Variation-Aware Static and Dynamic Writability Analysis for Voltage-Scaled Bit-Interleaved 8-T SRAMs

Daeyeon Kim, Vikas Chandra*, Robert Aitken*, David Blaauw, Dennis Sylvester
Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI
*ARM R&D, San Jose, CA
daeyeonk@umich.edu, {vikas.chandra, rob.aitken}@arm.com, {blaauw, dmcs}@umich.edu

Abstract— **As process technology scales, SRAM robustness is compromised. In addition, lowering the supply voltage to reduce power consumption further reduces the read and write margins. To maintain robustness, a new bitcell topology, 8-T bitcell, has been proposed and read where write operation can be separately optimized. However, it can aggravate the half select disturb when write word-line boosting is applied or the bitcell sizing is done to enable robust writability. The half select disturb issue limits the use of a bit-interleaved array configuration required for immunity to soft errors. The opposing characteristic between write operation and half select disturb generates a new constraint which should be carefully considered for robust operation of voltage-scaled bit-interleaved 8-T SRAMs. In this paper, we propose bit-interleaved writability analysis that captures the double-sided constraints placed on the word-line pulse width and voltage level to ensure writability while avoiding half select disturb issue. Using the proposed analysis, we investigate the effectiveness of word-line boosting and device sizing optimization on improving bitcell robustness in low voltage region. With 57.7% of area overhead and 0.1V of word-line boosting, we can achieve 4.6σ of V_{TH} mismatch tolerance at 0.6V and it shows 41% of energy saving.**

Keywords-SRAM, Memory, Dynamic Write Margin, Half select Disturb

I. INTRODUCTION

Static random access memory (SRAM) is an indispensable part of most modern VLSI designs and dominates silicon area in many applications. In scaled technologies, maintaining high SRAM yield becomes more challenging since they are particularly vulnerable to process variations due to 1) minimum(close to minimum) sized devices used in SRAM bitcells and 2) the large array sizes (10s of MB). At the same time, low power design is a key focus throughout the semiconductor industry. Since low voltage operation is one of the most effective ways to reduce power consumption due to its quadratic relationship to energy savings, lowering the minimum operating voltage (V_{min}) of SRAM has gained significant interest.

To mitigate variability and reduce V_{min}, it is important to understand SRAM failure modes and quantify immunity to failures. Static noise margin [1] has been widely used as a metric to estimate the immunity to read/write/hold failures. However, it overestimates read failures and underestimates write failures since it assumes infinitely long word-line pulses. For more accurate analysis, dynamic writability has been introduced [2][3][4]. In addition, soft error susceptibility of SRAM to particle strikes is a key issue in modern SRAM design [5][6]. To fix soft errors, bit-interleaved arrays are commonly used; however this leads to the possibility of half select disturb, which degrades robustness. As the supply voltage is lowered, mitigating soft error becomes more important because soft error vulnerability is more critical at low voltage [6]. Hence bit-interleaving array must be adopted and half select disturb issue must be carefully analyzed.

To ensure robust operation at low voltage in nanoscale technologies, 8-T bitcell has been proposed [7][8][9] which can be separately optimized for read and write since bitcells are not interleaved. However, when 8-T bitcells are interleaved for immunity to soft errors, half select disturb issue [10] arises and it limits the freedom to maximize its writability. In terms of dynamic writability, the longer pulse width is favorable to write operation while the shorter pulse width is favorable to immunity to half select disturb. These double-sided constraints placed on word-line pulse width make it difficult to determine appropriate word-line pulse width to maximize SRAM robustness. Also, when write assist method is adopted, half-select disturb is more likely to happen and it will decrease overall yield.

To address these issues, this paper proposes bit-interleaved writability analysis (both static and dynamic) for voltage-scaled 8-T bitcell using SRAM worst-case corner simulations. It captures the double-sided constraints to ensure successful write operation and immunity to half select disturb. In addition, we can obtain appropriate word-line pulse width using bit-interleaved dynamic writability analysis.

Compared with the prior works commonly used for SRAM robustness analysis, this work highlights the double-sided constraints on 8-T bitcell write operation which can be mistakenly not considered on the assumption that 8-T bitcell has an unlimited freedom to maximize writability. The common method to avoid the double-sided constraints between read and write at 6-T bitcell is to apply different word-line pulse widths and voltage levels for each operations and this is feasible because read and write at a 6-T bitcell cannot be done simultaneously. However, the new constraints between write and half select disturb in a bit-interleaved 8-T SRAM cannot be solved using those methods because write targeted cell and half selected cell always experience the same word-line pulse width and voltage level. Therefore, special regard is paid to this fact in this paper. Also, this work uses the dynamic writability

The authors at University of Michigan acknowledge the support of NSF, GSRC, and Army Research Laboratory.

analysis and therefore it does not overestimate its failures by assuming an infinitely long word-line pulse.

With this analysis method, we evaluate the effectiveness of two techniques to lower V_{min}: word-line boosting and device size optimization. Poor writability and high soft error susceptibility limit low voltage operation. An SRAM cell in a commercial 45nm low-power CMOS can tolerate only up to 1.7σ at 0.6V in terms of worst case V_{TH} mismatch which is not acceptable for yield. To achieve iso-robustness(4.5σ) as 1.0V, device sizes need to increase and word-line boosting is needed. We can achieve 4.6σ tolerance at 0.6V with 57.7% area overhead and 0.1V of word-line boosting. Compared with normal write operation at 1.0V without any technique, it shows 41% of energy saving per operation.

The rest of the paper is organized as follows. Section II overviews SRAM failure modes and provides background. Section III describes bit-interleaved writability. Section IV explores the Vmin lowering techniques using bit-interleaved writability. Section V summaries the work and concludes.

II. Background and Related Work

Write failure and read disturb are two major SRAM failure modes. To quantify the probability of these failures, the static noise margin method [1] has been used for more than twenty years. In addition to write failure and read disturb, soft error [5][6] and half select disturb [10] have arisen as sources of SRAM failures. This section reviews these SRAM failure modes and related work.

A. Write Failure and Read Disturb

Figure 1(a) describes the write operation of a 6 transistor (6-T) bitcell. As access transistors (AXL and AXR) are turned on, values on bit-lines (WBL and WBLB) are driven to internal nodes of a bitcell which attempt to flip both nodes. A write failure occurs when the internal nodes do not flip due to the access transistors being too weak or the pull up PMOS transistors being too strong.

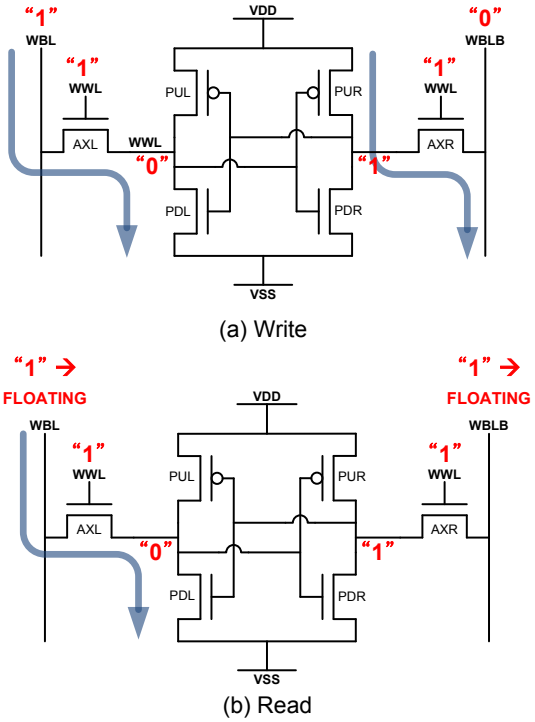


Figure 1. SRAM (a) write operation and (b) read operation

The read operation is depicted in Figure 1(b). Both bit-lines are pre-charged to logic "1" before reading. After the word-line pulse is asserted, one of the bit-line (WBL) falls rapidly due to active pull down while another bit-line (WBLB) falls very slowly due to the leakage of other bitcells connected to WBLB. The read operation is completed when a sense amplifier detects a sufficient voltage difference between the two bit-lines. A read disturb occurs when the internal nodes accidentally flip during read operation, caused by a voltage excursion from 0 due to an overly strong access transistor and weak pull-down transistor.

It is difficult to make both the read and write operations highly stable because strong access transistors are preferred for write operations while the opposite is true during read operations. To overcome this limitation, there has been many work using an 8-T bitcell [7][8][9] that decouples the read and write paths (Figure 2). The write operation is identical to the 6-T but the read operation is executed via a 2-T read path. With 8-T, devices on write and read paths can be optimized separately for each operation.

To estimate SRAM bitcell immunity to these failures, the static noise margin method [1] is typically used. This method assumes that two static noise voltage sources are inserted at the two internal nodes and then calculates the maximum noise voltage tolerance until an error occurs. The maximum noise voltage tolerance is called static noise margin (SNM) and the well-known butterfly curve is used to obtain SNM. While other work [11][12] have been recently proposed to supplement the SNM method, it remains the standard approach today.

B. Soft Error and Half Select Disturb

Soft errors are faults induced by a particle strike that upsets internal data states while the circuit itself is undamaged [5]. Even though it is unpredictable, soft error susceptibility is a critical reliability challenge for modern SRAM design [6]. Figure 3 shows three different scenarios when a soft error occurs. In Figure 3(a), a particle hit results in only one upset bit and it can be fixed with Hamming Single Error Correction/Double Error Detection (SECDED) codes [13]. Figure 3(b) and Figure 3(c) depict single event multi-bit upsets, which become more common in highly scaled technologies with smaller bitcells [14]. In Figure 3(b), all bits in a single

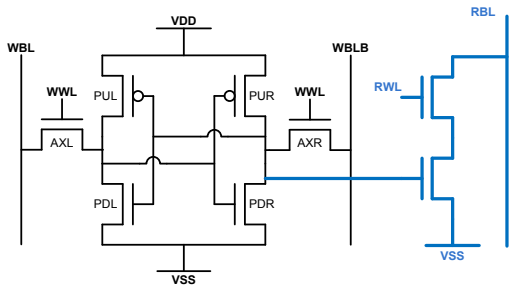


Figure 2. 8-T SRAM to decouple write and read

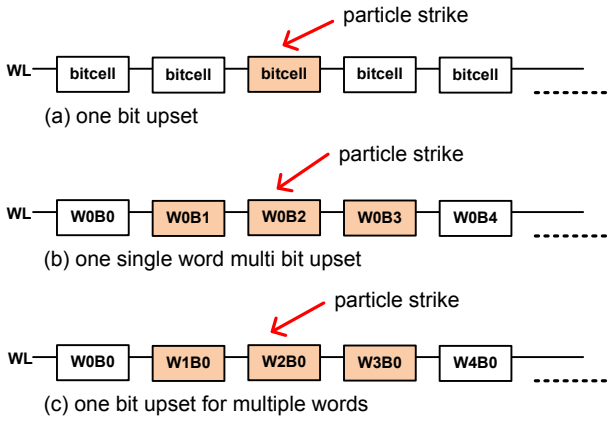


Figure 3. Three different scenarios of soft errors

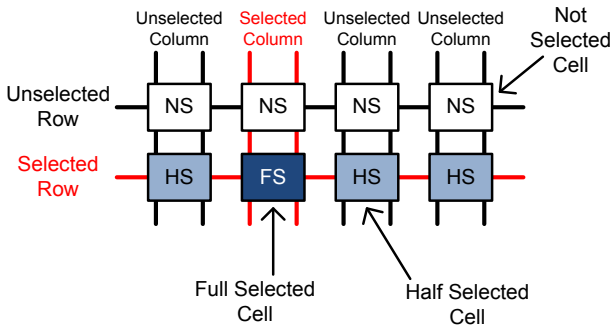


Figure 4. Half select disturb in bit-interleaved array

word are located next to each other and, therefore, a single word has multiple bits corrupted. This type of error cannot be fixed with SECDED codes and requires complicated approaches that incur large area penalties [14]. The easiest way to avoid single word multi-bit upsets is to interleave bits, such that logically adjacent bits are not physically adjacent. Figure 3(c) shows such a bit-interleaved array structure and in this case the multi-bit upset can be easily fixed with SECDED codes since each word contains only a single upset bit.

Although bit-interleaving is effective in avoiding single word multi-bit upset, it induces half select disturb problem. Figure 4 illustrates the half select disturb phenomenon in a bit-interleaved array. In a selected row containing 4 words, ¾ of columns are unselected, defined as "half selected", and as a result the access transistors in these cells are turned on and the internal data could flip(those half selected columns are in "6-T read-like" mode).

Several work [10][15] have focused on resolving this issue. Reference [10] proposed a local word-line scheme which does not allow bit-interleaving and reference [15] proposed the electron injection which requires process tweaking.

C. *Writability vs. Half Select Disturb in Bit-Interleved 8-T SRAMs*

In a 6-T SRAM, half select disturb is nearly identical to read disturb if bit-lines of unselected columns are floated. Therefore, half select disturb is unlikely to happen with appropriately sized 6-T SRAM if write assist method is not applied. However, if the 6-T portion of an 8-T SRAM is optimized for the write operation, half select disturb is more likely to happen and must be carefully considered when 8-T SRAM is designed. In addition, to regain robust writability at low voltage in nanoscale technologies, write assist methods such as word-line boosting are commonly used. Differently from read disturb in 6-T SRAM, half select disturb happens concurrently along with write operation so the write assist methods directly influence half select disturb and therefore the effect on yield must be analyzed before using these methods.

III. Writability Analysis For Bit-Interleaved SRAM

Static noise margin has long been used to estimate read-stability and writability. However, it assumes an infinitely long word-line pulse, making it optimistic for write and pessimistic for read operations compared with realistic SRAM operation. Recently, several papers [2][3][4] have considered dynamic writability to accurately assess SRAM writability. This section analyses SRAM bit-interleaved writability using worst case corner simulation. Because read operation is done through the 2-T read path of the 8-T bitcell, we consider the 6-T part of 8-T bitcell in the rest of the paper for write operation and the writability does not influence read operation.

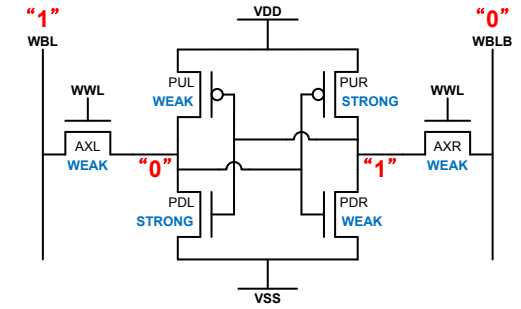
A. *SRAM Dynamic Writability Metric*

SRAM dynamic writability can be defined using the minimum duration write word-line pulse width required for a successful write operation, T_{CRIT} [2]. If T_{WL} is longer than T_{CRIT}, the write operation will be successful. However, a bitcell cannot be written for T_{WL} shorter than T_{CRIT} and this is referred to as dynamically limited write failure. If the bitcell cannot be written at all, even with an infinitely long word-line pulse, we refer to this case as statically limited write failure. T_{CRIT} is infinite in this case, allowing static write failure to be captured with the same metric.

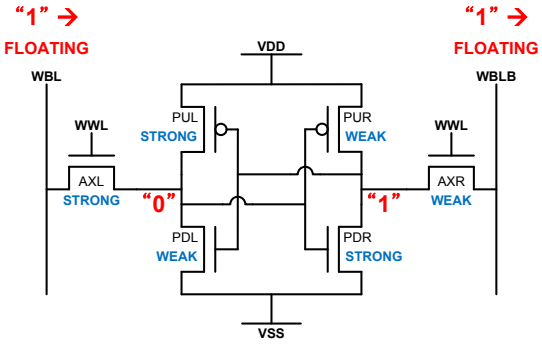
B. *SRAM Worst Case Corner Simulation*

SRAM worst case corner simulation is used to characterize the SRAM writability. The basic idea of this simulation is to find the maximum V_{TH} mismatch allowable before failure occurs, which is then used as the quantitative definition of writability [16].

A device becomes stronger or weaker when its V_{TH} decreases or increases, respectively. Initially, there is no V_{TH} skew for each device. To worsen the writability of s bitcell, V_{TH} of each device is skewed in appropriate directions. Figure 5(a) shows the worst case corner directions for each device in a write operation. Weak access transistors, AXL and AXR, worsen writability since it becomes difficult to drive the bit-line values onto the internal nodes through them. The strong left pull-down transistor (PDL) and the weak left pull-down transistor (PUL) tightly hold logic "0" and, therefore, writability is weakened. Similarly, the weak right pull-down transistor (PDR) and right pull-up transistor (PUR) worsen writability. The 6-T structure is symmetric and, therefore, we only consider a single state case (write "1" only) and it will reflect the other state too. Figure 5(b) shows the worst case corner directions for half select disturb. While an SRAM cell is designed not to have half select disturb in the absence of variations, transistor mismatch will incur such errors. The



(a) Worst Case Write

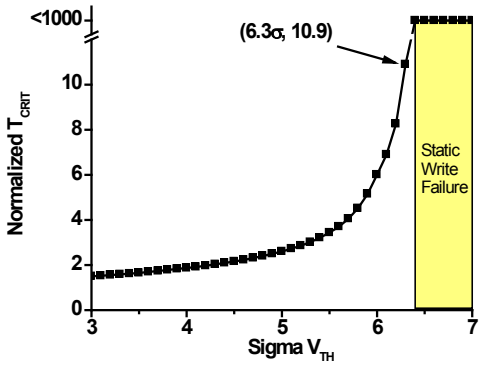


(b) Worst Case Half Select Disturb

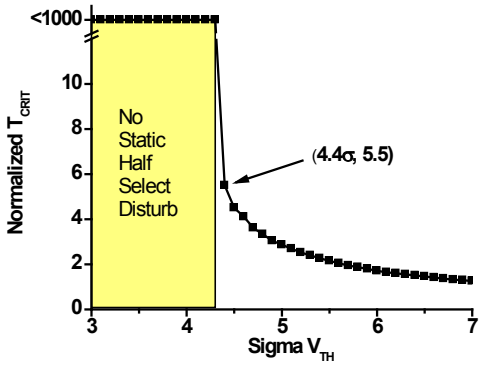
Figure 5. V_{TH} mismatch directions of each device (a) for write operation and (b) for half select disturb

worst case corner directions for the half select disturb differ from that of the write operation. To cause a disturbance, AXL should be strong enough to easily drive the logic "1" on write bit-line (WBL) to the internal node while AXR is weak such that the logic "1" on write bit-line bar (WBLB) does not help maintain the high state on the right internal node. The skewed directions of the internal four devices are set such that they do not strongly hold the internal nodes, allowing them to more easily flip. Initially, both bit-lines are pre-charged to logic "1". However, they float when write word-line (WWL) is asserted and the voltage levels of bit-lines are determined by how many bitcells are connected to each bit-line and the values stored in each bitcell. In our simulations, we assume 256 bitcells are connected to each bit-line with all other bitcells storing an opposite value to the target bitcell to ensure the worst case.

Figure 6 shows the worst case corner simulation result at 1.1V in a 45nm low-power CMOS process for (a) write operation and (b) half select disturb. All values are normalized to T_{CRIT} for the write operation at 0σ, i.e., no variation. The write operation (Figure 6(a)) can be successfully performed up to 6.3σ, however T_{CRIT} increases monotonically as V_{TH} mismatch increases indicating a steady degradation in write performance with variability. In the real simulation, because we cannot use the infinitely long word-line pulse, we assume that a pulse width of normalized 1000 as a practical limitation before static failure. Putting this in terms of static writability (SW, with infinitely long word-line pulse), SW at 1.1V is 6.3σ. On the other hand, the dynamic writability (DW) depends on T_{WL} (word-line pulse width). For example, if T_{WL} is allowed to be 3× the nominal value, DW at 1.1V is not 6.3σ but 5.2σ. While SW reveals a theoretical limitation, DW represents a more realistic view of actual writability.



(a) Write Corner



(b) Half Select Disturb Corner

Figure 6. Corner simulation results at 1.1V

Figure 6(b) depicts the half select corner result. Half select disturb does not occur up to 4.3σ even with an infinitely long word-line pulse hence 4.3σ is the static limitation of the half select disturb at 1.1V. As variation increases, the half select disturb likelihood increases, such that at 4.4σ, a 5.5X long word-line pulse is required to cause half select disturb. With more variation, the necessary word-line pulse width decreases, indicating that the cell becomes more vulnerable to half select.

C. Bit-Interleaved Writability Analysis

The previous subsection investigated static and dynamic writability and half select disturb. The simulation results show that a longer word-line pulse is simultaneously favorable for write and unfavorable for half select disturb. This is problematic when an SRAM array uses bit-interleaving for soft-error immunity. Without bit-interleaving, the dynamic writability can be improved at the expense of operation speed (e.g., by allowing for longer word-line pulses when variability is large). However, with bit-interleaving, such longer word-line pulses will generate half select disturbs, limiting overall array robustness. To analyze this tradeoff, the worst corner simulation results for the write operation and half select disturb are overlaid in Figure 7. With an infinitely long word-line pulse, the write operation tolerates up to 6.3σ variability while half select disturb starts to occur beyond 4.3σ. The bit-interleaved static writability (BSW) can be defined as the maximum V_{TH} mismatch until the write failure OR the half select disturb occurs, assuming an infinitely long word-line pulse. Therefore, BSW at 1.1V is 4.3σ. Up to 4.3σ, the write operation is successful if T_{WL} is larger than $T_{CRIT, WRITE}$ (T_{CRIT} of the write operation). However, BSW is pessimistic because the infinitely

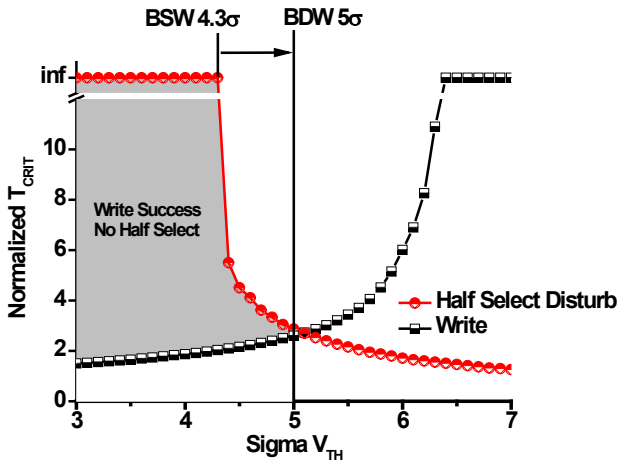


Figure 7. Bit-interleaved writability analysis at 1.1V

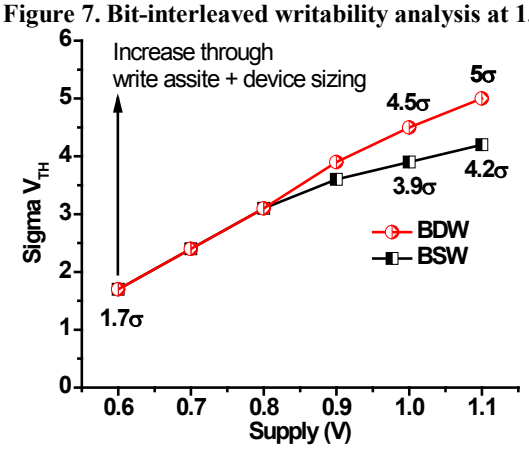


Figure 8. Bit-interleaved writability as voltage scales

long word-line pulse is unrealistic. To overcome this, the bit-interleaved dynamic writability (BDW) can be defined as the maximum V_{TH} mismatch until write failure and half select disturb occurs at a given word-line pulse width. At 4.4σ, if $T_{WL} > T_{CRIT, WRITE}$ yet smaller than $T_{CRIT, HALF}$ (T_{CRIT} of the half select disturb), the write operation can be successfully performed without incurring half select disturb. This leads to a BDW of 5σ at 1.1V. In this way, BDW best captures the tradeoff between writability and half select while capturing the negative correlation between these parameters.

IV. Writability Analysis at Near-Threshold

Figure 8 depicts the bit-interleaved writability with supply voltage scaling. Figure 8 clearly shows that the writability is very limited at near-threshold region. When BDW and BSW are overlapped, it has very poor writability and therefore the writability is statically limited before half-select happens. In this section, we investigate how to increase the writability using two writability enhancement techniques: word-line boosting and device sizing optimization.

A. Word-Line Boosting

The first approach to enhance the writability is word-line boosting. This is a commonly used technique for SRAM operation in low voltage regime [9][17]. In Figure 9, the two access transistors are over driven by the boosted word-line. By doing this, the current driving abilities of both transistors are enhanced, therefore, the SRAM cell becomes more favorable to the write operation. At the same time, the SRAM cell is more likely to experience read disturb and half select disturb with the word-line boosting.

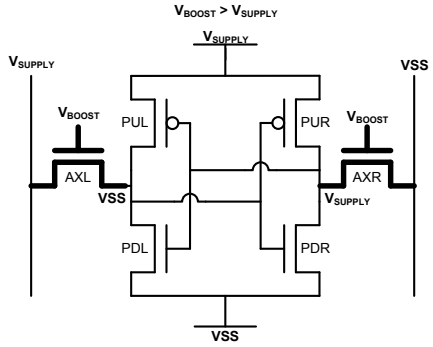


Figure 9. Word-line boosting

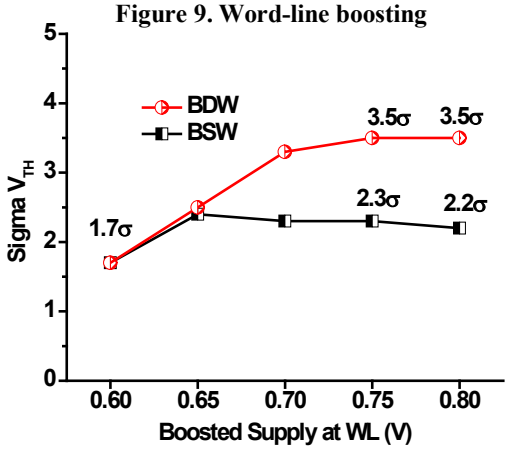


Figure 10. Bit-interleaved writability with word-line boosting at 0.6V

Figure 10 depicts the writability as the word-line boosting voltage increases when the supply voltage is 0.6V. Without boosting, both BDW and BSW are 1.7σ. After the boosted word-line is used, the writability is enhanced. However, BSW gets worse beyond 0.7V of the boosted supply because it make the bitcells more prone to half select disturb. On the other hand, BDW monotonically increases as the boosted supply increases but saturates sooner. In conclusion, the word-line boosting is effective up to 0.75V and 3.5σ of BDW is achieved.

B. Device Sizing Optimization

The second approach is device sizing optimization. Reference [18] shows that sizing optimization can achieve an iso-robustness condition while lowering the supply voltage, at the cost of density. Referring back to Figure 5(a), write operation is mainly driven by AXR and PUR since AXR can drive logic "0' into the internal node and PUR keeps logic "1" in the internal node. The writability can be enhanced by increasing the width of access transistors or increasing the length of pull-up devices. On the other hand, strong pull-down devices are favorable to avoid half select disturb. Here we increase the width of access transistor and pull-down transistor simultaneously. Increasing the length of pull-down devices is not used because it makes a notch in poly which is not

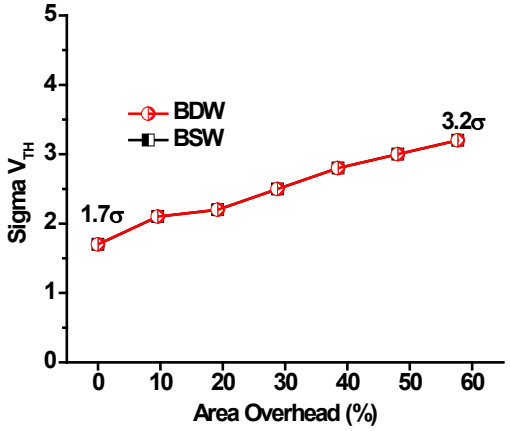


Figure 11. Bit-interleaved writability with device sizing at 0.6V

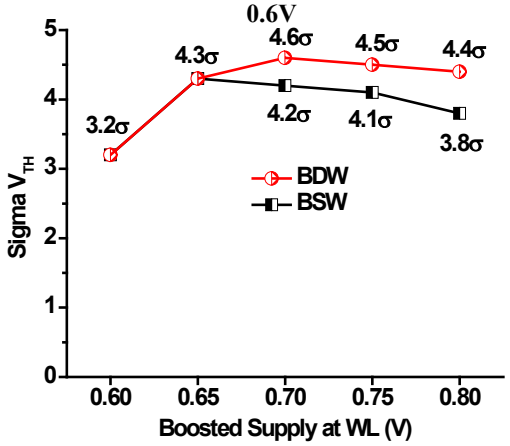


Figure 12. Bit-interleaved writability with word-line boosting and 57.7% area overhead

favorable for design for manufacturability (DFM). Area overhead is calculated based on layout.

With device sizing, BDW and BSW monotonically increase at the same time (Figure 11). Because the way of device sizing in this paper is favorable to both the writability and the half select disturb immunity, BSW also increases monotonically. With 57.7% of area overhead, BDW and BSW are extended to 3.2σ from 1.7σ.

C. *Dual Writability Enhancement*

In the previous subsections, two writability enhancement techniques are used. However, both techniques are practically limited to below 4σ. To achieve higher robustness, both techniques are applied simultaneously.

Figure 12 depicts how BSW and BDW change as the boosted word-line supply increases with a SRAM cell sized 57.7% larger than the nominal when the supply voltage is 0.6V. Since the two techniques are applied at the same time, BDW is extended to 4.6σ at 0.7V of boosted supply. Beyond 0.7V, the half select disturb overwhelms the writability so BDW decreases. This implies that higher word-line boosting does not guarantee better bit-interleaved writability. Also, we can clearly observe that BSW is too pessimistic and BDW reflects SRAM writability appropriately. In terms of energy consumption, it shows 41% saving over normal 1.0V operation.

V. CONCLUSION

In this paper, we discuss the writability and the half select disturb immunity of bit-interleaved 8-T SRAM arrays. The bit-interleaved static and dynamic writability analysis is proposed using worst case corner simulation to estimate the writability more precisely. At the end, two SRAM writability enhancement techniques are compared using the newly proposed analysis. The results show that device sizing and word-line boosting need to be used simultaneously to achieve higher robustness. To obtain the same robustness as 1.0V while lowering the supply voltage down to 0.6V, 0.1V of word-line boosting and 57.7% larger area are required. With these two techniques, we can successfully save the energy consumption per operation by 41%. In addition, the result confirms that higher word-line boosting does not guarantee better robustness because it lowers the half select immunity.

REFERENCES

[1] E. Seevinck, F.J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," IEEE Journal of Solid-State Circuits, Vol. 22, pp. 748-754, Oct. 1987

[2] J. Wang, S. Nalam, and B. H. Calhoun, "Analyzing static and dynamic write margin for nanometer SRAMs," International Symposium on Low Power Electronics and Design, pp. 129-134, Aug. 2008

[3] W. Dong, P. Li, and G.M. Huang, "SRAM Dynamic Stability: Theory, Variability and Analysis," IEEE/ACM International Conference on Computer-Aided Design, pp. 378-385, Nov. 2008

[4] S.O. Toh, Z. Guo, and B. Nikolic, "Dynamic SRAM stability characterization in 45nm CMOS," IEEE Symposium on VLSI Circuits, pp. 35-36, June 2010

[5] R.C. Baumann, "Soft errors in advanced semiconductor devices-part I: the three radiation sources," IEEE Transaction on Device and Materials Reliability, Vol. 1, pp.17-22, Mar. 2001

[6] V. Chandra, R. Aitken, "Impact of voltage scaling on nanoscale SRAM reliability," Design, Automation & Test in Europe, pp. 387-392, Apr, 2009

[7] L. Chang *et al.*, "Stable SRAM cell design for the 32nm down and beyond," IEEE Symposium on VLSI Circuits, pp. 128-129, June 2005

[8] L. Chang *et al.*, "A 5.3GHz 8T-SRAM with operation down to 0.41V in 65nm CMOS," IEEE Symposium on VLSI Circuits, pp. 252-253, June 2007

[9] A. Raychowdhury *et al.*, "PVT-and-aging adaptive wordline boosting for 8T SRAM power reduction," IEEE International Solid-State Circuits Conference, pp. 352-353, Feb. 2010

[10] R. Joshi *et al.*, "6.6+ GHz Low Vmin, read and half select disturb-free 1.2 Mb SRAM," IEEE Symposium on VLSI Circuits, pp. 250-251, June 2007

[11] E. Grossar *et al.*, "Read stability and writability analysis of SRAM cells for nanometer technologies," IEEE Journal of Solid-State Circuits, Vol. 41, pp. 2577-2588, Nov. 2006

[12] M. Wieckowski *et al.*, "A black box method for stability analysis of arbitrary SRAM cell structures," Design, Automation & Test in Europe, pp. 795-800, Mar. 2010

[13] R.W. Hamming, "Error detecting and error correcting codes," Bell System Technical Journal, Vol. 29, pp. 147-160, Apr. 1950

[14] R. Naseer, J. Draper, "Parallel double error correcting code design to mitigate multi-bit upsets in SRAMs," European Solid-States Circuits Conference, pp. 222-225, Sep. 2008

[15] K. Honda *et al.*, "Elimination of half select disturb in 8T-SRAM by local injected electron asymmetric pass gate transistor," IEEE Custom Integrated Circuits Conference, pp. 1-4, Sep. 2010

[16] G. Chen *et al.*, "Crosshairs SRAM – An adaptive memory for mitigating parametric failures," European Solid-States Circuits Conference, pp. 366-369, Sep. 2010

[17] V. Chandra, C. Pietrzyk, and R. Aitken, "On the efficacy of write-assist techniques in low voltage nanoscale SRAMs," Design, Automation & Test in Europe, pp.345-350, Mar. 2010

[18] G. K. Chen *et al.*, "Yield-driven near-threshold SRAM design," IEEE/ACM International Conference on Computer-Aided Design, pp. 660-666, Nov. 2007