in Section V we conclude.

II. Candidate Architectures

A. Differential 6T SRAM

In this study we compared a differential 6T bitcell to a single-ended 6T and an 8T bitcell, both of which have been presented as solutions for low supply voltage SRAM stability [4][5]. For the differential 6T bitcell in Figure 4a, a read is performed by precharging the bitlines attached to A3 and A6 and asserting WR, allowing the bitcell to drive the bitlines. A write is performed by driving opposite values onto the bitlines and asserting WR, overwriting the value held in the bitcell.

B. Single-ended 6T SRAM

The single-ended design is constructed by removing one access transistor from the differential 6T bitcell and converting

the other access transistor into a transmission gate, as shown in (Figure 4b. During write, the transmission gate is turned on and the power supply of B1 and B2 is drooped using an NMOS header and PMOS footer. Voltage drooping weakens the transistors which hold the bitcell's state, allowing the value on the bitline to overwrite the stored value. On a read, the bitline is precharged to V_{dd}, and the bitcell is allowed to drive the bitline. Since the read is one-sided, only B1 must be sized for read stability. However there is no differential output, so a differential sense amp can not be used.

C. 8T SRAM

The 8T bitcell shown in Figure 4c duplicates the read-out path with C7 and C8, thus isolating the read line from the value holding nodes [5]. This makes the 8T bitcell more robust to read upset failures, which is the primary failure mode of the differential 6T. However, the 8T bitcell requires an area overhead for extra transistors and cannot use differential sense amps. A write for the 8T bitcell is performed in the same way as in the differential 6T bitcell.

III. Iso-Robustness Scaling Analysis

The candidate SRAM bitcells were simulated at supply voltages in the near-threshold space and constrained to single-cycle and twenty-cycle latencies. The bitcells were sized until they matched the robustness of a differential 6T bitcell at 800mV. This bitcell was chosen for comparison because it showed adequate reliability [4]. Throughout the study, importance sampling was used to calculate the SRAM failure rates. In the following sections we will expand on importance sampling and the SRAM simulation and sizing methodology.

A. Importance Sampling

Importance sampling is a variance reduction technique where the area of interest is oversampled so it can be more accurately modeled. The importance samples are weighted appropriately to generate the same result as Monte Carlo sampling methods, but with reduced simulation runtime.

In Monte Carlo sampling, the number of passing bitcells is divided by the total number of iterations (n) to find the expected yield, as shown in Equation 1 [9][8][10]. Process parameters such as V_{th} and gate length, are selected from a probability density function (PDF), which represents the natural variation in the process parameter. As shown in Figure 5, the natural PDF of V_{th} in SRAM devices is modeled as a normal distribution.

Since caches contain many bitcells, the failure rate of each one must be very low in order to have high yield for the cache. For example, to have a 99% yield for a 16kB cache, the bitcell failure rate must be 6.28×10^{-7}. To calculate these low yields using Monte Carlo, a very large number of simulations must be performed, making this procedure computationally intensive. As shown in Equations 2 to 5, the importance sampling technique chooses a new sampling PDF for each transistor so that more failures are simulated. For our study, the V_{th} of each transistor is shifted by 4σ in either

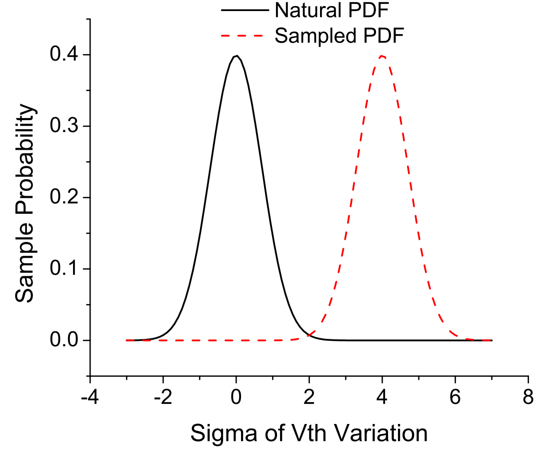


Fig. 5. V_{th} PDFs for Monte Carlo and importance sampling.

direction to introduce worst-case mismatch into the bitcell. The importance samples are then weighted by the ratio of the natural probability of the large V_{th} shift occurring in each transistor to the probability that these V_{th} shifts were sampled. The average of the weighted samples generates the probability that the bitcell will pass.

$$Y = \frac{1}{n}\sum_{n} f(x) \ where \ f(x) = \begin{cases} 1, & pass \\ 0, & fail \end{cases} \tag{1}$$

$$Y = \frac{1}{n}\sum_{g(x)} \frac{p(x)}{g(x)} \tag{2}$$

$$p(x) = \prod \frac{1}{\sigma\sqrt{2\pi}} exp\frac{(x-\mu)^2}{2\sigma^2} f(x) \tag{3}$$

$$g(x) = \prod \frac{1}{\sigma\sqrt{2\pi}} exp\frac{(x-\mu+4\sigma)^2}{2\sigma^2} \tag{4}$$

$$Y = \frac{1}{n}\sum_{g(x)} \frac{\prod \frac{1}{\sigma\sqrt{2\pi}} exp\frac{(x-\mu)^2}{2\sigma^2} f(x)}{\prod \frac{1}{\sigma\sqrt{2\pi}} exp\frac{(x-\mu+4\sigma)^2}{2\sigma^2}} \tag{5}$$

Since the sampling PDF and number of importance samples have large effects on experimental results, they were carefully chosen to maintain accuracy in the simulation while still reducing simulation runtime. A small V_{th} shift in the sample PDF would not introduce a large number of failures, thus negating the variance reduction effect of importance sampling. Oppositely, an overly large V_{th} shift introduces failures, but causes the sample weighting to be small and reduces the accuracy of the simulation. An 800mV differential 6T bitcell was studied to find the optimal sampling PDF. As shown in Figure 6, with less than a $4\sigma V_{th}$ shift the sample failure rate is very low. Above a $4\sigma V_{th}$ shift the calculated failure rate drops and is inaccurate. Therefore, a 4σ shift was chosen for our study. After a sufficient number of importance samples have been performed, the calculated failure rate converges to its final value. From Figure 7 we determine that 20,000 samples

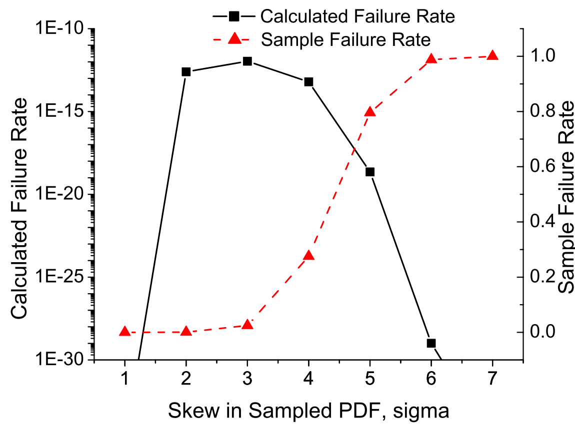


Fig. 6. Accuracy and variance reduction for V_{th} shift in sample PDF.

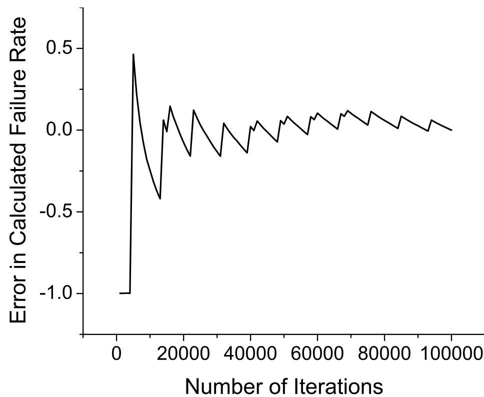


Fig. 7. Calculated failure rate convergence.

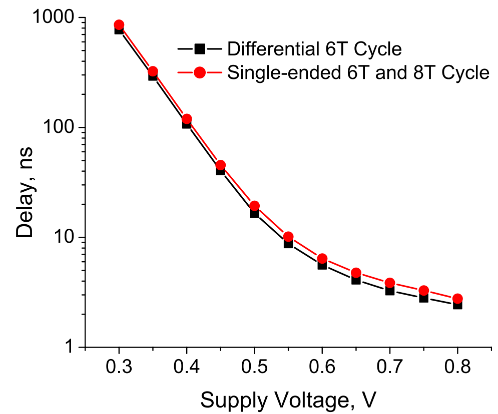


Fig. 8. Delay specifications of SRAM bitcells.

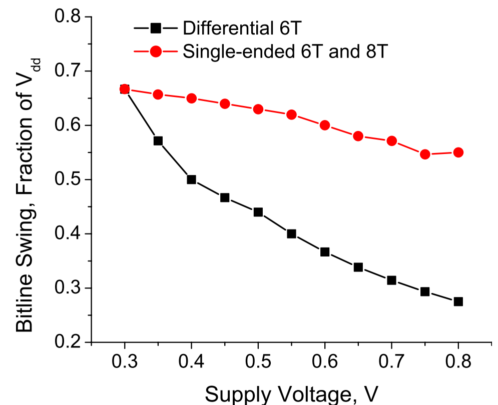


Fig. 9. Minimum sensible bitline voltage swing.

is sufficient for accurate results. To measure these failure rates with Monte Carlo, at least one trillion samples are needed.

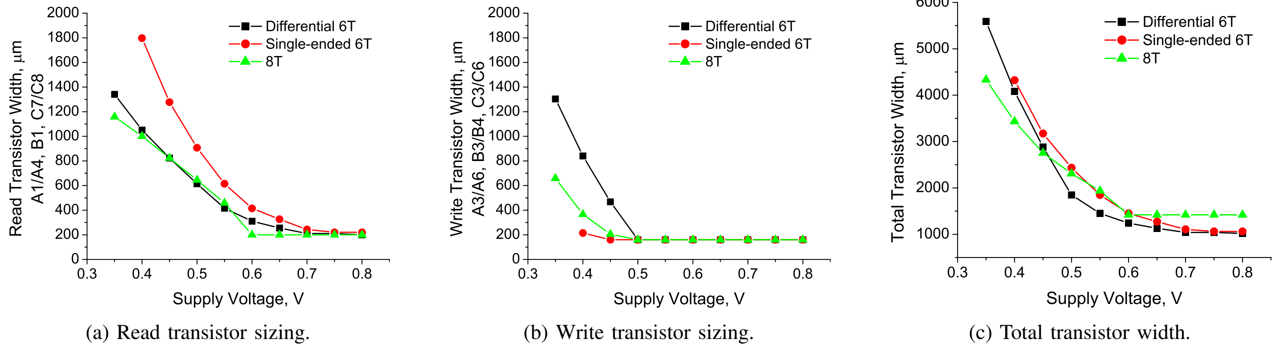
B. Sizing Optimization

We require that all bitcell and V_{dd} combinations have the same failure rate as a differential 6T bitcell at 800mV. We also require the delay to scale with V_{dd} as CMOS delay scales. Sense amplifiers or other read-out logic are considered to create iso-delay read-out paths of each architecture. Hold failures were demonstrated to have a much lower occurrence than other failure modes and are not included in the importance sampling study. At each V_{dd}, the appropriate transistors in each design are sized until the iso-robustness condition is met.

1) Read Upset: During a read operation on a differential 6T bitcell, noise is injected from the bitline through the access transistors to the node holding a zero value. Read upset occurs when the pull down device is not able to hold the node to the correct value. Therefore read upset tolerance depends heavily on the on-current ratio of the pull down to access transistors. The single-ended 6T bitcell may be sized asymmetrically, reducing the area overhead needed to prevent read upset. In addition, the bitcell may be sized with less concern for write stability since the write is stabilized by supply rail drooping. The 8T bitcell duplicates the read-out path, incurring an area penalty but blocking the path for noise injection to the value holding nodes. To meet the read upset failures condition, transistors A1/A4, B1, and C7 may be sized up in the differential 6T, single-ended 6T and 8T bitcells in Figures 4a to 4c, respectively.

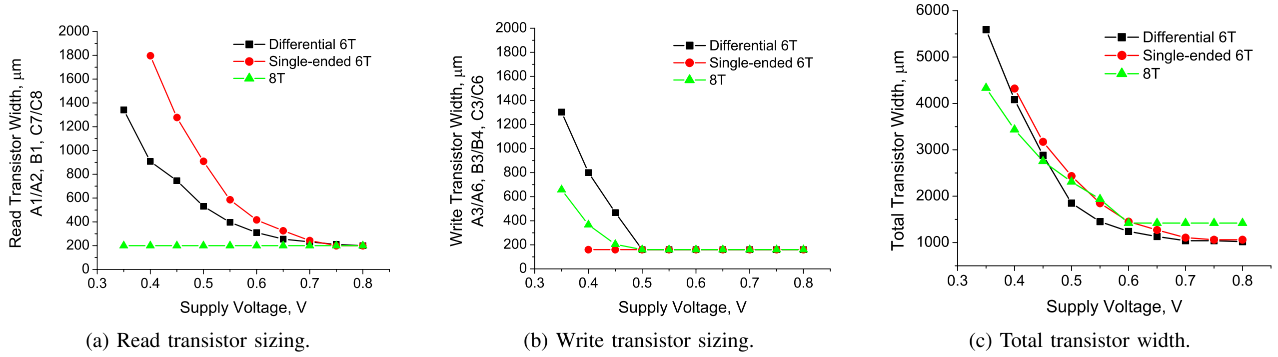
2) Read Access: In addition to retaining the correct value, the SRAM must be able to read out the value at the required performance specification. Since we wish to design a single-cycle L1 cache and twenty-cycle L2 cache, we require that the bitcell plus read-out logic delay scales with V_{dd} with a 64-bit Alpha processor. For the differential 6T bitcell a sense amp can be used. However, since the other architectures have single-ended reads CMOS logic is used for read-out. Monte Carlo simulation was used to analyze the minimum sensible voltage and delay for each type of read-out logic. The specifications used to constrain the bitcells are shown in Figures 8 and 9. To meet read access failures condition, transistors A1/A4, B1/B3/B4, and C7/C8 may be sized up.

3) Write Access: Write simulations require adequate access transistor strength to overwrite the value held in the bitcell. This is the opposite requirement of read upset prevention. For this reason, at lower voltages, 6T bitcells may have to be sized greatly to achieve both read and write stability or may not be able to achieve both. The other bitcell architectures attempt



(a) Read transistor sizing. (b) Write transistor sizing. (c) Total transistor width.

Fig. 10. Sizing for single-cycle delay.



(a) Read transistor sizing. (b) Write transistor sizing. (c) Total transistor width.

Fig. 11. Sizing for twenty-cycle delay.

to remove the conflict between read and write stability. The single-ended design droops the power rails, stabilizing write accesses allowing the bitcell to be sized optimally for read. In the 8T bitcell, the read path is separate and may be sized up without impairing write functionality. To meet the write failure condition, transistors A3/A4, B3/B4, C3/C6 may be sized up.

IV. Results

A. SRAM Bitcell Sizing for Single-cycle Latency

Transistor sizings for iso-robustness operation at a single-cycle latency constraint are shown in Figures 10a and 10b. As V_{dd} is scaled for the differential 6T, the pull down device must to sized to prevent read upset. This weakens write stability necessitating access transistor sizing below 500mV. The pull down device must be sized more aggressively in the single-ended 6T design since a transmission gate connects the bitcell with the bitline, increasing the amount of noise injection. The read devices in the 8T design must be sized to meet the delay constraint below 600mV. The differential 6T and 8T designs have the lowest total transistor width above and below 450mV, respectively, as shown in Figure 10c.

B. SRAM Bitcell Sizing for Twenty-cycle Latency

Twenty-cycle caches can be used for L2 caches or in designs where memory latency is not as critical as throughput. Since the differential and single-ended 6T bitcells areas are limited by read upset and write failures, their transistor sizings are similar as the delay specification is relaxed, as seen in Figures 11a and 11b. When the delay constraint is relaxed, the read-out path of the 8T bitcell does not have to be sized. For this reason the 8T bitcell has the lowest total transistor width below 550mV, as seen in Figure 11c.

C. SRAM Bitcell Area

Using the information on transistor widths, we calculate the layout area from previous layouts of each SRAM architecture [4][5][11]. All three bitcells are laid out using logic DRC rules. The areas are then normalized to the bitcell area of a minimum-sized commercial differential 6T bitcell. The 8T bitcell is larger than the differential 6T bitcell for the same transistor width because of additional contacts. As shown in Figures 12 and 13, the differential 6T is smallest except for the twenty-cycle sub-450mV case where the 8T is smallest. For a single-cycle latency, all bitcells must be sized significantly as V_{dd} approaches V_{th}, making iso-robustness SRAM impractical. For a twenty-cycle latency at 350mV, the 8T bitcell gives a 16% reduction in area over the differential 6T design.

D. Energy Impact of Supply Voltage Scaling in SRAM

As was shown in Figure 1, for each circuit there exists a V_{dd} where total energy consumption per cycle is minimized (V_{min}). This voltage is highly dependent on the ratio of dynamic to leakage energy. For caches, V_{min} increases as the size and delay of the cache increase and as associativity, cache

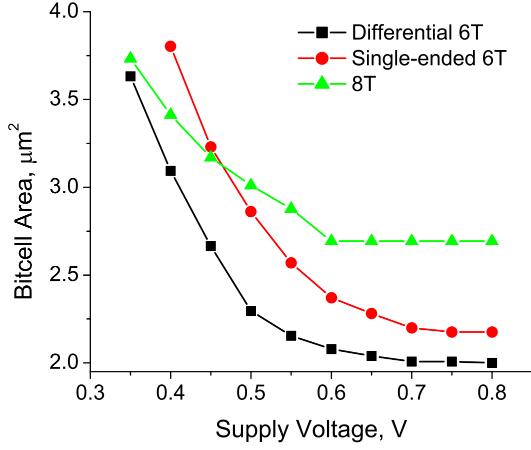


Fig. 12. SRAM bitcell area for single-cycle delay.

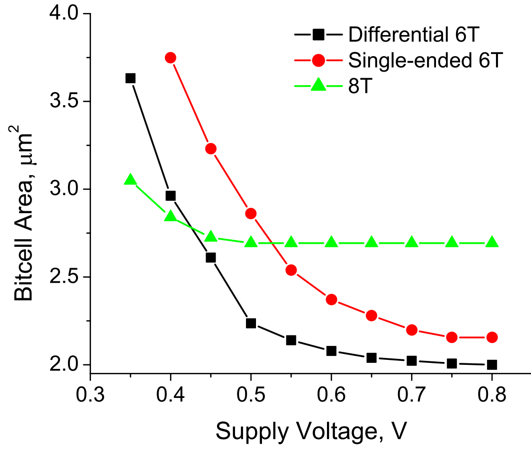


Fig. 13. SRAM bitcell area for twenty-cycle delay.

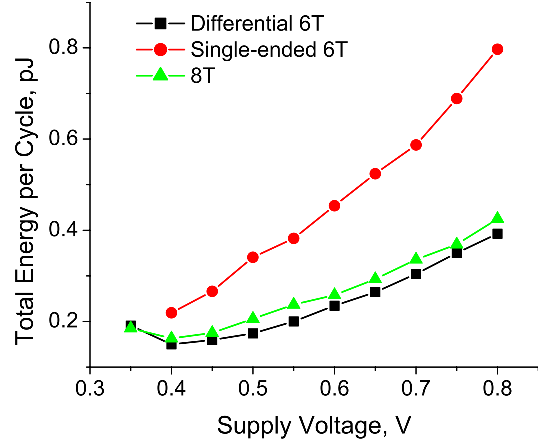


Fig. 14. Energy of a 1kB, direct-mapped, single-cycle L1 cache.

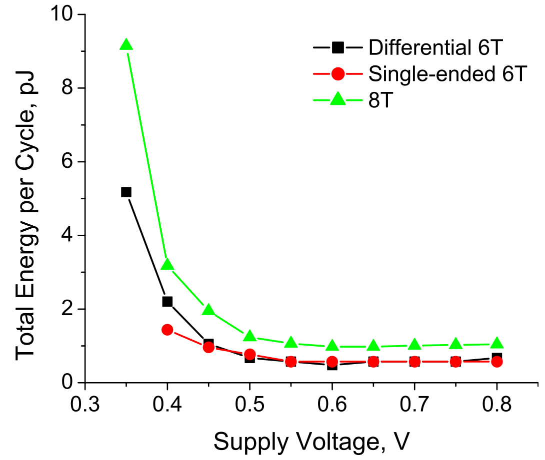


Fig. 15. Energy of a 64kB, 8-way associative, twenty-cycle L2 cache.

line length and access rate decrease. Delay, and thus leakage, increases as the square root of bank size because of added capacitance. Increasing the cache line length, number of ways and access rate all increase the dynamic activity of the cache. These observations allow us to establish a relationship between V_{min} and these cache parameters, shown in Equation 6.

$$V_{min} \propto \frac{Number\ of\ Leaking\ Bitcells \times Delay}{Number\ of\ Accessed\ Bitcells} \propto \frac{Size \times \sqrt{Bank\ Size}}{Ways \times Access\ Rate \times Cache\ line} \quad (6)$$

For small caches, where the value in Equation 6 is low, dynamic energy consumption dominates total energy consumption. As seen in Figure 14, the V_{min} for a 1kB, direct mapped L1 cache is near 400mV. The differential 6T bitcell has the lowest energy consumption above 400mV. The single-ended 6T has the highest dynamic energy consumption because it has transitions on two write-lines for read and write, added bitline cap due to transmission gates, and transitions of the virtual power supplies during a write. Although the 8T bitcell has two word lines and three bitlines, they do not all transition for each access so the dynamic energy consumption is relatively low. Below 400mV the 8T is energy optimal because of smaller transistor sizes for equal robustness.

For iso-robustness SRAM, V_{min} may be much higher than previous reported because at low V_{dd} the bitcells are larger and leakier than in previous designs. For the single-ended 6T and mux-memory, the reported V_{min} values are 350mV and 450mV, respectively [4][6]. In our study, V_{min} can be as high as 600mV for common cache configurations where the cache size is large and dynamic activity is low. For the 64kB, 8-way associative L2 cache in Figure 15, leakage energy dominates total energy. There is little change in the energy consumption as the caches are scaled down to the V_{min} at 600mV. The single-ended 6T bitcell has the lowest energy and leakage of the candidate architectures. The 8T bitcell energy is the highest because it has additional transistors and is the leakiest.

In Figure 16, V_{min} results for a variety of common cache configurations are plotted against the ratio in Equation 6 for a differential 6T bitcell. For all the cache configurations tested, the V_{min} of L1 caches was below 450mV, and the V_{min} of L2 caches was between 550mV and 600mV.

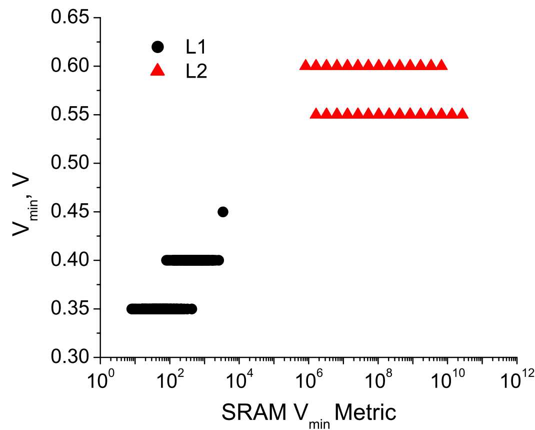


Fig. 16. Effect of cache parameters on Vmin.

V. Conclusion

Using importance sampling for yield estimation, we have compared a 6T, single-ended 6T with power rail drooping and an 8T bitcell at an iso-robustness and iso-delay condition. Importance sampling was successfully used to reduce the runtime of yield estimation. At higher V_{dd}, the differential 6T bitcell is the smallest. The 8T bitcell becomes smaller below a V_{dd} of 450mV at a twenty-cycle latency. As V_{dd} approaches V_{th}, all bitcells must be sized greatly to maintain robustness unless delay is relaxed, making large arrays impractical. The differential 6T bitcell has the lowest dynamic energy consumption at most supply voltages. The single-ended 6T bitcell has the lowest leakage per cycle. V_{min} increases with cache size and bank size, and decreases with associativity, activity factor and cache line length. For common cache configurations, V_{min} may be above V_{th} and significantly higher than previously reported. By comparing SRAM bitcells at an iso-robustness and iso-delay condition, the best SRAM architecture and sizing for a design can be quickly and accurately chosen.

References

[1] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner. "Theoretical and practical limits of dynamic voltage scaling." In *Design Automation Conference*, pp. 868–873, 2004.

[2] A. Wang and A. Chandrakasan. "A 180mV FFT processor using subthreshold circuit techniques." In *IEEE International Solid-State Circuits Conference*, vol. 1, pp. 292–529, 15-19 Feb.2004.

[3] N. Lindert, T. Sugii, S. Tang, and C. Hu. "Dynamic threshold pass-transistor logic for improved delay at lower power supply voltages." In *IEEE Journal of Solid-State Circuits*, vol. 34(1), pp. 85–89, Jan. 1999.

[4] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson. "A sub-200mV 6T SRAM in 0.13μm CMOS." In *IEEE International Solid-State Circuits Conference*, pp. 332–606, 11-15 Feb. 2007.

[5] L. Chang, D.M. Fried, J. Hergenrother, J.W. Sleight, R.H. Dennard, R.K. Montoye, et al. "Stable SRAM cell design for the 32nm node and beyond." In *Symposium on VLSI Technology*, pages 128–129, 14-16 June 2005.

[6] B.H. Calhoun and A. Chandrakasan. "A 256kb sub-threshold SRAM in 65nm CMOS." In *IEEE International Solid-State Circuits Conference*, pages 2592–2601, Feb. 6-9, 2006.

[7] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS." In *Computer-Aided Design of Integrated Circuits and Systems*, vol. 24(12), pp.1859–1880, Dec. 2005.

[8] P. Shahabuddin. "Importance sampling or the simulation of highly reliable Markovian systems." In *Management Science*, vol. 40.3(3), pp.333–352, 1994.

[9] R. Kanj, R. Joshi, and S. Nassif. "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events." In *Design Automation Conference*, pp. 69–72, 24-28 July 2006.

[10] P. Heidelberger. "Fast simulation of rare events in queueing and reliability models." In *ACM Modeling and Computer Simulation*, vol. 5.1, pp. 43–85, 1995.

[11] F. Arnaud, B. Duriez, B. Tavel, L. Pain, J. Todeschini, M. Jurdit, et al. "Low cost 65nm CMOS platform for low power & general purpose applications." In *Symposium on VLSI Technology*, pp. 10–11, 15-17 June 2004.