Improving SRAM Vmin and Yield by Using Variation-Aware BTI Stress

Jiajing Wang*, Satyanand Nalam, Zhenyu(Jerry) Qi, Randy W. Mann, Mircea Stan, and Benton H. Calhoun
*Intel Corporation, Hillsboro, OR 97124
University of Virginia, 351 McCormick Rd., Charlottesville, VA 22904

***Abstract*- We propose a novel method that exploits BTI to partially offset variation and thus improve SRAM Vmin and yield. We show correlation between a bitcell's power-up state and its static noise margin. By applying stress with periodic re-power-up, device mismatch can be compensated by BTI induced changes. The proposed method has no extra design and area cost. It can be applied during burn-in test to offset manufacturing variation and/or used during the lifetime of the chip to offset variation from real-time aging and hence continue to improve the margins. Simulations in 45nm show that write, read, and hold Vmin at 6σ can be reduced by 128, 75, and 91 mV, respectively. Measurements from a 16Kb 45nm SRAM demonstrate the improvement of Vmin and yield.**

I. INTRODUCTION

Bias temperature instability (BTI) is one of the major aging issues that weaken transistors over time. Since SRAM is highly vulnerable to mismatch, BTI induced variation can cause SRAM yield degradation and often requires built-in guard-bands to compensate for it. Many studies have examined the impact of BTI on SRAM yield and proposed techniques to reduce BTI effects [1][2][3][4]. In lieu of only considering BTI as a variation source, we argue in this paper that it can also be a means of variation compensation. More interestingly, due to its dependence on the stress condition, we can exploit BTI to combat random variation (i.e. mismatch) by applying different stress conditions for different transistors. For an SRAM bitcell with mismatch, when we only put the stronger transistors under stress, BTI induced change can offset mismatch and thus contribute to balancing the cell and improving operation margins. Therefore, we propose to exploit BTI to improve SRAM yield. To the best of our knowledge, this is the first time that the utilization of BTI for yield and Vmin improvement is reported.

The key to offsetting mismatch with BTI is to only stress the relatively stronger transistors in the cell. An SRAM cell's power-up state due to process variation has been exploited as a random ID generator [5]. Here, we use power-up to identify mismatch and apply the power-up state as a BTI stress pattern to compensate detrimental variation. We show that the power-up state is consistent with the more stable state (in terms of static margin) determined by large mismatch in the more imbalanced cells regardless of thermal noise and power-up speed. While those more imbalanced cells are healed by BTI, the originally more balanced cells might be deteriorated by over-stress. We thus further propose to periodically re-power-up to adapt to changes in stability during stress.

The effectiveness of the BTI-based method for improving SRAM noise margins and Vmin will be quantified in this work by simulations and measurements from a 45nm bulk SRAM. One advantage of the proposed method over existing yield improvement methods is its zero design cost and easy implementation. Though various assist methods [6][7] and alternative cell topologies [8][9] have shown their effectiveness for yield and Vmin improvement, many of them require additional circuits that lead to some area, delay and/or power overhead. Careful design efforts must be taken to implement them in order to achieve better improvement while reducing overhead. In contrast, the proposed method does not need any circuit change. It only requires periodically running two easy steps: power-up the SRAM array and then stress it in standby mode. Burn-in is an integral part of chip production, which eliminates infant mortality by utilizing accelerated aging under stress conditions [10]. Our approach can be easily integrated into the burn-in test and use burn-in stress to offset mismatch. It can also be periodically used during the lifetime of the chip to offset variation from real-time aging and hence continue to improve the margins.

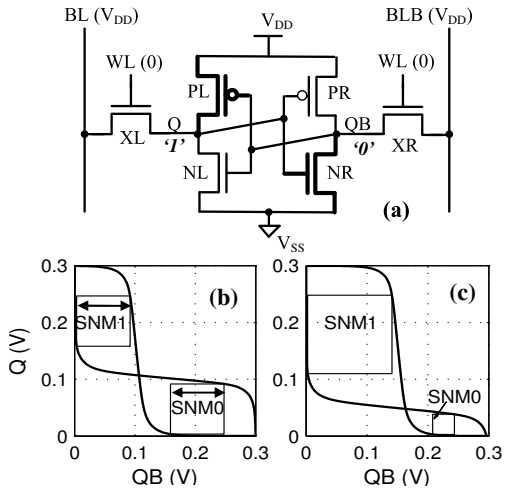

Fig. 1 (a) 6T SRAM cell; (b) SNM1=SNM0 for the balanced cell and (c) one imbalanced cell is more stable for holding '1' due to variation (i.e. SNM1>SNM0), in which state PL and NR are stressed by NBTI and PBTI respectively.

II. VARIATION-AWARENESS FROM POWER-UP STATE

Fig. 1(a) shows the schematic of a conventional 6T SRAM cell. Butterfly curve based static noise margin (SNM) is widely used to measure cell stability [11]. We denote SNM0 and SNM1 as the noise margin for '0' and '1'. The final SNM is the minimum of these two margins. Without mismatch, the SRAM cell has the same SNM for holding '0' and '1', i.e. SNM1=SNM0 as illustrated in Fig. 1(b). Local variation

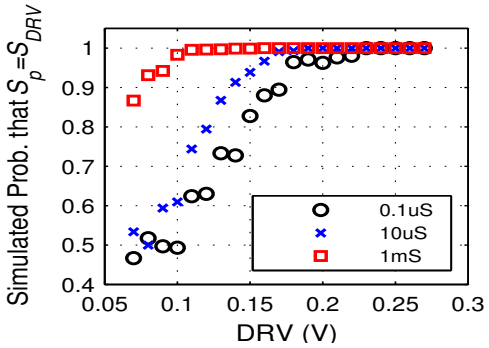

Fig. 2 Probability that S_p=S_{DRV} for different power-up speeds from simulation. Cells with more mismatches (higher DRV) always power up in their more stable state, and exhibit less sensitivity to the rate of power up.

favors one data state. For example, when the left-side inverter has a stronger pull-up FET (PL) and/or the right-side inverter has a stronger pull-down FET (NR) due to variation, the cell becomes more stable for holding '1', i.e. SNM1>SNM0 as illustrated in Fig.(c).

Power-up is a transient event, during which the cell supply voltage (V_{DD}) is raised up from 0 to the nominal value. After power-up, each SRAM cell is automatically settled to its more stable state, which is mainly determined by intrinsic mismatch and thermal noise. We denote S_p as the power-up state. The data retention voltage (DRV, the minimum V_{DD} for preserving data), is an alternative static metric for cell stability [12]. Lower DRV implies higher stability. We denote S_{DRV} as the more stable state measured by DRV (S_{DRV}=1 if DRV0<DRV1; otherwise S_{DRV}=0). Fig. 2 shows the probability that S_p=S_{DRV} from Monte Carlo simulation for different power-up speeds. More balanced cells (i.e. those with lower DRV) are more sensitive to transient sources, such as the power-up speed and thermal noise, and do not always power up in the more stable state that is solely caused by device variation. Imbalanced cells that have higher DRV due to larger mismatch consistently power up in the more stable state indicated by DRV, regardless of the V_{DD} ramp-up speed. SRAM Vmin or yield is limited by the worst cells, and powering-up consistently sets those cells in their more stable state. This ensures that the relatively stronger transistors are always stressed so that mismatch can be counteracted by BTI. In the next two sections, we analyze the impact of the proposed BTI stress on noise margins and particularly concentrate on the stability improvement for a given BTI induced V_T shift (ΔV_T).

III. Mismatch Compensation Using BTI

We first use one mismatched cell with S_p=1 as an example. Similar analysis can be performed for S_p=0. When S_p=1, as shown in Fig. 1(a), only transistors PL and NR operate in strong inversion and thus experience negative BTI (NBTI) and positive BTI (PBTI), respectively. ΔV_T of PL and NR due to BTI results in a better strength balance between the inverters and thus improves cell stability. We assume that the same amount of ΔV_T occurs in NMOS and PMOS under the same stress condition. In reality, depending on the process, NBTI and PBTI might generate different amount of ΔV_T under the

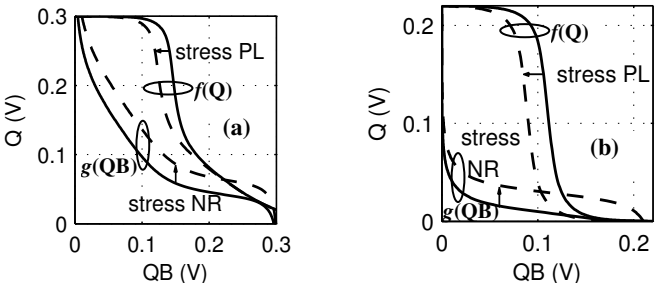

Fig. 3 Butterfly curves before and after stressing PL and/or NR for (a) read at V_{DD}=0.3V and (b) hold at V_{DD}=0.22V.

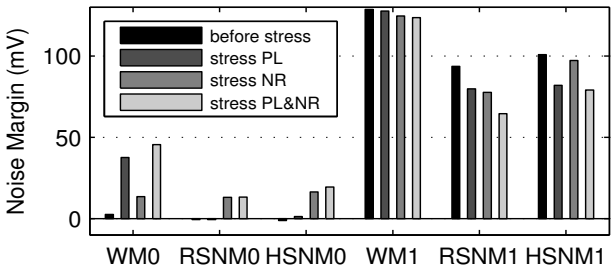

Fig. 4 With 40mV V_T shift on PL and/or NR after stress, WM0, RSNM0, and HSNM0 at low voltage increase. WM and RSNM are simulated at V_{DD}=300mV; HSNM is simulated at V_{DD}=220mV.

same temperature and voltage bias. While NBTI results in a larger P ΔV_T for thin gate oxide devices, PBTI also becomes significant for high-K devices in 45nm and beyond [13].

Fig. 3(a) plots the butterfly curves for read before and after stress. As shown with the solid curves, the cell has zero read SNM0 (RSNM0) (i.e. a read failure) before stress at V_{DD}=300mV. The dashed curves show the change when 40mV ΔV_T occurs on both PL and NR after stressing under the power-up state. The upper region of the curve f(Q) shifts to the left since PL is weakened. The lower region of f(Q) does not change since in this region current through PL is negligible and the curve shape is mainly determined by XL and NL, which are not impacted by stress with the power-up state. The curve g(QB) is primarily shifted up because NR becomes weaker. The shift of f(Q) and g(QB) creates a positive margin in the lower lobe, i.e. RSNM0>0. Similarly, Fig. 3 (b) plots the butterfly curves for hold before and after stress when V_{DD}=220mV. Before stress, hold SNM0 (HSNM0) equals to 0; after stress, f(Q) shifts left due to the weaker PL and g(QB) shifts up due to the weaker NR. This results in a positive HSNM0. Fig. 4 plots the noise margin values at low voltage (read/write at 300mV and hold at 220mV) when only NBTI or PBTI is considered (i.e. only PL or NR is stressed). One observation is that RSNM0 and HSNM0 improve little if only NBTI exists. However, they can be effectively improved by PBTI. The adoption of the high-k metal-gate devices increases PBTI effects [13] and thus SRAMs with these new high-k processes will benefit more from our approach.

We use a WL voltage sweep [14] to measure static write margin (WM). The example cell has about 4mV WM0 when V_{DD}=300mV. WM0 increases to 38mV when only considering NBTI and PL ΔV_T=40mV. It increases to 14mV when only

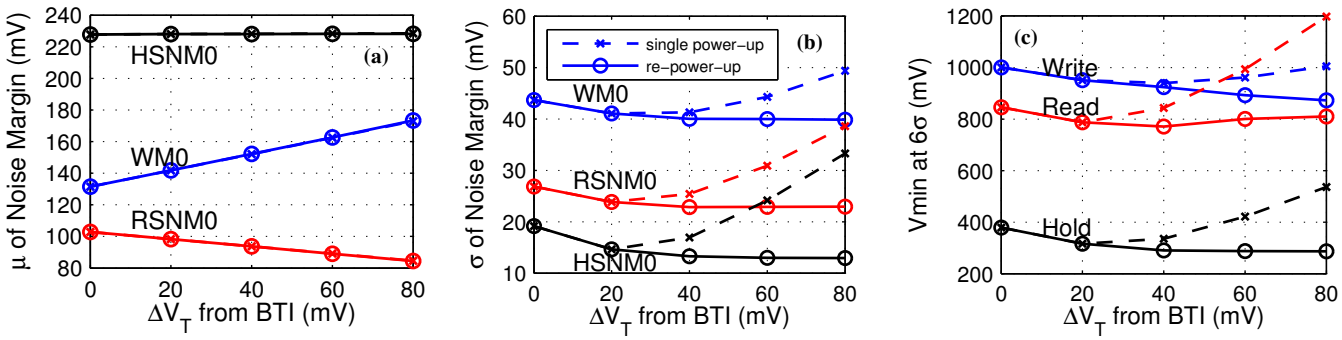

Fig. 5 Simulated (a) μ and (b) σ of WM0, RSNM0 & HSNM0 at V_{DD}=0.6V; (c) estimated Vmin at 6σ against BTI-induced ΔV_T on both NMOS and PMOS. Single power-up only improves Vmin when ΔV_T is small while periodic re-power-up improves Vmin up to 128mV.

NR ΔV_T=40mV. When NBTI and PBTI occur simultaneously, WM0 becomes 46mV. Thus the proposed stress method offers great improvement for write margin.

The key observation is that stressing in the S_p state improves strongly imbalanced cells, i.e., the yield limiting bits. Though this stress also causes slightly less stability at S_p (e.g. RSNM1, HSNM1 & WM1 decrease after stress in Fig. 4), the overall stability improves as long as it is limited by $\bar{S}_p$.

IV. Overstress Avoidance With Re-powerup

More imbalanced cells need a larger amount of ΔV_T from BTI to offset the bigger intrinsic mismatch. Simultaneously, other initially less imbalanced cells in the same array are also stressed. Those cells only need a smaller ΔV_T to achieve balance. After their initial mismatch is compensated, additional ΔV_T from BTI would actually cause new mismatch in the opposite direction and reduce noise margins. For example, if the worst cell needs 80mV BTI-induced ΔV_T to offset its mismatch, then the initially balanced cell is overstressed and 80mV mismatch is actually generated by BTI. One way to avoid damaging good cells is to identify all the bad cells and only stress bad ones. For example, we can perform column-based stress and only power-up the columns with bad cells. But this method requires more test time and control circuits. Instead, we propose a simple and efficient method to reduce the overstress risk. We periodically re-power-up the cell to set it to its most stable state before continuing stress, which adapts to changes in stability. An even number of re-power-up operations can cancel out the stress effects on the initially balanced cells because the left and right devices are alternately stressed. Next, we will statistically quantify the margin improvement and compare the two BTI stress methods: stress with a single-power-up and stress with periodic re-power-up.

Fig. 5(a) and (b) show the simulated mean (μ) and standard deviation (σ) of WM0, RSNM0, and HSNM0 at V_{DD}=0.6V from Monte Carlo versus the total amount of BTI-induced ΔV_T for $\Delta V_T \leq$80mV. Similar to the previous analysis, we assume N and P devices experience the same amount of ΔV_T after stress. Note for the re-power-up method, ΔV_T might be distributed on both sides of devices. To generate the same amount of ΔV_T, the stress time with re-power-up would be longer than single-power-up since recovery occurs between re-power-ups. In this simulation, we assume re-power-up occurs after every 20mV of ΔV_T. μ of noise margin in the re-power-up mode is similar with that in the single power-up mode. σ of noise margin in the single power-up mode decreases when ΔV_T is small, but it increases and finally exceeds the before-stress value when ΔV_T becomes larger. As expected, overstress limits the achievable improvement without re-power-up. However, with the proposed re-power-up method, σ further decreases at larger ΔV_T, resulting in more improvement in yield. Note that the pull-up FETs (PL & PR) are indeed gradually weakened while mismatch between them is offset by NBTI. The pull-down FETs (NL & NR) are also more balanced but weaker due to PBTI. However, the pass-gate FETs (XL & XR) do not experience stress. μ_{WM} increases with ΔV_T because the strength ratio between the pass-gate and the pull-up FETs increases while μ_{RSNM} decreases because the strength ratio between the pull-down and the pass-gate FETs decreases. Since HSNM is more sensitive to the relative strength between the pull-up and pull-down FETs, μ_{HSNM} remains almost the same. Nevertheless, σ for read, write, and hold noise margin all decrease at larger ΔV_T with the re-power-up method. For the worst case at 6σ (i.e. μ-6σ), WM0, RSNM0 and HSNM0 improve 49.6%, 8.4% and 33.4%, respectively. By using a statistical model for SRAM Vmin [12], we can also estimate Vmin values for larger SRAMs. Fig. 5(c) shows that the stress with re-power-up method can reduce write, read, and hold Vmin at 6σ by 128, 75, and 91mV, respectively.

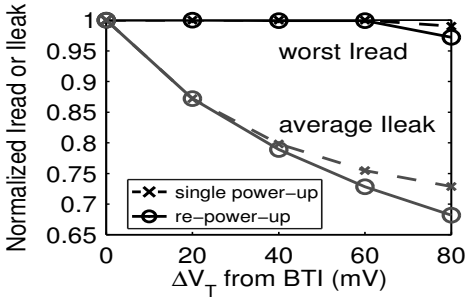

Fig. 6 Simulated worst Iread is degraded by 4% while the average Ileak improves by up to 31% after the stress with re-power-up.

We also evaluate the impact on leakage current (Ileak) and read current (Iread) and plot the simulation results in Fig. 6.

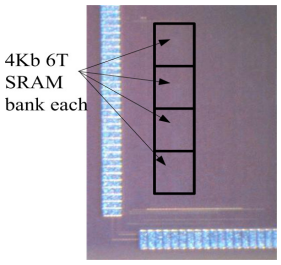

Fig. 7 Die photo of the 45nm test chip with a 16Kb 6T SRAM.

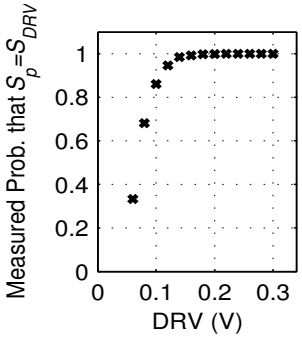

Fig. 8 Measured probability that S_p=S_{DRV} against DRV.

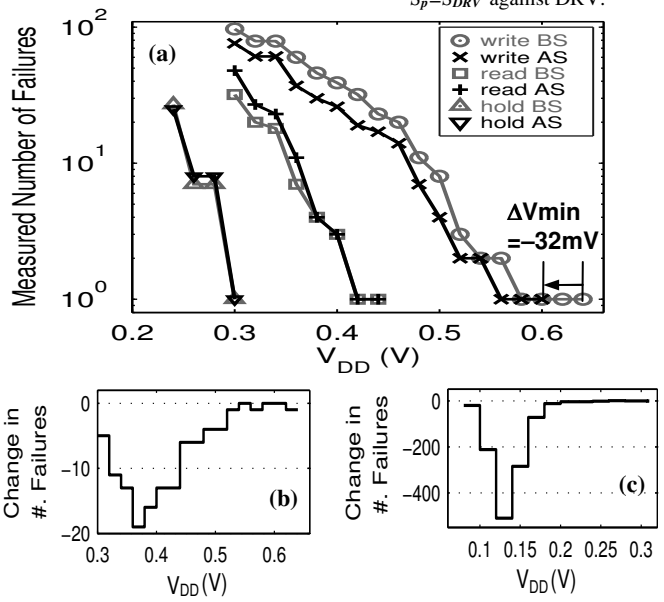

Fig. 9 (a) Measured failures before stress (BS) and after stress (AS) for write, read, and hold; (b) total failures and (c) hold failures vs. voltage decrease AS.

The average Ileak at nominal V_{DD} is reduced by up to 31%. The worst Iread is only slightly degraded (<4%) because it correlates more strongly with the unchanged pass-gate NMOS.

Note that we assume all the transistors see the mean BTI-induced ΔV_T change. We do not consider BTI induced ΔV_T variance, which has been shown to relate to the mean of BTI-induced ΔV_T and gate area for NBTI [3]. More accurate analysis requires accurate BTI models. It is also interesting to quantify the improvement in SNM as a function of specific stress conditions (e.g. stress time, voltage, and temperature). These are left for future work.

V. 45NM TEST CHIP MEASUREMENTS

A custom 16Kb 6T SRAM (logic rule compliant) chip (Fig. 7) is implemented in a 45nm bulk CMOS technology. The chip is stressed at 1.7× of the nominal V_{DD} and 45°C for 24 hours with periodic re-power-up every 2 hours. Fig. 8 shows that the measured S_p matches the measured S_{DRV} especially for those cells with larger DRV, verifying Fig. 2. Hence those cells are correctly stressed with their more-stable state after power-up. Fig. 9(a) shows the measured write, read, and hold failures versus V_{DD} before and after stress. The first failure occurs on write, and the chip Vmin drops 32mV after stress. The write failures are reduced across the entire V_{DD} range. The voltages of the first read and hold failure remain almost unchanged. This might be caused by the insignificant PBTI effect in this 45nm technology which employs SiO_2 devices. As shown in Fig. 4, read and hold SNM are much more sensitive to PBTI than NBTI. We expect that more improvement for read and hold can be achieved for SRAMs in high-K metal-gate processes. As estimated in Fig. 5(c), more BTI-induced ΔV_T for larger SRAMs can reduce Vmin more. Fig. 9(b) shows that the measured total failures are reduced across the range of 0.3V to 0.64V. Fig. 9(c) shows that the measured hold failures at lower V_{DD} also decrease.

VI. CONCLUSION

We propose to compensate SRAM mismatch by using a power up test to identify and allow selected stress on the yield limiting bitcells. Repeating power-up during stress can adapt to changes in cell stability and avoid overcompensation. This method improves SRAM Vmin/yield and leakage current with only slight delay degradation and no area cost. It can be applied during burn-in to offset manufacturing variation or periodically used during the lifetime of the chip to offset variation from real-time aging and hence continue to improve the margins.

ACKNOWLEDGMENT

The authors thank Freescale Semiconductor, Inc. for chip fabrication.

REFERENCES

[1] A.T. Krishnan, V. Reddy, D. Aldrich et al, "SRAM cell static noise margin and Vmin sensitivity to transistor degradation," IEDM'06, Dec. 2006, pp. 1–4.

[2] J.C. Lin, A. S. Oates, and C. H. Yu, "Time dependent vccmin degradation of SRAM fabricated with high-k gate dielectrics," IRPS'07, pp.439–444.

[3] G. L. Rosa, W. L. Ng, S. Rauch et al, "Impact of NBTI induced statistical variation to SRAM cell stability," IRPS'06, pp. 274-282.

[4] S.Kumar *et al.*, "Impact of NBTI on SRAM read stability and design for reliability," ISQED, 2006.

[5] Y. Su, J. Holleman, and B. Otis, "A 1.6pj/bit 96% stable chip-id generating circuit using process variations," ISSCC 2007, pp. 406–407.

[6] S. Ohbayashi et al., "A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits," IEEE J. Solid-State Circuits (JSSC), vol. 42, no. 4, 2007.

[7] H. Pilo, C. Barwin, G. Braceras et al., "An SRAM design in 65nm and 45nm technology nodes featuring read and write-assist circuits to expand operating voltage," IEEE Symp. on VLSI Circuits, 2006, pp. 15–16.

[8] Y. Morita, H. Fujiwara, H. Noguchi et al, "An area-conscious low-voltage-oriented 8T-SRAM design under DVS environment," IEEE Symp. on VLSI Circuits, 2007, pp. 256–257.

[9] S. Nalam and B. H. Calhoun, "Asymmetric sizing in a 45nm 5T SRAM to improve read stability over 6T," In CICC, pp. 709–712, Sept. 2009.

[10]R.-P. Vollertsen, "Burn-in," IEEE International Integrated Reliability Workshop Final Report, pp. 167--173, 1999.

[11]E. Seevinck, F. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," IEEE JSSC, vol. 22, no. 5, pp. 748–754, 1987.

[12]J. Wang, A. Singhee, R. A. Rutenbar, and B. H. Calhoun, "Statistical modeling for the minimum standby supply voltage of a full SRAM array," ESSCIRC, 2007, pp. 400–403.

[13] S. Pae, M. Agostinelli, M. Brazier et al, "BTI reliability of 45 nm high-k + metal-gate process technology," IRPS'08, pp. 352–357.

[14] Z. Guo,.A. Carlson, L.-T. Pang et al, "Large-scale read/write margin measurement in 45nm CMOS SRAM arrays," IEEE Symp. on VLSI Circuits, Jun. 18–20, 2008, pp. 42–43.