Yield-Driven Near-Threshold SRAM Design

Gregory Chen, *Student Member, IEEE*, Dennis Sylvester, *Senior Member, IEEE*, David Blaauw, *Senior Member, IEEE*, and Trevor Mudge, *Fellow, IEEE*

***Abstract*—Voltage scaling is desirable in static RAM (SRAM) to reduce energy consumption. However, commercial SRAM is susceptible to functional failures when V_{DD} is scaled down. Although several published SRAM designs scale V_{DD} to 200–300 mV, these designs do not sufficiently consider SRAM robustness, limiting them to small arrays because of yield constraints, and may not correctly target the minimum energy operation point. We examine the effects on area and energy for the differential 6T and 8T bit cells as V_{DD} is scaled down, and the bit cells are either sized and doped, or assisted appropriately to maintain the same yield as with full V_{DD}. SRAM robustness is calculated using importance sampling, resulting in a seven-order run-time improvement over Monte Carlo sampling. Scaling 6T and 8T SRAM V_{DD} down to 500 mV and scaling 8T SRAM to 300 mV results in a 50% and 83% dynamic energy reduction, respectively, with no reduction in robustness and low area overhead, but increased leakage per bit. Using this information, we calculate the supply voltage for a *minimum total energy* operation (V_{MIN}) based on activity factor and find that it is significantly higher for SRAM than for logic.**

***Index Terms*—Low power, near threshold, robustness, static RAM (SRAM), threshold voltage tuning.**

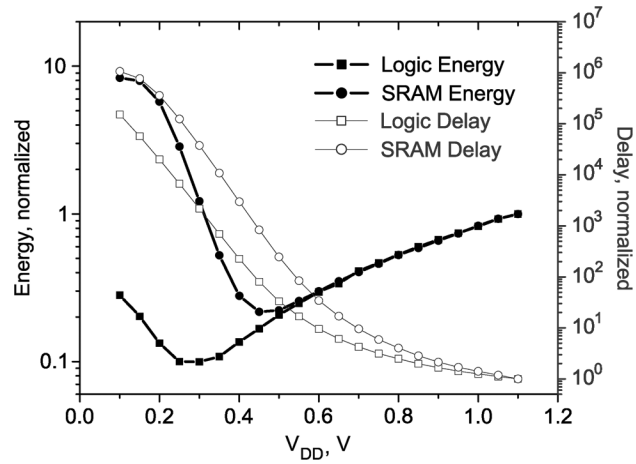

Fig. 1. Voltage scaling quadratically reduces dynamic energy and linearly reduces leakage power. Since delay increases exponentially, leakage energy per instruction increases and dominates total energy at low V_{DD}. These competing trends result in a V_{DD} that minimizes total energy per instruction, denoted by V_{MIN}.

I. Introduction

REDUCTION of energy consumption is desirable in microprocessors to enable longer battery life and adequate heat dissipation. A simple and effective way to reduce energy is to scale down supply voltage. This delivers a quadratic saving in dynamic energy consumption and a linear reduction in leakage power [1]–[3]. As shown in Fig. 1, as V_{DD} is scaled down into the near-threshold region, between 400 and 700 mV, the energy per operation is significantly reduced and delay degrades gracefully [1], [2]. As V_{DD} is scaled further, delay increases dramatically, and total energy per cycle increases because leakage energy dominates. Leakage energy per computation increases as V_{DD} is scaled down, even though leakage power decreases, since it is proportional to delay, which increases exponentially in the subthreshold region. There exists a supply voltage where the total energy per operation is minimized (V_{MIN}). V_{MIN} heavily depends on the ratio of dynamic-to-leakage energy for the circuit. Compared to combinational logic, which commonly has subthreshold V_{MIN}, caches have more idle circuitry and a lower activity rate. This increases the ratio of leakage to the dynamic energy and subsequently increases V_{MIN} into the near-threshold region for common cache configurations. In this paper, we target SRAM designs to robustly operate near V_{MIN} in the near-threshold region.

As V_{DD} is scaled down, the ON/OFF current ratio for devices is reduced and the noise margins diminish. Typically, CMOS circuitry maintains adequate robustness in the presence of these effects. However, static RAM (SRAM) becomes more prone to functional failures at low V_{DD}, as evidenced by the reduction in the static noise margin (SNM) shown in Fig. 2(a) [4]. In addition, at low V_{DD}, bit cells are more susceptible to V_{TH} variation caused by random dopant fluctuation (RDF). RDF shifts the V_{TH} of each transistor independently, causing mismatch within bit cells and greatly reducing the SNM, as shown in Fig. 2(b). When the SRAM V_{DD} is scaled down, these V_{TH} shifts have a greater impact on device currents. At near-threshold supply voltages, RDF is the dominant form of process variation, and the foremost reason for poor robustness in subthreshold and near-threshold SRAM. Making SRAM transistors larger can increase the SRAM robustness since nonuniformities in channel doping average out, resulting in more uniform device V_{TH}s [5]. The cost of increased device sizing is larger SRAM area and higher energy.

One proposed solution for near-threshold SRAM is the 8T bit cell [6]. The 8T bit cell connects two additional stacked negative-channel FETs (NFETs) to the differential 6T structure to isolate the read and write accesses. A separate read wordline (WL) and read bitline (BL) are employed to perform a single-ended read on the bit cell with no risk of upsetting the bit cell's value. The separate read structure allows the other six

Manuscript received October 27, 2008; revised March 01, 2009. First published September 15, 2009; current version published October 27, 2010. This work was supported in part by the National Science Foundation.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: grgkchen@umich.edu; dmcs@umich.edu; blaauw@umich.edu; tnm@umich.edu).

Digital Object Identifier 10.1109/TVLSI.2009.2025766

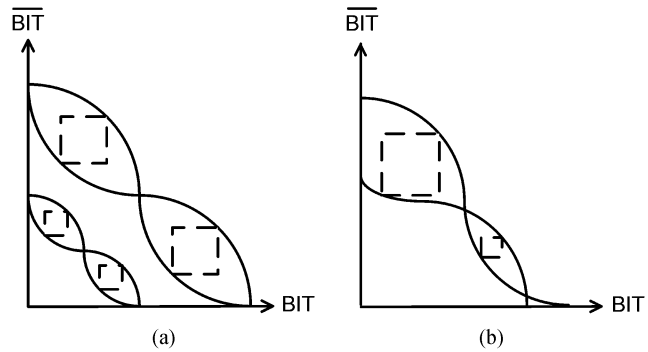

Fig. 2. Voltage scaling reduces SNM. (a) Ideal SNM scales with $V_{\rm DD}$. (b) RDF induced $V_{\rm TH}$ variation causes mismatch in SRAM bit cells, reducing robustness and making the SNM smaller. At near threshold, SRAM robustness is more sensitive to $V_{\rm TH}$ variation because drain current is more sensitive to gate overdrive.

devices to be sized and doped appropriately to ensure write stability. A typical 8T bit cell is over 33% larger than a differential 6T bit cell, but may have higher array efficiency [6]. Further solutions for high robustness SRAM use read and write assist circuits [7]–[11]. These circuits modulate the WL, BL, or supply voltages to prevent functional failure. These techniques have the advantage of keeping bit cell density high, but may require extra overhead such as additional peripheral devices or voltage sources.

Numerous ultralow energy SRAMs reduce energy by scaling $V_{\rm DD}$ to subthreshold levels [11]–[15]. A single-ended 6T SRAM has been demonstrated, which is functional below 200 mV with a 40% area penalty [11]. A multiplexer tree can be used to read data values and improve read stability [12]. A 10T bit cell was designed with assist circuitry to improve BL sensing [13]. Incorporating a Schmitt trigger into the cross-coupled inverters can prevent read failures and improve hold margins [14]. Also, multiple-$V_{\rm TH}$ designs exist for improving robustness and reducing leakage [15]. Many of these ultralow energy SRAMs exhibit insufficient robustness for commercial designs, where SRAM sizes reach megabytes, limiting them to small arrays and sensor applications. Also, increased delay, and thus leakage, for these architectures may cause subthreshold operation to be suboptimal for minimum energy operation.

In our study, we take a new look at the existing differential 6T and 8T bit cell architectures by thoroughly comparing the two designs in robustness, area, delay, and energy in the superthreshold and near-threshold voltage regimes, in order to explore energy savings through voltage scaling [6]. In our study, we constrain all bit cells at all $V_{\rm DD}$s to have equal robustness to a commercial differential 6T at a 1-V supply. As $V_{\rm DD}$ is scaled down, either the bit cell doping and sizing will be adjusted or assist circuits will be employed to meet these constraints.

To calculate robustness, we model RDF-induced random process variation. The effects of process variation may be measured through either SNM measurement, corner case analysis, Monte Carlo simulation, or analytical modeling [4], [16]. However, SNM analysis does not consider the dynamic nature of noise injection. Corner case analysis is pessimistic, resulting in over-optimized bit cells and unnecessary area and power. Monte Carlo simulation is extremely computationally intensive for SRAM because the acceptable failure rate is low. Alternatively, we calculate the SRAM robustness using importance sampling. We sample heavily in the failure region of interest, reducing the number of samples needed to characterize the failure modes [17]. The resulting samples are weighted using device $V_{\rm TH}$ probabilities to calculate the SRAM yield [18]. Importance sampling allows us to accurately and efficiently calculate bit cell yield.

We find that halving supply voltages from 1 V to 500 mV for differential 6T bit cells halves dynamic energy with either a 40% area overhead or a 200× delay penalty for maintaining robustness. Halving supply voltages for 8T bit cells also halves dynamic energy with no area overhead and preserved cache latency. The 8T bit cell can be further scaled to 300 mV to cut dynamic energy by 83% with a negligible area overhead. Using this information, we find the $V_{\rm MIN}$ and energy at $V_{\rm MIN}(E_{\rm MIN})$. $V_{\rm MIN}$ can be as low as 300 mV for 8T L1 caches with high access rates, and as high as 950 mV for L2 caches with low access rates.

In this paper, we contribute a framework for selecting an appropriate SRAM architecture given a set of design constraints, including near-threshold robustness. For the first time, we show that $V_{\rm MIN}$ for SRAM is significantly higher than voltages targeted in previous designs, and hence guide the focus of new SRAM research for energy efficiency.

The rest of this paper is organized as follows. Section II discusses the topology and operation of the candidate architectures. Section III examines the simulation setup and importance sampling methodology. We present our results in Section IV and Section V concludes the paper.

II. Candidate SRAM Architectures

A. Differential 6T Bit Cell

For the differential 6T bit cell shown in Fig. 2(a), a read is performed by precharging and floating the BLs (BL and $\bar{BL}$) in the desired columns at $V_{\rm DD}$, and asserting the WL in the desired rows. The bit cell pulls down either BL or $\bar{BL}$, depending on the bit cell's state, and the voltage differential is detected using a sense amplifier. A write is performed by driving opposite values onto the BLs and asserting WL, overwriting the value held in the bit cell.

Bit cells are susceptible to four prominent failure modes: read upset, write, timing, and hold. During a read operation on a 6T bit cell, noise is injected from the BL through the pass gate transistors to the node holding a zero value. Read upset occurs when the voltage transient on the zero node causes the bit cell value to flip. Read upset tolerance heavily depends on the cell ratio (on-current ratio of the pull down to pass gate transistors) as well as the feedback from the cross-coupled inverters. Write stability requires adequate pass gate transistor strength to overwrite the value held in the bit cell. The most critical transistors for a write are the pass gate device connected to the BL at a ZERO value and the pull up PMOS holding the bitcell node to ONE. The requirements for both read and write stability place contradicting requirements on pass gate strength. For this reason, at lower voltages, 6T bit cells must be sized up substantially or

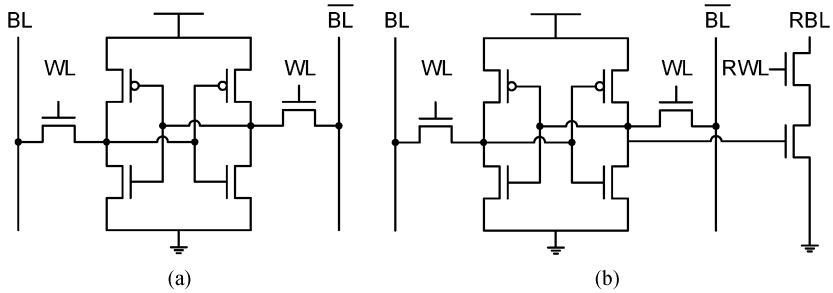

Fig. 3. Candidate bit cells: (a) differential 6T and (b) 8T.

doped differently to achieve both read and write stability, or they may not be able to achieve both.

B. SRAM Assist Circuits

As an alternative to sizing the 6T bit cell, assist circuits can be used to prevent failure. Read upset can be prevented by lowering the WL voltage in relation to the SRAM array V_{DD} [7]. This reduces the cell ratio of the bit cell, but increases delay, hurts write stability, and requires an additional voltage source. To prevent write failures, a dual-V_{DD} WL or additional write assist circuitry can be employed. In a dual-V_{DD} scheme, WL voltage is only reduced when a read access is performed. This requires additional decoding and a more complex WL driver to select between two WL voltages.

Dual-V_{DD} WL and other schemes have also been proposed to enhance write robustness. During a write operation, the WL voltage can be increased above the SRAM array V_{DD}, thereby increasing the effective pass gate strength [8]. Another way to increase pass gate strength and write stability is to pull the BL to a negative voltage [9]. The negative BL voltage must not turn on unaccessed devices on the same BL and must not cause intolerable junction leakage. Both the dual-V_{DD} and negative BL techniques require an additional voltage source. Another write assist method droops the SRAM array V_{DD} and GND during a write [10], [11]. This reduces the strength of the cross-coupled inverters that hold the bit cell state, facilitating write. Voltage drooping can be implemented with diode drops in shared headers and footers. The drooped supplies must be shared in a row or column, and unaccessed drooped bit cells must retain their state.

C. 8T Bitcell

The 8T in Fig. 3(b) uses two additional transistors over the differential 6T bit cell to isolate the read and write paths [6]. This enables separate optimization of the read and write mechanisms. The two stacked NFETs are connected to additional read word and read bitlines (RWL and RBL), as well as one bit cell node to perform a single-ended read. This read circuitry eliminates the read upset failure mode. A write operation is performed similarly to a write in the differential 6T bit cell; however, since the pass gate devices and cross-coupled inverters are not used for reading, they can be optimized solely for write.

The 8T bit cell has the same timing failure mode as the differential 6T. However, since the 8T read is single-ended, differential sense amplifiers cannot be used to improve delay and minimize RBL swing. For our study, we sense an 8T read using the same sense amplifier structure with one input tied to a reference voltage. The reference voltage must be sufficiently below V_{DD} to sense the read of a 1. This necessitates that the BL falls below the reference voltage to sense a 0, increasing the delay and BL swing for an 8T read. In our study, this delay must be recuperated by optimizing the stacked NFETs used for reading. Increasing the strength of these devices does not exacerbate other failure modes; however, it incurs area and energy penalties.

III. Reliability Analysis Using Importance Sampling

A. Scaling Methodology for Iso-Robustness, Low V_{DD} Operation

When SRAM bit cells are naively scaled into the near-threshold V_{DD} region, significant energy gains are achieved, but RDF and other process variations lead to functional failures and low yield. In our study, we examine the robustness of 6T and 8T SRAM in a 65-nm process when V_{DD} is scaled to the near-threshold region. We constrain all bit cells at all V_{DD}s to have the same robustness as the differential 6T bit cell at 1 V with sizes taken from commercial designs. To meet these constraints, as the bit cells are scaled into the near-threshold region, the bit cell device's geometry and dopings are optimized, or assist circuits are tuned. The delay, density, and energy of the final bit cells are compared to find the advantages and disadvantages of all designs.

B. Sizing and Doping Methodology

For our sizing and doping study, we adjust device strengths to prevent functional failure when V_{DD} is reduced. We constrain the bit cell delay to scale with logic, such that memory latency (in cycles) is not affected when V_{DD} is scaled. The WL driver, BL driver, and bit cell delays are monitored in this study. The bit cell delay for a read is measured at the time when adequate BL swing is developed for differential or single-ended sensing with a commercial current-mode sense amplifier.

In modern SRAM designs, V_{TH} is optimized separately from V_{TH} for logic to improve robustness and performance. The 65-nm process used in this study has a nominal V_{DD} of 1.1 V, and uses separate NFET V_{TH}s of 560 and 520 mV for the pass gate and pull down devices, respectively. These V_{TH}s are carefully chosen by manufacturers to enhance performance at nominal V_{DD}; however, as V_{DD} is scaled down, the criticality of failure modes, and thus the optimal V_{TH}s change. Optimizing V_{TH} can help SRAM meet delay requirements as well as control the current ratios between devices to balance probabilities of

different failure modes. When V_{TH} is tuned, σV_{TH} is calculated appropriately according to the device models. Circuit designers have limited flexibility to tune V_{TH}; therefore, in this study, we optimize the device geometry alone and also geometry with individual device V_{TH}s. Reasonable limits are placed on V_{TH} to ensure realistic doping concentrations and tolerable leakage power.

C. Assist Circuit Methodology

In our study of assist circuits, we maintain bit cell robustness as V_{DD} is scaled down by adjusting the peripheral circuits. Assist circuits are unnecessary for the 8T bitcell because there is no read upset failure mode, and write stability can be maintained with minimal sizing. The 6T bitcell design in our study is taken from a commercial design optimized for superthreshold operation, and no device sizing or V_{TH} tuning is performed. To maintain read stability, a dual-V_{DD} WL with reduced read voltage is used. This read assist circuit incurs a delay penalty, precluding isolatency voltage scaling, so there is no delay constraint for the assist circuit study. For write robustness, three methods will be compared: overdriven WL, negative BL, and supply rail drooping. In the latter two cases, the assist circuits must be adjusted appropriately not to disturb the unaccessed bit cells. The resulting decrease in bit cell performance and changes in energy consumption are measured.

D. Robustness Calculation Using Importance Sampling

At this point, an accurate metric of SRAM robustness is necessary to determine when optimization is complete. SRAM robustness is often measured using SNM because it is relatively easy to compute. However, SNM does not consider the dynamic nature of noise injection into bit cells. Since the probability of injecting the same amount of noise changes as V_{DD} is scaled, SNM does not translate directly to SRAM yield. Corner cases can also be used to measure robustness; however, in general, the supplied corner cases only consider global variation and not device mismatch. Since mismatch has a strong effect on the SRAM yield, these cases are not sufficient. Corner cases involving mismatch can be performed, but have several drawbacks. First, different transistors have differing criticality for SRAM functionality, but in corner case analysis the same amount of variation is placed on each device, making the analysis incomplete. Second, calculating the SRAM yield based on corner case simulations is nontrivial.

For a complete look at SRAM, reliability sampling methods like Monte Carlo are necessary. In Monte Carlo sampling, the number of passing bit cells is divided by the total number of iterations (n) to find the expected yield, as shown in (1) [17]–[19]. 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 Fig. 4, the PDF of V_{TH} in SRAM devices is modeled as a normal distribution. Since caches contain many bit cells, 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 small 8-kB SRAM, the bit cell failure rate must be 1.53×10^{-7}. To calculate this bit cell yield using Monte Carlo, at least 10 million simulations must be performed, making this procedure computationally intensive. For larger caches, the required bit cell failure rate is even lower and complete Monte Carlo analysis is almost infeasible

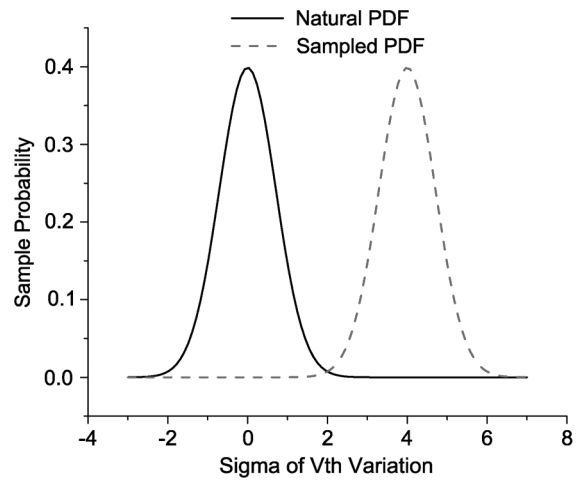

Fig. 4. For importance sampling, device V_{TH}s are selected from the SPDF. The SPDF is created by shifting the natural V_{TH} PDFs into the failure region by introducing mismatch between devices. By selecting V_{TH} values from this region of interest, the SRAM failure modes can be analyzed more quickly.

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

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

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

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

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

For our study, we choose importance sampling as an efficient and accurate way of calculating the SRAM robustness. As shown in Fig. 4 and (2)--(5), the importance sampling technique chooses a new sampling PDF (SPDF) for each transistor so that more failures are simulated. The V_{TH} of each transistor is shifted by the value sampled from the PDF plus 4σ, to be justified later, to introduce enough mismatch into the bit cell to increase the probability of failure. Since the natural occurrence of these highly skewed devices is rare, the importance samples are then weighted by the ratio of the probability of the large V_{TH} shift in each transistor to the probability that these V_{TH} shifts were sampled. These weighted values are then used to calculate the bit cell yield. This method allows us to accurately measure the region of interest where SRAM can fail with greatly reduced computational complexity.

Since the sampling PDF and number of importance samples have a large impact on experimental results, they were carefully chosen to maintain accuracy in the simulation while still reducing simulation run-time. A small V_{TH} shift in the SPDF

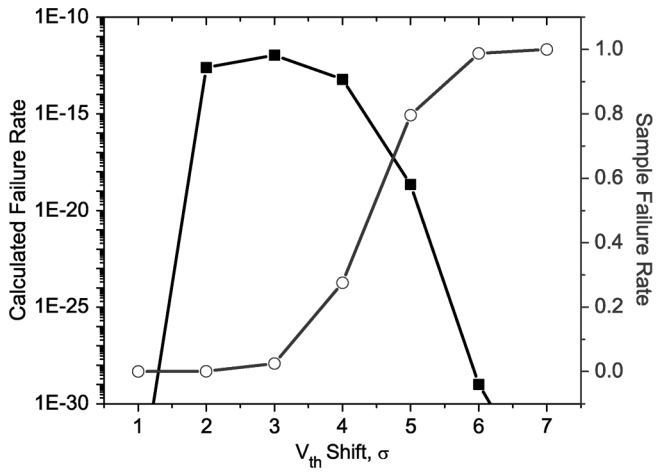

Fig. 5. Natural PDF is shifted into the failure region to create the SPDF. If the SPDF is too similar to the PDF, then simulation run-time to calculate yield is long because few failures are seen. If the SPDF greatly varies from the PDF, then more samples are necessary for accurate yield calculations.

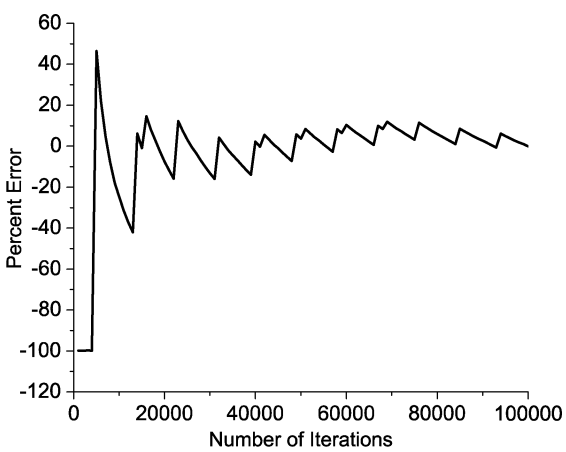

Fig. 6. After a sufficient number of importance samples have been simulated, the calculated yield converges to the correct value.

would not introduce a large number of failures, thus negating the variance reduction effect of importance sampling. Conversely, an excessively large V_{TH} shift introduces failures, but causes the sample weighting to be small and reduces the accuracy of the simulation. A differential 6T bit cell is studied to find the optimal sampling PDF. As shown in Fig. 5, with less than a 4σ V_{TH} shift, the sample failure rate is very low. Above a 4σ V_{TH} shift, the calculated failure rate drops and is inaccurate. Therefore, a 4σ shift is chosen for our study. After a sufficient number of importance samples have been performed, the calculated failure rate converges to its final value. Fig. 6 shows that the calculated failure rate converges and more samples are taken. We determine that 20 000 samples are sufficient for accurate results. To measure the failure rates in our study with Monte Carlo, at least 10^{12} samples are needed, making importance sampling 50 million times faster.

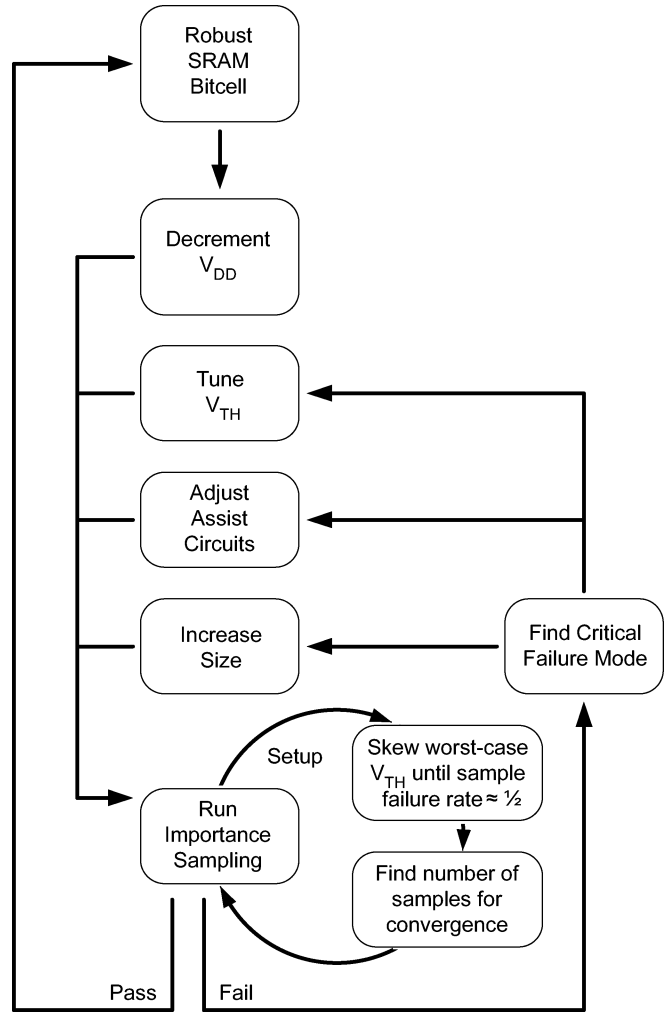

Fig. 7. Design methodology for yield-driven near-threshold SRAM.

IV. Experimental Results

A. Bit Cell Sizing and Doping in Near-Threshold SRAM

We examine the area and energy of 6T and 8T bit cells when V_{DD} is scaled down, robustness is maintained through sizing, and delay is constrained to scale with logic. The analysis is performed with and without the ability to individually adjust the pass gate, pull down, and pull up device V_{TH}s. The isorobustness bit cell sizings are plotted in Fig. 8. V_{TH} tuning dramatically reduces the required bit cell area at low voltage. V_{TH} is often set higher for pass gate devices than for pull down devices to prevent read upset failures at superthreshold V_{DD}. However, when V_{DD} is scaled to near-threshold voltages, this V_{TH} selection makes the pass gates too weak for write robustness. This effect is especially strong when the pass gates enter subthreshold operation, but other devices are still in the near-threshold regime. If V_{TH} is a fixed parameter set as a value optimized for superthreshold operation, then SRAM bit cells must be sized by 400% at 500 mV, and voltage scaling to the subthreshold regions is not practical for isorobustness operation. Tuning V_{TH} enables balancing of the SRAM failure modes. In

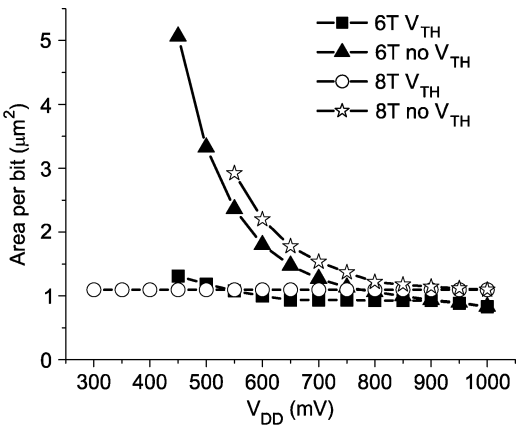

Fig. 8. Bit cells can be sized up to maintain robustness when V_{DD} is scaled. The density of isorobustness subthreshold SRAM is improved when V_{TH} tuning is used.

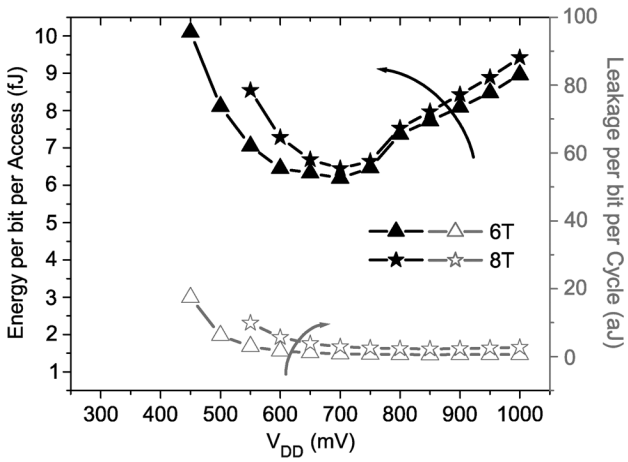

Fig. 9. When bit cells are sized for robustness without V_{TH} tuning, the energy benefits from isorobustness scaling of 65-nm SRAM to the near-threshold region are limited because WL and BL capacitances from upsized devices are prohibitive.

our study, the highest density, robust SRAM is achieved by increasing pass gate V_{TH} to prevent read upset for near-threshold SRAM. As V_{DD} is further scaled, the optimal pass gate V_{TH} is lower because write failures become critical. When V_{TH} is tuned, the robustness can be maintained in the 6T SRAM at 500 mV with a 40% area penalty. Across all voltages studied, V_{TH} tuning with minimal sizing is sufficient to maintain robustness in the 8T SRAM, enabling high-density low-voltage memory.

V_{DD} scaling from 1 to 500 mV reduces dynamic energy by more than 50% for all bit cells studied, with a 61% reduction for 8T SRAM with V_{TH} tuning, as shown in Figs. 9 and 10. In our study, we consider energy from the WL drivers, BL drivers, and bit cells only. If other memory peripheries, such as the decoder and sense amplifiers, are voltage scaled with the SRAM bit cells, then energy gains greater than those reported are possible. Without V_{TH} tuning and at low voltages, bit cells must be aggressively sized to control relative device strengths under RDF variation to maintain robustness. Device sizing substantially increases WL and BL capacitances, thus reducing the energy benefit of voltage scaling. For subthreshold robustness without V_{TH}

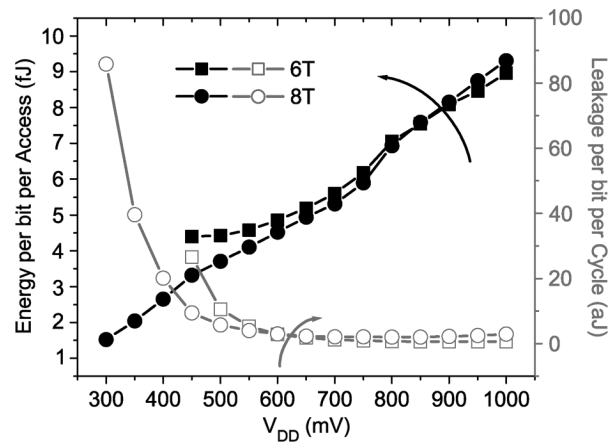

Fig. 10. When bit cells are sized and V_{TH} is tuned for robustness, dynamic energy can be reduced by as much as 83% in isorobustness SRAM using voltage scaling.

tuning, devices must be sized up to a level where the energy benefit is eliminated. When V_{TH} tuning is used, less dramatic sizing is needed, keeping capacitance and energy lower. Using V_{TH} tuning for near-threshold 550-mV SRAM reduces dynamic energy by 44% and 56% for 6T and 8T SRAMs, respectively, over the fixed V_{TH} case. For 8T SRAM, isorobustness operation at 300 mV is obtained with little device sizing, and an 83% energy reduction is achieved.

Above 500 mV, leakage energy per cycle is relatively constant, whereas below 500 mV, leakage increases dramatically. Although leakage power scales down linearly with V_{DD}, leakage energy per cycle is also proportional to delay, which increases exponentially in the subthreshold region. Since dynamic energy decreases and leakage increases when V_{DD} is scaled down, a minimum energy point (E_{MIN}) is achieved at some intermediate voltage (V_{MIN}) [1]. As seen in Figs. 11 and 12, V_{MIN} and E_{MIN} are heavily dependent on the activity factor, which we define as the average fraction of bit cells accessed per cycle. For L1 caches, which are generally small with high activity, the total energy is almost entirely dynamic, making voltage scaling a desirable method for energy reduction. Based on typical memory access patterns, an 8T eight-way 1 kB L1 cache could have an activity factor of 10^{-2} and a V_{MIN} of 450 mV. In L2 caches, which are larger with lower activity, the benefits of voltage scaling are reduced and V_{MIN} rises. A large L2 cache could easily have an activity factor lower than 10^{-6}, making voltage scaling below 850 mV detrimental.

B. Assist Circuits for Near-Threshold SRAM

Assist circuits can be used to increase SRAM robustness and enable near-threshold operation. The dual-V_{DD} WL, negative BL, and supply drooping assist circuits considered in this study increase the SRAM stability by modifying control or supply voltages during accesses. The voltage levels necessary for isorobustness operation are shown in Fig. 13. Of the three write assist circuits studied, only overdriven WL enables subthreshold SRAM. A functional minimum-sized SRAM cell with no V_{TH} tuning and an SRAM array V_{DD} of 300 mV requires a write

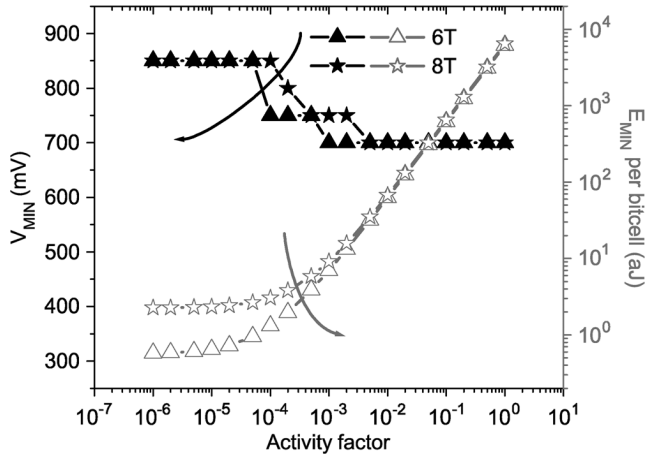

Fig. 11. When bit cells are sized for robustness without V_{TH} tuning, the supply voltage for minimum energy computing is above 700 mV, because the large device sizes needed to maintain robustness in near-threshold SRAM result in large capacitances and switching energy.

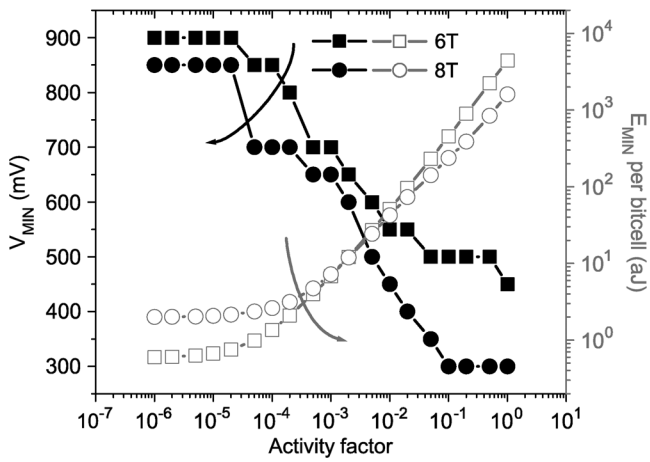

Fig. 12. When bit cells are sized and V_{TH} is tuned for robustness, 8T L1 caches with high activity factor can benefit from voltage scaling to 300 mV.

WL voltage of 650 mV. This near-threshold WL voltage requires additional access energy and precludes unaccessed bit cells on the WL. Negative BL and supply drooping can keep the robustness high when V_{DD} is scaled to 650 and 600 mV, respectively. Below these voltages, the aggressive assist circuits needed to maintain robustness disturb unaccessed bit cells. For the negative BL scheme, when the BL is driven below GND for a write, pass gates of unaccessed bitcells are turned partially on and can cause erroneous writes. For unaccessed bit cells with supply drooping, process variation and supply transients cause the loss of state.

An underdriven WL helps to prevent read upset failures but also reduces performance as shown in Fig. 14. At 500 mV, the WL voltage must be reduced to 250 mV to have the same robustness as the unassisted bit cell with a 1 V supply, resulting in a 200× increase in bit cell delay. This excessively large delay also manifests itself as intolerable leakage energy as shown in Fig. 15. As a result, when assisted SRAM circuits are scaled to the near-threshold region, leakage energy dominates and the

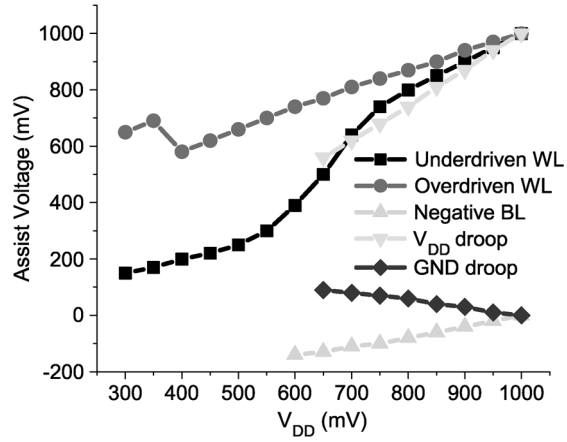

Fig. 13. Assist circuits modulate SRAM voltages and can maintain robustness as V_{DD} is scaled down. Underdriving the WL during read prevents read upset. Overdriving the WL during write can enable isorobustness subthreshold SRAM. Negative BL and supply drooping disturb unaccessed bit cells below 600 mV.

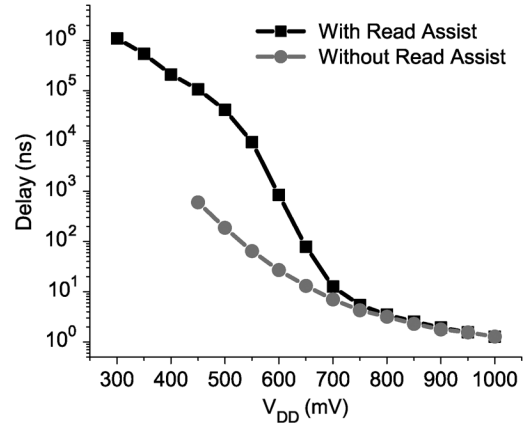

Fig. 14. Delay of isorobustness 6T bit cells is significantly greater with read assist than with sizing and doping.

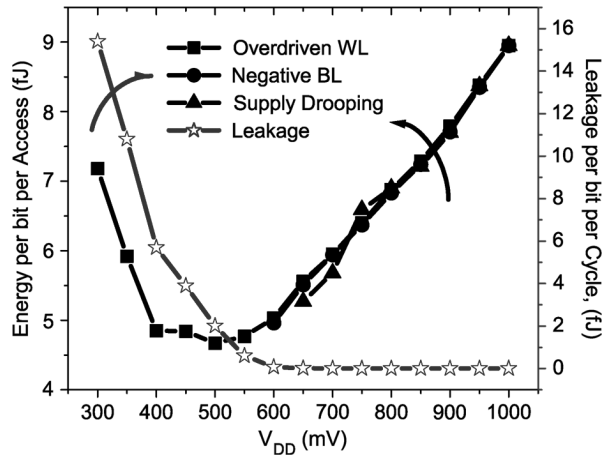

Fig. 15. When assist circuits are used to maintain SRAM robustness, dynamic energy can be reduced by 50% by halving V_{DD}.

V_{MIN} never falls below 600 mV, regardless of the activity factor. The active energy for the three write assist circuits is almost the same. The V_{MIN} for all assist circuits is shown in Fig. 16.

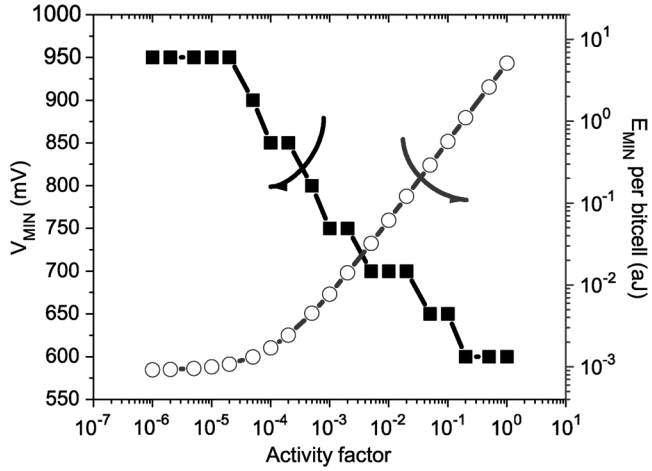

Fig. 16. When assist circuits are used to maintain SRAM robustness, V_{MIN} for caches with high activity factor can be as low as 600 mV.

V. Conclusion

We compared 6T and 8T bit cells in various voltage domains with an isorobustness condition. Our study is enabled by using importance sampling to accurately calculate SRAM yield 50 million times faster than with Monte Carlo sampling. We find that energy gains of 50% can be achieved for small caches by halving V_{DD} to 500 mV with no decrease in robustness and a small area overhead. At 300 mV, 8T SRAM with low V_{TH} devices can deliver an 83% energy reduction over the nominal case. For L1 caches, the supply voltage for minimum energy isorobustness operation can be as low as 300 mV, making voltage scaling a desirable technique for low-energy computing. Assist circuits can only enable isorobustness SRAM to scale to 600 mV before delay and leakage become prohibitive. The method shown in this paper assesses design tradeoffs in SRAM quickly and accurately, allowing a designer to select an appropriate SRAM architecture and sizing.

References

[1] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. 41st ACM/IEEE Des. Autom. Conf.* , 2004, pp. 868–873.

[2] A. Wang and A. Chandrakasan, "A 180 mV FFT processor using subthreshold circuit techniques," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2004, vol. 1, pp. 292–529.

[3] N. Lindert, T. Sugii, S. Tang, and C. Hu, "Dynamic threshold pass-transistor logic for improved delay at lower power supply voltages," *IEEE J. Solid-State Circuits*, vol. 34, no. 1, pp. 85–89, Jan. 1999.

[4] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, Oct. 1987.

[5] M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," in *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, Dec. 1998, pp. 915–918.

[6] L. Chang, D. M. Fried, J. Hergenrother, J. W. Sleight, R. H. Dennard, R. K. Montoye, L. Sekaric, S. J. McNab, A. W. Topol, C. D. Adams, K. W. Guarini, and W. Haensch, "Stable SRAM cell design for the 32 nm node and beyond," in *Symp. VLSI Technol., Dig. Tech. Papers*, Jun. 2005, pp. 128–129.

[7] S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, T. Yoshihara, M. Igarashi, M. Takeuchi, H. Kawashima, Y. Yamaguchi, K. Tsukamoto, M. Inuishi, H. Makino, K. Ishibashi, and H. Shinohara, "A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits," *IEEE J. Solid-State Circuits*, vol. 42, no. 4, pp. 820–829, Apr. 2007.

[8] O. Hirabayashi, A. Kawasumi, A. Suzuki, Y. Takeyama, K. Kushida, T. Sasaki, A. Katayama, G. Fukano, Y. Fujimura, T. Nakazato, Y. Shizuki, N. Kushiyama, and T. Yabe, "A process-variation-tolerant dual-power-supply SRAM with 0.179 μm^2 cell in 40 nm CMOS using level-programmable wordline driver," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2009, pp. 458–459.

[9] D. P. Wang, H. J. Liao, H. Yamauchi, Y. H. Chen, Y. L. Lin, S. H. Lin, D. C. Liu, H. C. Chang, and W. Hwang, "A 45 nm dual-port SRAM with write and read capability enhancement at low voltage," in *Proc. IEEE Int. SOC Conf.*, Sep. 2007, pp. 211–214.

[10] H. Pilo, C. Barwin, G. Braceras, C. Browning, S. Lamphier, and F. Towler, "An SRAM design in 65-nm technology node featuring read and write-assist circuits to expand operating voltage," *IEEE J. Solid-State Circuits*, vol. 42, no. 4, pp. 813–819, Apr. 2007.

[11] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200 mV 6T SRAM in 0.13 μm CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 332–606.

[12] B. H. Calhoun and A. Chandrakasan, "A 256 kb sub-threshold SRAM in 65 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2006, pp. 628–629.

[13] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, "A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 330–606.

[14] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mV robust Schmitt Trigger based subthreshold SRAM," *IEEE J. Solid-State Circuits*, vol. 42, no. 10, pp. 2303–2313, Oct. 2007.

[15] C. H. Kim and K. Roy, "Dynamic Vt SRAM: A leakage tolerant cache memory for low voltage microprocessors," in *Proc. Int. Symp. Low Power Electron. Des.*, 2002, pp. 251–254.

[16] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 12, pp. 1859–1880, Dec. 2005.

[17] P. Shahabuddin, "Importance sampling or the simulation of highly reliable Markovian systems," *Manage. Sci.*, vol. 40.3, no. 3, pp. 333–352, 1994.

[18] 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 *Proc. 43rd ACM/IEEE Des. Autom. Conf.*, Jul. 2006, pp. 69–72.

[19] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," *ACM Trans. Model. Comput. Simul.*, vol. 5, no. 1, pp. 43–85, 1995.

Gregory Kengho Chen (S'07) received the B.S. and M.S. degrees in electrical engineering from the University of Michigan, Ann Arbor, in 2006 and 2009, where he is currently pursuing the Ph.D. degree.

His research interests include SRAM, power management, and energy harvesting.

Dennis Sylvester (S'95–M'00–SM'04) received the Ph.D. degree in electrical engineering from the University of California, Berkeley, where his dissertation research was recognized by the David J. Sakrison Memorial Prize as the most outstanding research in the UC-Berkeley EECS Department.

He is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. He previously held research staff positions in the Advanced Technology Group of Synopsys, Mountain View, CA, Hewlett-Packard Laboratories, Palo Alto, CA, and a visiting professorship in Electrical and Computer Engineering at the National University of Singapore. He has published over 200 articles along with one book and several book chapters in his field of research, which includes low-power circuit design and design automation techniques, design-for-manufacturability, and interconnect modeling. He also serves as a consultant and technical advisory board member for electronic design automation and semiconductor firms in these areas.

Dr. Sylvester was a recipient of an NSF CAREER Award, the Beatrice Winner Award at ISSCC, an IBM Faculty Award, an SRC Inventor Recognition Award, and numerous best paper awards and nominations. He was also the recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He has served on the technical program committee of major design automation and circuit design conferences, the executive committee of the ACM/IEEE Design Automation Conference, and the steering committee of the ACM/IEEE International Symposium on Physical Design. He is currently an Associate Editor for IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and previously served as Associate Editor for IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He is a member of ACM and Eta Kappa Nu.

David Blaauw (M'94) received the B.S. degree in physics and computer science from Duke University, Durham, NC, and the Ph.D. degree in computer science from the University of Illinois, Urbana, in 1991.

Until August 2001, he worked for Motorola, Inc., Austin, TX, where he was the manager of the High Performance Design Technology Group. Since August 2001, he has been on the faculty at the University of Michigan where he is a Professor. He has published over 300 papers and holds 29 patents. His research has a three fold focus. He has investigated adaptive computing to reduce margins and improve energy efficiency using an approach called Razor for which he received the Richard Newton GSRC Industrial Impact Award. He has active research in resilient circuit design for wearout and error prone silicon. His latest work is focused on ultra-low power computing using near-threshold and subthreshold computing for millimeter sensor systems. This work recently lead to a processor design with record low power consumption which was selected as one of the year's most significant innovations in Technology Review.

Dr. Blaauw was a recipient of an extensive number of best paper awards and nominations, the Motorola High Impact Technology Award in 1996, and the Motorola Innovation Award in 1997. In recent years, he received the Richard Newton GSRC Industrial Impact Award in 2008, the Analysts Choice Award from Microprocessor Review for Innovation in 2007, the University of Michigan Henry Russel Award for teaching and research in 2004, and the IBM Faculty Award in 2003. His research has been featured in the MIT Technology review in 2008 and in IEEE Spectrum in 2009. He was general chair of the IEEE International Symposium on Low Power, technical program chair for the ACM/IEEE Design Automation Conference, and a member of Program Committee of the IEEE International Solid-State Circuits Conference.

Trevor Mudge (S'74–M'77–SM'84–F'95) received the Ph.D. degree in computer science from the University of Illinois, Urbana, in 1977.

Since, he has been at the University of Michigan, Ann Arbor. He became the Bredt Family Professor of Electrical Engineering and Computer Science after ten years as the Director of the Advanced Computer Architecture Laboratory—a group of 10 faculty and 80 graduate students. He has coauthored numerous papers on computer architecture, programming languages, VLSI design. He has also chaired 40 theses.

Dr. Mudge is a member of the ACM, the IET, and the British Computer Society.