Crosshairs SRAM – An Adaptive Memory for Mitigating Parametric Failures

Gregory Chen, Michael Wieckowski, David Blaauw, Dennis Sylvester
University of Michigan, Ann Arbor, Michigan
{grgkchen, wieckows, blaauw, dmcs}@umich.edu

***Abstract*— We propose Crosshairs SRAM to adaptively fix parametric failures and increase yield. It mitigates process variation by tuning VDD and GND of each bitcell inverter independently from its cross-coupled counterpart. It targets failing cells at the intersection of individually-tuned orthogonal VDD and GND rails. We implement 70 32kb test arrays in 45nm CMOS with little modification to a commercial 6T design and no increase in bitcell area. Crosshairs improves performance by 13% and fixes an average of 70% of parametric failures for reasonable initial failure rates lower than 0.1%.**

I. INTRODUCTION

The current trends of larger caches and greater processor parallelism increase the amount of SRAM per chip, making SRAM failures a dominant factor in processor yield. Technology scaling increases parametric failures (PFs), including timing and stability failures, due to excessive process variation. For example, shrinking devices amplify the effects of random dopant fluctuation [1], and lithographic double patterning increases gate length variation [2]. As a result, SRAM requires higher levels of error correction coding (ECC) [3] and redundancy [4] to satisfy yield requirements. We propose the Crosshairs method to detect and adaptively correct PFs. Crosshairs tunes the SRAM's power and ground supply networks to mitigate excessive variation. It improves yield with respect to timing and stability constraints. The Crosshairs bitcell has the same area, transistors, and number of metal layers as a commercial design.

II. CROSSHAIRS SRAM METHOD

A. *Controlling Bitcell Power Supplies*

To identify PFs, a BIST performs March tests on the SRAM. When it detects a failure, it determines the nature of mismatch in the bitcell by checking if write-ZERO/read-ZERO or write-ONE/read-ONE accesses failed. The BIST then uses this information to tune VDD and GND of each bitcell inverter with respect to its cross-coupled counterpart, cancelling process variation and restoring bitcell functionality.

Each bitcell has connections to left and right vertical power rails (VDDL and VDDR) and horizontal ground rails (GNDL and GNDR). It is identical to a commercial differential 6T design except that the vertical VDD rail is split into VDDL

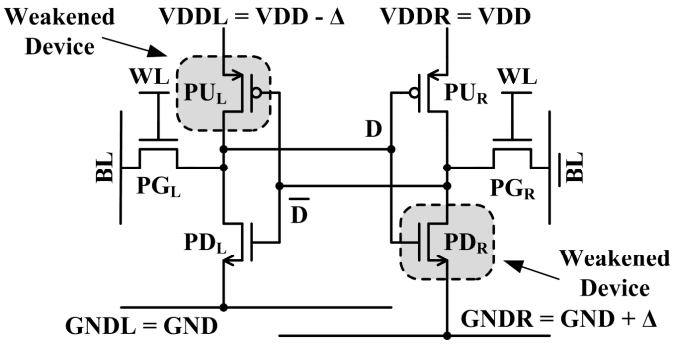

Figure 1. Crosshairs recovers parametric failures (PFs) by separately tuning the VDD and GND supplies of each inverter within a bitcell.

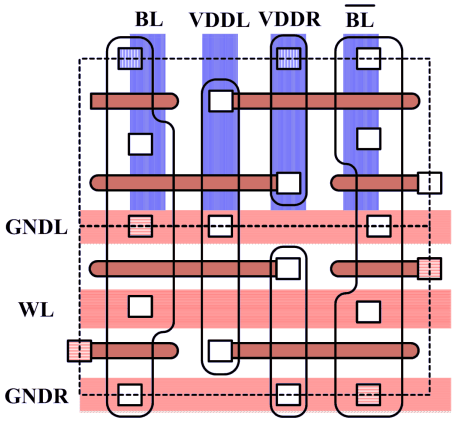

Figure 2. The Crosshairs bitcell is a minimally modified commercial differential 6T design that does not require larger area or more metal layers.

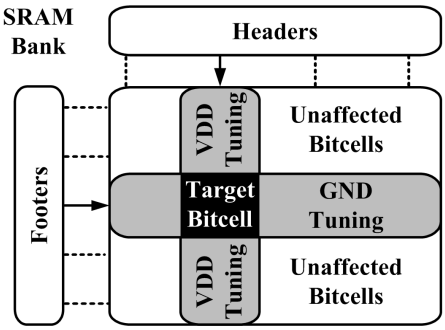

Figure 3. Crosshairs shares VDD within a column and GND within a row. The orthogonal, tuned supplies target failing bitcells at their intersection.

and VDDR (Fig. 1). This modification does not require a larger bitcell or more metal layers (Fig. 2).

Crosshairs eliminates PFs by adjusting VDD in the column and GND in the row where a PF occurs (Fig. 3). Thus, the orthogonal tuned supply rails target PFs at their intersection. Each column shares VDDR and VDDL rails and adjacent rows share GNDR and GNDL rails (Fig. 4). To tune VDD, PMOS headers connect VDDR and VDDL to one of two global power supplies (VDD_HI and VDD_LO). Similarly, NMOS footers connect GNDR and GNDL to either GND_HI or GND_LO (Fig. 5). The BIST generates control signals for the headers and footers. An on-chip linear regulator can generate the global VDDs and GNDs. The voltage difference between these global supplies is defined as the Crosshairs tuning voltage.

B. *Fixing Parametric Failures*

Tuning the supplies of each bitcell inverter with respect to its cross-coupled counterpart cancels process variation and eliminates PFs. Initially the stronger VDD and GND (VDD_HI and GND_LO) supply all bitcell inverters. When writing a ZERO to node D in Fig. 1, the left pass gate (PG_L) overpowers the left pull up (PU_L), pulling D low enough to initiate the write mechanism. Process variation can create a write PF by making PG_L too weak with respect to PU_L. To increase write margin and fix this PF, Crosshairs weakens PU_L by connecting VDDL to VDD_LO and GNDR to GND_HI, reducing the likelihood of a write PF by 9× based on importance sampling Monte Carlo SPICE simulations [5].

When reading a ZERO from node D, charge from the bitline (BL) is injected onto D, potentially causing a read-upset PF by overwriting the value to a ONE. This is more probable if process variation causes the left pull down (PD_L) to be too weak with respect to PG_L or the timing constraint. To correct this, Crosshairs weakens the right pull down (PD_R) device by connecting GNDR to GND_HI. Similarly the left pull up (PU_L) device is weakened by connecting VDDL to VDD_LO. In this configuration the bitcell holds a stronger ZERO and the probability of a read PF decreases by 3×, based on importance sampling [5]. A larger simulated static noise margin (SNM) reflects the increase in read stability (Fig. 6a).

As seen from the previous examples of write and read PFs, Crosshairs uses the same voltage configuration to improve both read-ZERO and write-ZERO margins. Similarly, lowering VDDR and raising GNDL increases read-ONE and write-ONE margins. Thus, to properly apply the Crosshairs algorithm, the BIST does not need to determine whether the write or read access failed. Rather, it must determine only whether a ONE or ZERO access failed. This allows the BIST to gather all the information it needs about process variation in the SRAM array using simple March test algorithms.

C. *Net Reduction in SRAM failures*

Crosshairs tuning can positively or negatively impact non-PF cells in the same column or row as a PF. It adjusts the VDD rails in each column based on the process variation in the PF cell, which does not necessarily reflect the variation in the

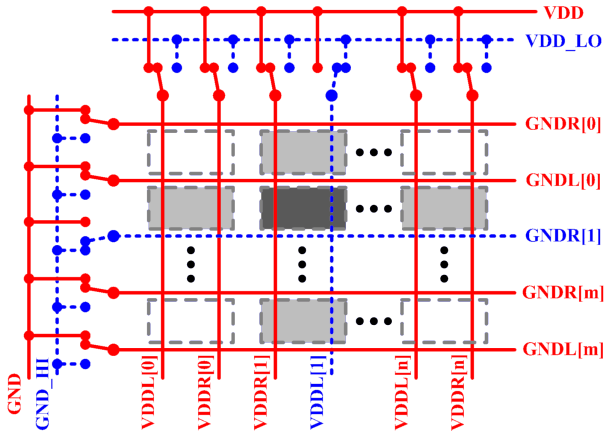

Figure 4. Header and footer cells control VDD and GND potentials, increasing array effiiciency.

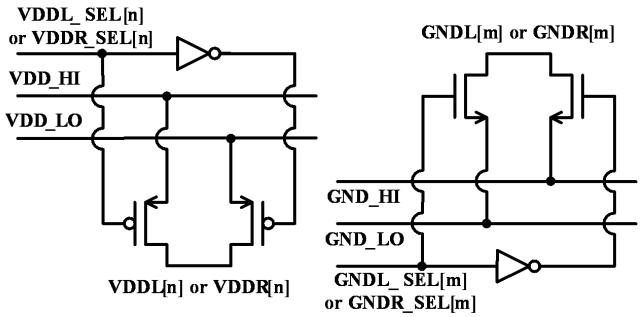

Figure 5. Headers connect each VDD column (VDDL, VDDR) to one of two global supplies (VDD_HI, VDD_LO). Simiarly footers connect each GND row (GNDL, GNDR) to global grounds (GND_HI, GND_LO).

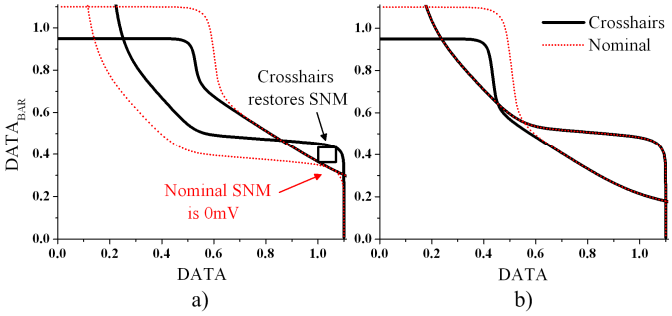

Figure 6. a) Crosshairs restores simulated read SNM in a PF cell. b) It has less impact on non-PF cells in the same column or row as a PF cell because either VDD or GND is tuned but not both.

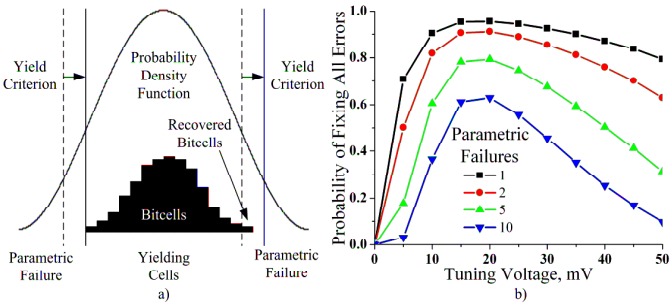

Figure 7. Crosshairs has a high probability of fixing PFs and a low probability of creating PFs, resulting in a net reduction in simulated failures.

other cells. Similarly, Crosshairs will impact the GNDs of cells in the same row as a PF cell. However, Crosshairs will not tune both VDD and GND in non-PF cells, greatly decreasing the potential negative impact on stability margins (Fig. 6b).

The distribution of each bitcell with process parameters is weighted at the mean, with few cells at tails of the distribution that fail yield critera (Fig. 7a). Applying Crosshairs tuning relaxes the yield criterion on one end of the distribution, but tightens it at the other tail. The conditional PDFs for PF and non-PF cells dictate that Crosshairs is likely to fix a given PF, but unlikely to cause new errors. This probability is calculated using importance sampling as 95.5% for a 128x256 array with a 20mV tuning voltage [5] (Fig. 7b) .

D. Header and Footer Sizing

Headers and footers require proper sizing to prevent IR drop from impacting robustness but should be small for a low area overhead. Fig. 8 shows a simulation demonstrating the affect of these sizes on robustness. The plot shows robustness in terms of the maximum V_{TH} mismatch that the bitcell can tolerate without functional failure. For each VDD column, we select a header width of 2× the bitcell PU device. Further increasing header size achieves only modest improvements in stability. Crosshairs requires only a small header since only one accessed cell per column draws current from the VDD rail. However, every bit can simultaneously draw current from the same GND rail. As such, the footer size is 2× the total PD width for one word. The presented array uses a 128-bit word and footer size decreases proportionally with word length.

III. MEASUREMENT RESULTS

We fabricated and measured 70 chips with 128x256 32kb Crosshairs SRAM banks in a 45nm CMOS process (Fig. 9). We designed Crosshairs with feedback from the foundry to violate logic design rules, as is typical for SRAM. This allows the bitcell to match the area of a commercial differential 6T SRAM design. Crosshairs decreases array efficiency by 12.5% because of additional peripheral circuits.

A. Recovering Timing Failures

Local process variation creates slow bitcells, which then dictate the overall performance of an SRAM array. By targeting slow bitcells, Crosshairs mitigates process variation and increases array performance by 13% at a tuning voltage of 20mV (Fig 10). It achieves the optimal performance at a tuning voltage that also minimizes the number of simulated and measured stability failures. It creates these performance gains with less than a 2.5% leakage overhead.

B. Recovering Stability Failures

We record bitcell functionality with no latency requirement for 70 test chips. To measure the impact of Crosshairs on stability, we must first observe some initial PFs. Since nominal PFs are rare, for testing purposes we artificially generate PFs through VDD scaling. Then we recover the resulting PFs using Crosshairs to demonstrate the method's effectiveness. Crosshairs fixes 9 out of 9 initial PFs in an array

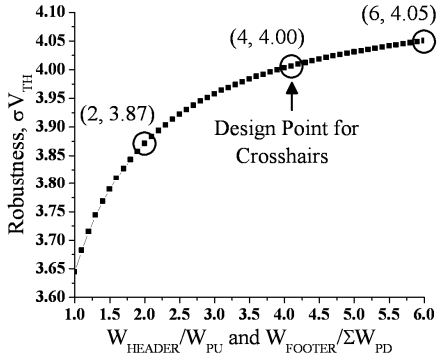

Figure 8. Headers and footers are optimally sized to create stable virtual supply rails without excess area overhead, based on simulations.

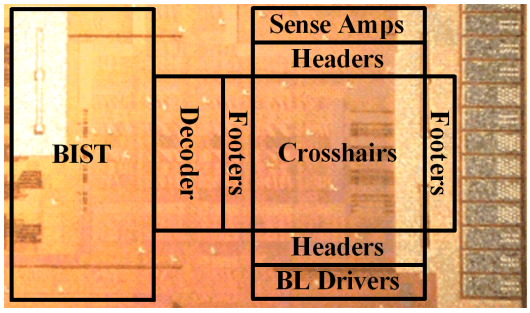

Figure 9. 45nm chip micrograph including a 32kb SRAM array and BIST.

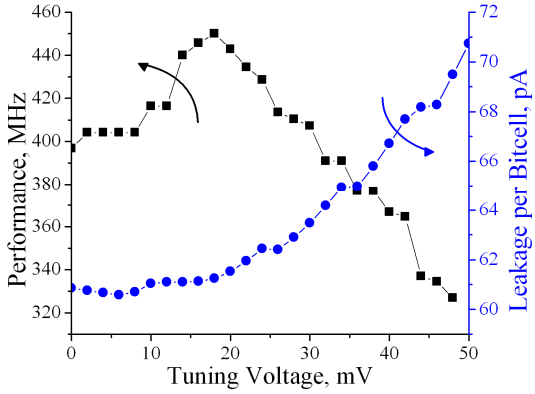

Figure 10. Measured results show that Crosshairs can improve array performance by 13% and has a modest leakage overhead.

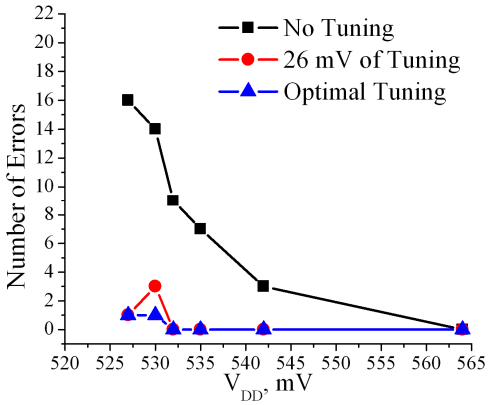

Figure 11. Crosshairs recovers measured PFs induced by VDD scaling from the nominal voltage of 1100mV. Using a fixed tuning voltage reduces the overhead for power supply generation with little decrease in effectiveness.

at a VDD of 532mV with a 26mV tuning voltage (Fig. 11). It recovers 15 out of 16 PFs at a VDD of 526mV. The optimal Crosshairs tuning voltage is between 20 and 26mV (Fig. 12). These voltages are high enough to fix PFs without creating new failures in tuned rows or columns. In this tuning range, Crosshairs fixes nearly all PFs in arrays with between 1 and 16 initial VDD-scaling-induced PFs.

ECC and redundancy can fix a limited number of PFs based on the spatial distribution of failures (Fig. 13). Single-error-correct double-error-detect (SECDED) ECC can only fix one PF per word. Higher levels of ECC incur additional area and performance penalties. One measured SRAM array did not yield with ECC. Crosshairs recovers this array at tuning voltages ranging from 10mV to 50mV. Using redundancy, each column or row with a PF requires an additional redundant row or column, incurring an area and complexity penalty. Crosshairs with a tuning voltage of 26mV reduces the average number of required redundant rows for 100% yield from 4.56 to 1.95. It reduces the required number of redundant columns from 4.46 to 1.91. In addition, the proposed method can be used on top of ECC and redundancy.

Fig. 14 presents the number of recovered PFs versus initial PFs for all measured chips with a fixed tuning voltage of 26mV. The number of chips at each data point is represented by circle size. Crosshairs fixes an average of 70% of PFs for reasonable initial failure rates lower than 0.1%.

IV. Conclusions

Crosshairs recovers 70% of PFs in 70 128x256 test arrays by tuning VDD and GND of each SRAM bitcell inverter with respect to its cross-coupled counterpart. These gains are achieved with little modification to a commercial 45nm 6T design and no increase in bitcell area. Crosshairs increases SRAM yield and eliminates or reduces the overheads for other yield improvement techniques such as ECC or redundancy.

Acknowledgements

The authors thank STMicroelectronics for fabrication and support of this project.

References

[1] A. Agarwal, B.C. Paul, S. Mukhopadhyay, and K. Roy, "Process variation in embedded memories: failure analysis and variation aware architecture," *IEEE Journal of Solid-State Circuits,* vol.40, no.9, pp. 1804- 1814, Sept. 2005.

[2] K. Jeong, and A.B. Kahng, "Timing analysis and optimization implications of bimodal CD distribution in double patterning lithography," *Asia and South Pacific Design Automation Conference,* pp.486-491, Jan. 2009.

[3] C.L. Chen, and M.Y. Hsiao, "Error-Correcting Codes for Semiconductor Memory Applications: A State-of-the-Art Review," *IBM Journal of Research and Development,* vol.28, no.2, pp.124-134, March 1984.

[4] J. P. Bickford, R. Rosner, E. Hedberg, J.W. Yoder, and T.S. Barnett, "SRAM Redundancy - Silicon Area versus Number of Repairs Trade-off," *Advanced Semiconductor Manufacturing Conference,* pp.387-392, May 2008.

[5] G.K. Chen, D. Blaauw, T. Mudge, D. Sylvester, and N.S. Kim, "Yield-driven near-threshold SRAM design," *IEEE/ACM International Conference on Computer-Aided Design,* pp.660-666, 4-8 Nov. 2007.

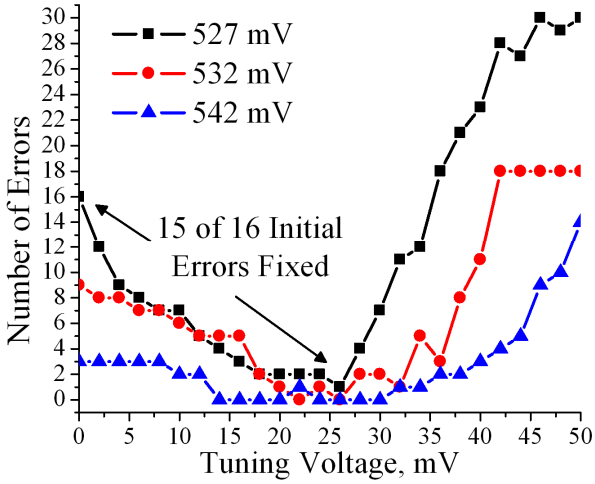

Figure 12. The measured optimal tuning voltage is between 20 and 26mV.

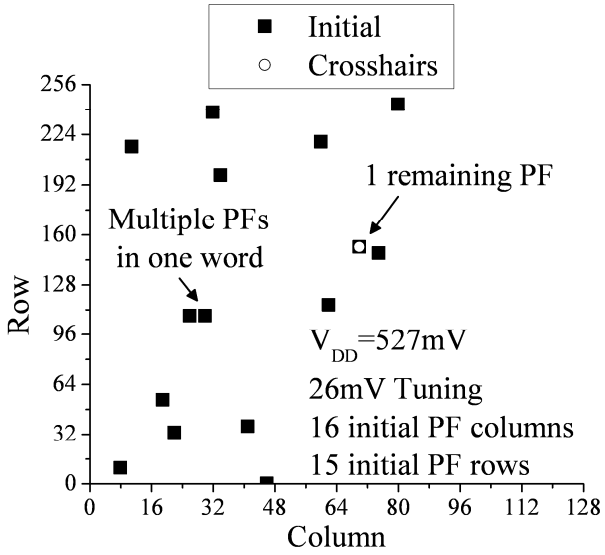

Figure 13. Crosshairs can eliminate multiple PFs within one word, unlike SECDED ECC. It can also eliminate PFs spanning many SRAM columns and rows, which would require many redundant columns or rows.

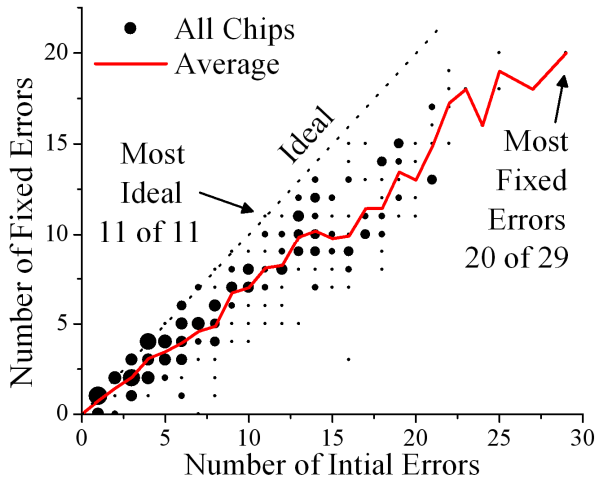

Figure 14. Crosshairs reduces the number of PFs in 70 32kb test arrays using a fixed 26mV tuning voltage. It recovers an average of 70% of PFs for reasonable initial failure rates lower than 0.1%.