SRAM Assist Techniques for Operation in a Wide Voltage Range in 28-nm CMOS

Brian Zimmer, *Student Member, IEEE*, Seng Oon Toh, *Member, IEEE*, Huy Vo, Yunsup Lee, *Student Member, IEEE*, Olivier Thomas, Krste Asanović, *Senior Member, IEEE*, and Borivoje Nikolić, *Senior Member, IEEE*

***Abstract*—Reducing static random-access memory (SRAM) operational voltage (Vmin) can greatly improve energy efficiency, yet SRAM Vmin does not scale with technology due to increased process variability. Assist techniques have been shown to improve the operation of SRAM, but previous investigations of assist techniques at design time have either relied on static metrics that do not account for important transient effects or make specific assumptions about failure distributions. This paper uses importance sampling of dynamic failure metrics to quantify and analyze the effect of different assist techniques, array organization, and timing on Vmin at design time. This approach demonstrates that the most effective technique for reducing SRAM Vmin is the negative bitline write assist, resulting in a Vmin of 600 mV for a 28-nm LP process in the typical corner.**

***Index Terms*—Assist techniques, importance sampling, low-voltage static random-access memory (SRAM), SRAM.**

I. Introduction

CIRCUIT OPERATION over a wide range of supply voltages enables maximum energy efficiency for a given performance requirement, yet increased variability in deeply scaled technologies prevents static random-access memory (SRAM) from achieving as low an operational voltage (Vmin) as logic. Assist techniques that dynamically change the operating characteristics of bitcells, such as boosting the wordline voltage above the cell voltage, can lower Vmin for SRAM.

A comprehensive analysis of the effectiveness of assist techniques was performed in [1] using static metrics, which have been shown to be a poor match to silicon failures [2]. Transient simulations are required to quantify the effect of implementation details that have a large effect on failure rates, such as the shape of the boosted wordline waveform. Other works have resolved this problem by investigating the impact of assist techniques on dynamic writeability by assuming that failure metrics can be fit to a particular distribution [3]. Accelerated Monte Carlo techniques such as importance sampling (IS) do not make any assumptions about failure distributions and can track SRAM tail cells [4]. IS has been used to analyze how cell sizing, doping, and assists can minimize SRAM energy, but limited focus was given to assist techniques [5]. An alternative to IS based on statistical classifiers was used to analyze Vmin, but focused on dynamic writeability only and provided little investigation of assists [6]. This work provides an in-depth analysis of assist techniques' potential to reduce Vmin using IS for dynamic metrics.

Manuscript received July 17, 2012; revised October 2, 2012; accepted October 27, 2012. Date of current version February 1, 2013. This work was supported in part by DoE Award DE-SC0003624, AMD and Intel ARO. This brief was recommended by Associate Editor M. Alioto.

B. Zimmer, H. Vo, Y. Lee, K. Asanović, and B. Nikolić are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley CA 94720, USA (e-mail: bmzimmer@eecs.berkeley.edu).

S. O. Toh is with AMD Inc., Santa Clara, CA 95054, USA.

O. Thomas is with CEA/LETI-MINATEC, 38054 Grenoble, France.

Color versions of one or more of the figures in this brief are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSII.2012.2231015

II. Failure Metrics

Several metrics based on transient simulations can be used to take dynamic effects into account by essentially imitating typical SRAM read and write operations. This paper defines three modes of failure: readability, writeability, and read stability. Readability failures occur when a read bitline discharge in a specified time is less than the offset of the sense amplifier. This study assumes a differential read using a sense amplifier for six-transistor (6T) cells and a domino-style read with a skewed inverter for eight-transistor (8T) cells. Unlike I_{read}, this readability metric takes into account the reduced V_{DS} on the pass gate as the bitline discharges. Writeability failures occur when the internal node voltage does not reach the desired write value. Read stability failures occur when bitcell contents flip accidentally during a read condition. Half-select stability failures occur when bitcells on unaccessed columns of an interleaved array flip accidentally during a write operation. For readability and writeability, the result is checked at the end of the clock period to emulate typical SRAM operation where back-to-back access is supported. This pessimistic measure, which accounts for successive reads, results in an average Vmin increase of 30 mV. Each mode of failure is measured by transient simulations of netlists that accurately model read or write signal timing and capacitances.

For 90% yield on a 1-MB array, the probability of a single bit failure, or the bit error rate (BER), must be less than 10^{-9}. Vmin is defined as the minimum operating voltage for which the BER of all three types of failures is $< 10^{-9}$. This paper focuses on the 6T cell, then extends the results to the 8T cell. Fig. 1 shows both cells and some naming conventions: PU for pull-up, PD for pull-down, PG for pass-gate, RPD for read-pull-down, and RPG for read-pass-gate.

III. Analysis Methodology

A. Importance Sampling

Monte Carlo simulations can be used to measure the effect of variability, but an unfeasible number of simulations would be required to find a rare failure event. IS enables enormous speedup for Monte Carlo analysis of rare events [7]. Fig. 2(a) shows a graphical representation of the main idea behind IS for a simplified case of two devices, where the two axes represent

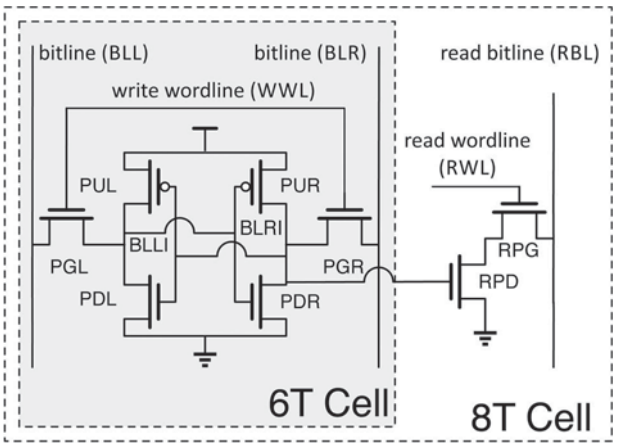

Fig. 1. SRAM bitcell schematics and naming conventions.

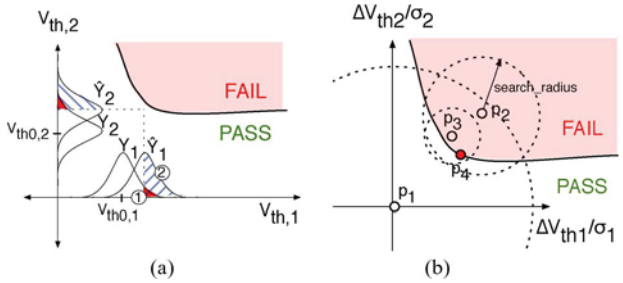

Fig. 2. Two-step algorithm to determine BER. (a) Importance sampling provides more information about the failure region by shifting the mean of each device's threshold voltage. (b) Graphical example of the variable-radius algorithm for hypothetical two-device circuit.

device threshold voltages. Under the assumption that threshold voltage deviations due to random dopant fluctuation can be modeled with a Gaussian distribution, Monte Carlo will simulate the two-device circuit with threshold probability density functions (PDF) given by Y_1 and Y_2, where the mean is the device's nominal value and the σ is given by $A_{Vt}/\sqrt{W * L}$. This enables investigation of a technology before full statistical models are available as long as an approximate $\sigma_{V_{th}}$ is known. In general, this method can be applied to any statistically modeled technology parameter. Ordinary Monte Carlo will only sample failure events with the very small probability shown by region ①. For IS, the mean of Y_1 and Y_2 is changed to create a new PDF, labeled $\hat{Y}_1$ and $\hat{Y}_2$, so Monte Carlo samples failure events with the probability given by region ②. To determine how often these failures would occur without the artificial shift, these samples are unbiased [7].

B. Variable-Radius Most Probable Failure Point Search

Finding the optimal sampling distribution is complex for a multi-dimensional design space. For quick convergence, the set of mean shifts must be the multi-dimensional most-probable failure point (MPFP), which is the point closest in distance to the origin. This paper uses a method that performs uniform sampling of a variable-radius n-dimensional sphere around points, a similar idea to [8].

Fig. 2(b) shows a simple case of a two-device circuit graphically. The space being searched is defined as shifts in thresholds for each device (normalized by their standard deviation), and the desired point is the closest point to the origin. Initially, a large search of 5σ in all directions is uniformly sampled from the origin (p1) and the closest failure point p2 is found. Larger initial searches are used if no failures are found after the first search. Sampling continues with a decreased search space until no closer points are found. Lastly, IS is run to determine the final BER.

TABLE I
MOST PROBABLE FAILURE POINT FOR WRITEABILITY AT 0.8 V

PUL ΔV_{th}	PDL ΔV_{th}	PGL ΔV_{th}	PUR ΔV_{th}	PDR ΔV_{th}	PGR ΔV_{th}
-.013σ	-1.39σ	1.62σ	-2.97σ	0.0292σ	5.64σ

Most importantly, this algorithm can be readily adapted to different circuits with different numbers of variables. All that is needed to run this algorithm and return a probability of failure is a netlist describing which thresholds can be shifted and failure metrics that return a pass/fail signal for a given threshold shift. This methodology finds the BER for any dynamic metric in approximately 5 min of simulation time on a modern computer, where most time is spent running small Monte Carlo transient simulations. Designer usability, rather than absolute performance operation, was the goal of this implementation.

This simulation methodology allows for rapid design-space exploration without sacrificing accuracy and provides intuition about the cause of failures.

C. Example: Analyzing Results

The MPFP approach will not only determine failure probability, it will also determine the relative strengths of particular devices that cause failure, and therefore explains why a cell fails. For example, consider the MPFP that this algorithm found for writeability failure determination at 0.8 V in Table I, with a predicted BER of $2 \cdot 10^{-11}$. The intuitive explanation of this MPFP is that if Monte Carlo was run for a very long time, most failures will have devices with thresholds shifted by around these amounts. Assuming device models are correct, real silicon cells would also fail for similar device characteristics. σ is around 25 mV, so the MPFP has a right PG with a threshold that is about 140 mV larger than normal.

In this example, the goal is to write a 0 to the right side [bitline-right-internal (BLRI)]. The cell is most likely to fail when pass-gate-right (PGR) is very weak and pull-up-right (PUR) is very strong, because the voltage divider between PGR and PUR will keep BLRI high instead of pulling it to 0 as desired. In addition, bitline-left-internal needs to be pulled high, but strong pull-down-left and weak pass-gate-left prevent this. These failure mechanisms change with different assists and at different voltages, making the intuition given by this approach invaluable to designing effective assists.

D. Verification

These results closely match Monte Carlo simulation, as shown in Fig. 3. Note that this IS implementation assumes that the only source of variation is the V_{th} variation, yet can be seen to track full MC well. BER smaller than 10^{-4} causes an excessive MC runtime. Other studies have shown that IS matches MC for longer simulations [7].

IV. SRAM FAILURE ANALYSIS

Using the methodology described, it is possible to analyze how different design decisions affect BER and Vmin and use

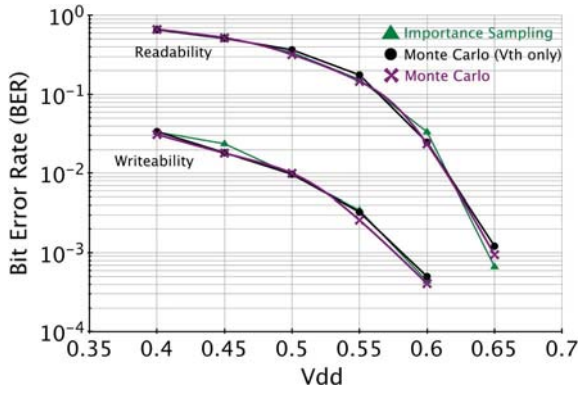

Fig. 3. Comparison Monte Carlo and importance sampling BER estimates for 90% accuracy and confidence.

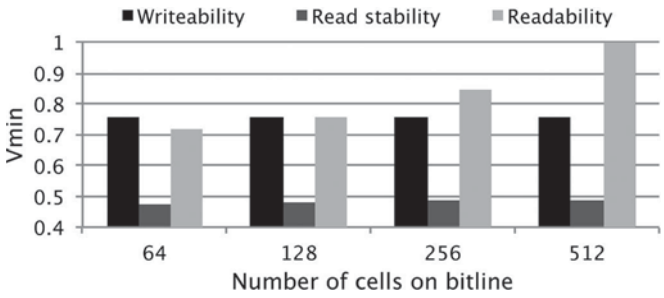

Fig. 4. Effect of bitline capacitance on BER.

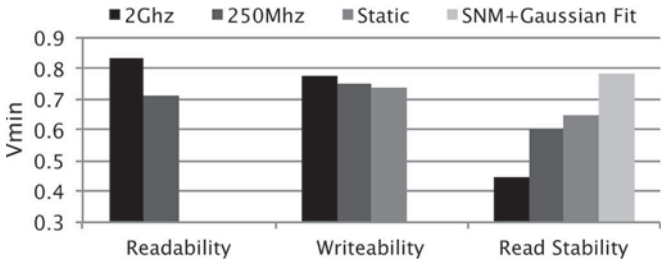

Fig. 5. Effect of the clock period on Vmin.

the MPFP to provide intuitive explanations for these quantitative findings.

A. Effect of Array Organization and Timing

Bitline capacitance, determined by the number of cells on a bitline, has a large effect on readability only, shown in Fig. 4. Writeability is unaffected as the bitlines are assumed to be held by a device, and stability shows negligible improvement for shorter bitlines because the low-node bitline can easy decrease voltage to match the internal value. Also, changing the frequency of SRAM operation will change the error rate for all three modes of failure, as shown in Fig. 5. Most importantly, this figure shows the large Vmin discrepancy between traditional static metrics and this methodology's read stability. At 2 Ghz, Vmin for read stability is about 450 mV, while a static test would predict a Vmin of 650 mV. Furthermore, fitting a Gaussian distribution to the static noise margin (SNM) instead of using IS would predict a Vmin of 780 mV. Designing using pessimistic metrics such as SNM and distribution approximations results in a drastic overdesign of the bitcell for stability.

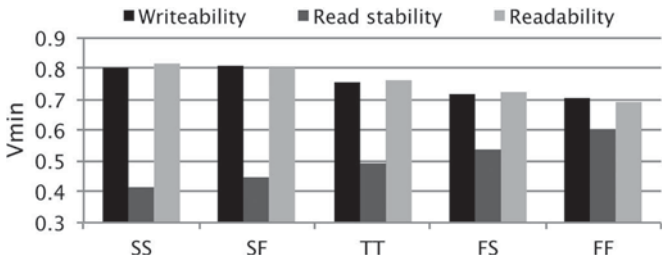

Fig. 6. Effect of process corners on Vmin.

B. Effect of Corners

A summary of the effects of corners is shown in Fig. 6. Readability and writeability are anti-correlated with stability, and process corners have the largest effect on stability. This behavior suggests that design-time optimization of a stability and writeability tradeoff would be dangerous.

C. Effect of Assist Techniques

Circuit-level assist techniques can be used on a cycle-by-cycle basis to improve Vmin. For the investigation of each assist technique, the following assumptions were made: corner: TT, design: HD 28 nm 6T 0.120 μm^2 cell [9], nominal voltage: 1 V, period: 50 FO4 ($\approx$1 ns at 1 V), wordline pulse width: 25 FO4 (1/2 of period), sense-amplifier offset: 0.1 V, bitline capacitance: 15 fF (128 cells). The effect of changing all of the above assumptions has been investigated in earlier sections. Note that the period will track the process FO4 as the supply is reduced to match the assumption that the SRAM operating frequency will be set by the critical path of a processor.

Different degrees of assist are defined as a percentage of the supply voltage (as opposed to the absolute quantity) because assist voltages are generally set by voltage dividers or charge redistribution and therefore are proportional to Vdd. Discussion of side effects assumes that Vdd runs vertically (parallel with bitlines) and GND runs horizontally (parallel with wordlines). Energy and area overhead are implementation dependent so are not quantified.

The voltage waveforms for a variety of assist techniques that target each mode of failure—readability, writeability, and read stability—are summarized in Fig. 7. Each technique is explained below.

The results of a thorough design-space exploration of effective assist techniques results are summarized in Fig. 8 for readability and writeability assists. Each Vmin measure also includes any stability consequences caused by the assist. Because all assists can be applied on a per-operation basis, the results for write Vmin and read Vmin are independent.

1) Effect of Negative Cell GND as a Readability Assist: Reducing the voltage of GND has been shown to improve readability [10]. Negative GND is the most effective of all readability assist techniques as it increases the Vgs on both the PD and PG by pulling the internal node holding 0 below ground. Unfortunately, this technique has a very high energy cost for 6T arrays, because each cell has two GND lines running horizontally, which have a large capacitance.

2) Effect of Wordline Boost as a Readability Assist: Using a wordline boost can improve both the readability and writeability of cells, but will drastically diminish the read stability of all half-selected cells in interleaved arrays, as shown in

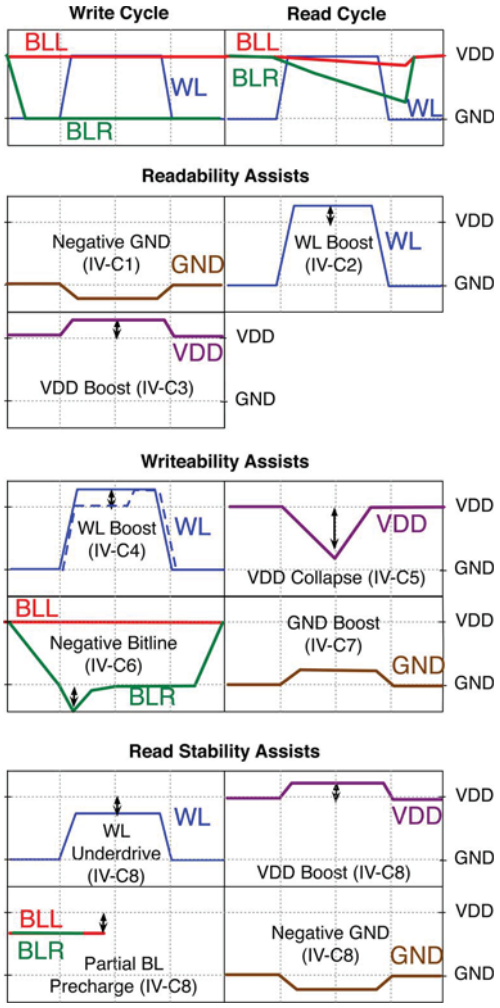

Fig. 7. Summary of assist techniques: negative GND, WL boost, Vdd boost, Vdd collapse, negative BL, GND boost, WL underdrive, partial BL precharge.

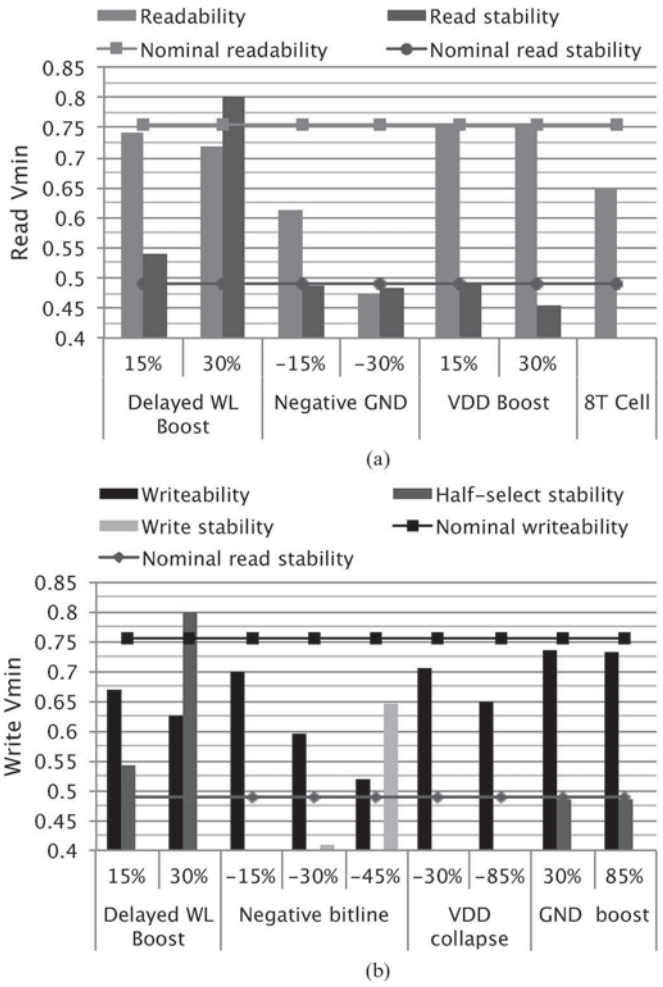

Fig. 8. Impact of assist techniques on Vmin. (a) Readability. (b) Writeability.

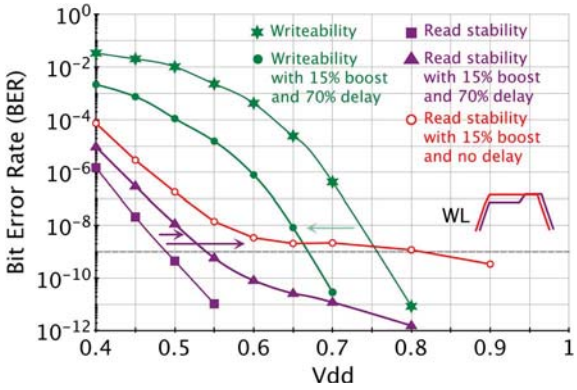

Fig. 9. Wordline boost improves writeability while reducing read stability.

Fig. 9. To avoid the half-select issue, Sinangil *et al.* [11] delays the boost until partway through the wordline pulse, so that half-selected cells have already started reading and the bitline voltage matches the internal voltage more closely. Fig. 9 shows that while delayed boost helps, the tradeoff between writeability and read stability remains very sensitive for wordline boosting. Adaptive circuitry must be used in order to tune this assist [12].

3) Effect of Cell Vdd Boost as a Readability Assist: For readability, both the PG and PD devices need to be strong. The Vdd boost improves the strength of the PD, but not the PG. Because the PD is sized larger than the PG for stability reasons, the probability that the PD limits read current more than the PG is much lower, suggesting that using wordline boost to improve the PG strength would be more effective.

4) Effect of Wordline Boost as a Writeability Assist: Wordline boost improves writeability by strengthening the PG, but hurts stability, as discussed above.

5) Effect of Cell Vdd Collapse as a Writeability Assist: Vdd collapse decreases write Vmin by decreasing the strength of the cross-coupled inverters [12]. However, for our cell, this assist is much less effective than negative bitline, because while Vdd collapse helps the PG pull the high-node low by weakening the high-node PU, a writeability failure can still occur if the PU on the low-node side is weak.

The dangerous consequence of Vdd collapse is potential violation of the data retention voltage of unaccessed SRAMs in the same column. However, IS tests of this condition show a BER $< 10^{-9}$ for all cases of interest.

6) Effect of Negative Bitline as a Writeability Assist: Negative bitline improves writeability by increasing the Vgs on the PG [10]. Simulation results show this as the most effective write assist, because it strengthens both the PG and helps pull the high-node low while also strengthening the low-node PU to complete the write operation. The IS simulation testbench uses a flying capacitor as opposed to a negative regulated voltage to accurately match typical implementations. However,

by decreasing one of the bitlines below GND, a non-zero Vgs will appear across the PG of unaccessed rows. If the internal node of an unaccessed bitcell on this side is high, then the value of the cell could flip, causing a write stability error. An IS test found that the BER for this case was $< 10^{-9}$ for boost amounts $\leq 30\%$.

7) Effect of Cell GND Boost as a Writeability Assist: GND boost weakens the cross-coupled inverters, improving writeability [13]. However, the effect of a large GND boost saturates after 30%, because the NMOS PG must pull the low internal node high for this assist to work, but can only pull up to around Vdd-V_{th}. This limitation does not exist for Vdd collapse, as NMOS can pass low voltages without limitation.

8) Effect of Stability Assists: Stability assists are not useful for the cell under investigation as readability and writeability Vmin dominate. However, for different cells and processes, stability can be a major concern.

Investigations at very low voltage show WL underdrive to be the most effective stability assist. Both Vdd boost and negative GND attempt to limit the voltage bump on the low side, but this effect is countered by shifting the switching threshold of the high-side inverter, cancelling most gains and making these techniques ineffective.

Pilo *et al.* [14] uses a regulator to precharge bitlines to around 70% of Vdd to improve yield from 5 to 5.7 sigma or, equivalently, from a BER of $2.9 \cdot 10^{-7}$ to $6 \cdot 10^{-9}$. IS analysis of read stability confirmed their results, with a BER improvement of around 1.5 orders of magnitude at 0.7 V. However, the assist becomes less helpful at low supplies and only achieves a Vmin reduction of 25 mV. Note that readability is slightly diminished due to the decreased Vds on the PG.

9) Other Techniques to Improve Readability: Readability has traditionally not been an SRAM design metric, as it depends on peripheral circuitry, while read stability and writeability do not. Variations cause a wide variability in read current. Shorter bitlines, as shown in Fig. 4, allow smaller read currents to provide the same required voltage difference but increases area overhead as sensing circuitry is amortized over fewer cells.

10) Leakage: The technology under investigation is a low-power process, so leakage was negligible even for a worst-case column of 512 cells. However, leakage can easily be taken into account by this methodology by using Monte Carlo to characterize the leakage current of N worst-case cells as a lognormal and including it into IS as an additional variable described by the fit distribution.

11) Assist Methods for 8T Cells: 8T bitcells have the same writeability assists as 6T bitcells. If 8T cells are interleaved, write operations will cause read stress on half-selected cells, producing the same read stability tradeoffs as the 6T cell. Readability assists are generally no longer needed, due to the much improved read path. Fig. 8(a) compares readability of the reference 6T array and bitcell to a domino-style read for an 8T cell with the same number of bitcells on a column.

V. Summary and Conclusion

The bitcell analyzed here is limited by both readability and writeability, but not stability. Using a combination of negative GND line to improve readability and a negative BL to improve writeability would lower Vmin by 175 mV. However, implementing negative GND involves the very difficult task of regulating a negative voltage that needs to sink every column's read current. Hence, to improve readability, modifications of the read path, such as using shorter bitlines, implementing hierarchical bitlines to minimize local bitline capacitance, using lower threshold devices, or lowering sense-amplifier offset, are needed to lower Vmin. Assists that improve stability have a very detrimental effect on readability and writeability, so cell sizing (sizing the PD stronger than the PG) remains the best option for maintaining stability.

Acknowledgment

The authors wish to acknowledge the contributions of the students, faculty and sponsors of the Berkeley Wireless Research Center.

References

[1] R. W. Mann, J. Wang, S. Nalam, S. Khanna, G. Braceras, H. Pilo, and B. H. Calhoun, "Impact of circuit assist methods on margin and performance in 6T SRAM," *Solid State Electron.*, vol. 54, no. 11, pp. 1398–1407, Nov. 2010.

[2] S. O. Toh, Z. Guo, T.-J. K. Liu, and B. Nikolić, "Characterization of dynamic SRAM stability in 45 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 46, no. 11, pp. 2702–2712, Nov. 2011.

[3] V. Chandra, C. Pietrzyk, and R. Aitken, "On the efficacy of write-assist techniques in low voltage nanoscale SRAMs," in *Proc. Des. Autom. Test Eur.*, 2010, pp. 345–350.

[4] 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. ACM/IEEE Des. Autom. Conf.*, 2006, pp. 69–72.

[5] G. Chen, D. Sylvester, D. Blaauw, and T. Mudge, "Yield-driven near-threshold SRAM design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 11, pp. 1590–1598, Nov. 2010.

[6] S. Nalam, V. Chandra, R. C. Aitken, and B. H. Calhoun, "Dynamic write limited minimum operating voltage for nanoscale SRAMs," in *Proc. Des. Autom. Test Eur.*, 2011, pp. 1–6.

[7] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2008, pp. 322–329.

[8] M. Qazi, M. Tikekar, L. Dolecek, D. Shah, and A. Chandrakasan, "Loop flattening & spherical sampling: Highly efficient model reduction techniques for SRAM yield analysis," in *Proc. Des., Autom. Test Eur.*, Mar. 2010, pp. 801–806.

[9] F. Arnaud, A. Thean, M. Eller, M. Lipinski, Y. W. Teh, M. Ostermayr, K. Kang, N. S. Kim, K. Ohuchi, J.-P. Han, D. R. Nair, J. Lian, S. Uchimura, S. Kohler, S. Miyaki, P. Ferreira, J.-H. Park, M. Hamaguchi, K. Miyashita, R. Augur, Q. Zhang, K. Strahrenberg, S. ElGhouli, J. Bonnouvrier, F. Matsuoka, R. Lindsay, J. Sudijono, F. S. Johnson, J. H. Ku, M. Sekine, A. Steegen, and R. Sampson, "Competitive and cost effective high-k based 28 nm CMOS technology for low power applications," in *Proc. IEEE IEDM*, Dec. 2009, pp. 1–4.

[10] M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Ohbayashi, Y. Nakase, and H. Shinohara, "A 45 nm 0.6 V cross-point 8T SRAM with negative biased read/write assist," in *Proc. IEEE Symp. VLSI Circuits*, 2009, pp. 158–159.

[11] M. Sinangil, H. Mair, and A. P. Chandrakasan, "A 28 nm high-density 6T SRAM with optimized peripheral-assist circuits for operation down to 0.6 V," in *Proc. Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2011, pp. 260–262.

[12] E. Karl, Y. Wang, Y.-G. Ng, Z. Guo, F. Hamzaoglu, U. Bhattacharya, K. Zhang, K. Mistry, and M. Bohr, "A 4.6 GHz 162 Mb SRAM design in 22 nm tri-gate CMOS technology with integrated active VMIN-enhancing assist circuitry," in *Proc. Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2012, pp. 230–232.

[13] A. Bhavnagarwala, S. Kosonocky, C. Radens, Y. Chan, K. Stawiasz, U. Srinivasan, S. P. Kowalczyk, and M. M. Ziegler, "A sub-600 mV, fluctuation tolerant 65 nm CMOS SRAM array with dynamic cell biasing," in *Proc. IEEE Symp. VLSI Circuits Dig.*, 2007, pp. 78–79.

[14] H. Pilo, I. Arsovski, K. Batson, G. Braceras, J. Gabric, R. Houle, S. Lamphier, C. Radens, and A. Seferagic, "A 64 Mb SRAM in 32 nm high-k metal-gate SOI technology with 0.7 V operation enabled by stability, write-ability and read-ability enhancements," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 97–106, Jan. 2012.