A 256kb 6T Self-Tuning SRAM with Extended 0.38V-1.2V Operating Range using Multiple Read/Write Assists and \( V_{\text{MIN}} \) Tracking Canary Sensors

Arijit Banerjee, Ningxi Liu, Harsh N. Patel, and Benton H. Calhoun
University of Virginia
Charlottesville, VA 22904, USA
{ab9ca, nl6cg, hnpatel, bcalhoun}@virginia.edu

Abstract—A closed loop self-tuning 256kb 6T SRAM with 0.38V-1.2V extended operating range using combined read and write assists and canary-based \( V_{\text{MIN}} \) tracking is presented. 337X and 4.3X power reductions are achieved using multiple assists and \( V_{\text{MIN}} \) tracking, respectively; combining both saves 1444X in active power and 12.4X in leakage at the 0.38V.

Keywords—self-tuning SRAM; combined assists; canary SRAM, \( V_{\text{MIN}} \) tracking;

I. INTRODUCTION

This paper presents an adaptive, closed loop memory system that leverages combinations of bias-based peripheral assists (CPA) for both read and write to expand the operating range of a 256kb 6T SRAM by over 67% to cover from 1.2V down to 0.38V. Assists are used in reverse to tune canary bitcells that allow a closed loop control of the \( V_{\text{DD}} \) to track the minimum operating voltage (\( V_{\text{MIN}} \)) at a desired operating frequency. The design uses CPA together with canary based \( V_{\text{MIN}} \) tracking to maximize the operating range that is compatible with the sub-threshold logic (6T SRAM usually has higher \( V_{\text{MIN}} \) than logic circuits across process, voltage, and temperature (PVT) variations [1][2][3][4]) and to minimize guard-banding. The design is thereby optimized for meeting the low power, and varying frequency needs of highly variable Internet of Everything (IoE) applications while retaining the density of 6T cells.

Since battery powered or energy harvested IoE devices mostly operate at lower frequencies (~10 kHz to 10 MHz) [5][6], there is a need to expand the 6T SRAM operating range to lower voltages to achieve low power operation. Bias-based assist techniques can lower SRAM \( V_{\text{MIN}} \) [1][2][4], but selecting the best CPA depends on the \( V_{\text{DD}} \) and can affect the power / performance tradeoff. Fig 1 (a) shows the measured cumulative distribution functions for the SRAM with three peripheral assists: (1) \( V_{\text{DD}} \) boosting (VDDB) for low-voltage readability and half-select [1][4] read-stability; (2) wordline (WL) boosting (WLBB); and (3) negative bitline (NBL) for writeability. Using all the three assists achieves 240mV of \( V_{\text{MIN}} \) improvement (at 90th percentile) and beats using other single or combinations (Fig 1 (a)), but using fewer assists can save power overhead when the target \( V_{\text{DD}} \) is higher for a given frequency.

Fig 1 (b) shows the measured Shmoo plot with the CPA-extended range highlighted for the 256kb SRAM. Using assists alone requires guard-banding to ensure that all chips function across PVT, reducing the potential power savings.

To maximize the benefits of CPA, runtime SRAM \( V_{\text{MIN}} \) determination [7] reduces the guard-banding of SRAM \( V_{\text{MIN}} \) at a given frequency. However, this technique has a huge penalty in the number of cycles for writing and reading the whole SRAM and in total energy for using a built-in-self-test (BIST). On the other hand, a smaller sized canary SRAM based \( V_{\text{MIN}} \) tracking [8] enables each chip to function at or near its \( V_{\text{MIN}} \) for much lesser clock cycles and energy.

A. Block Diagram of the System

Fig 2 (b) shows our full SRAM system comprising a 256kb SRAM in 4 sub-arrays (mats) each with 4 banks of 128x128 6T bitcells and 1 row of 128 canary bitcells per bank (2kb canary bitcells total), an assist controller (ASC), a frequency-to-digital converter (FDC), and a built-in self-test (BIST) block for the core SRAM and the canary bitcells (CBIST). The canary cells share the peripheral circuits such as write drivers, sense amplifiers, precharge circuits etc. with the SRAM array but have dedicated reverse assist (RA) controls [8] that tune writeability and readability of the canaries by degrading the canary WL signal using eight programmable settings.

John Poulton and C. Thomas Gray
Nvidia Corporation
Durham, NC 27713, USA
{jpoulton, tgray}@nvidia.com

Fig. 1. a) Measured CDF of 256kb SRAM \( V_{\text{MIN}} \) showing 90th percentile \( V_{\text{MIN}} \) improvement of 240mV using combined assists [\( V_{\text{DD}} \) boosting (VDDB), WL boosting (WLBB), negative bitline (NBL)], and b) measured \( V_{\text{DD}} \) Shmoo.
The CBIST tests the canary to provide the number of failures to the ASC.

B. Self-tuning Strategy of the System

Fig 2 (c) and (d) present the self-tuning strategy for canary-based SRAM $V_{\text{MIN}}$ tracking and dynamic control over assists, and $V_{\text{DD}}$ selection. When tuning is enabled (TRACK=1), the FDC converts the input clock (CLK_IN) frequency to a 16-bit digitized output (FDCOUT) and initializes an (off-chip) Low-Dropout (LDO) regulator to an initial $V_{\text{DD}}$ for the given frequency. Then, the ASC chooses an assist configuration for the current $V_{\text{DD}}$ from a look-up table (LUT) for flexibly optimizing assist selection based on measured characterization across $V_{\text{DD}}$. The ASC then iterates to find the target $V_{\text{MIN}}$ for the given frequency based on the canary outputs. The CBIST executes canary write and read operations across all canary addresses, calculates the number of canary failures ($F_c$), then compares $F_c$ with a canary failure threshold value ($F_{\text{th}}$) to generate a pass/fail signal (SPF). If the CBIST passes, the ASC reduces $V_{\text{DD}}$ by changing a 4-bit signal (LDOCTRL) controlling the off-chip LDO. The ASC repeats this process until the CBIST fails, then it raises $V_{\text{DD}}$ to the last operational $V_{\text{DD}}$, which completes the closed-loop tracking for $V_{\text{MIN}}$. The SRAM retains its data through this process, and tuning can be re-run when the frequency changes or to periodically adjust for temperature changes.

Fig. 3. Experimental setup for the chip measurements.
The assists expand the operating range, the canary feedback is critical to ensure that $V_{DD}$ scaling stops before the core SRAM fails. The RA [8] forces failures in canary bits ahead of the core bits at eight programmable reverse assist settings (RAS). Since the canaries are core SRAM cells with RAS applied, their failure distribution is a shifted version of the core cells that tracks with frequency and temperature [8]. This allows us to set $F_{th}$ based on CBIST results from a few dies to calibrate the canary failures relative to the core SRAM cells, thus all the chips are able to track their $V_{MIN}$. 

### III. EXPERIMENTAL SETUP

Fig 3 shows the experimental setup for the measurement of data. A DC voltage source supplies power to the SRAM PCB. The digital pattern generator (PGLA) generates a waveform that controls the SRAM chip. An external clock source drives the PGLA to generate a clock signal to the PCB and the chip. Overall, the PGLA is controlled by a laptop computer for waveform generation and data collection.

### IV. MEASUREMENT AND RESULTS

Fig 4 (a), (b), and (c) show the measured tunable range for canaries and the SRAM $V_{MIN}$ across temperature and frequency. Fig 4 (d) shows the distribution of the $V_{MIN}$ reduction using CPA and $V_{MIN}$ tracking across 30 dies. The ASC sets the $F_0$ and uses an LUT to select the RAS and sense amplifier delay based on the current $V_{DD}$, which allows the user to tradeoff guard-band margin with power savings. Fig 4 (a), (b), and (c) show settled $V_{MIN}$ values based on settings that aggressively reduce the margin to maximize the power savings, but the flexible system allows including an arbitrary margin.

CPA and canary based $V_{MIN}$ tracking work together to allow each chip self-tuning to its $V_{MIN}$ for a given frequency, including expanded operating range and power savings for low $V_{DD}$ IoE applications. Fig 2 (a) shows an annotated die photo of the SRAM chip. The area overheads are 0.77% for the canary bits in each SRAM bank and 1.8% for the system components without BISTs in this design. The combined assist overhead in the SRAM is less than 2.8%.

Fig 5. a) Measured active power reduction of SRAM and BIST with combined peripheral assists and $V_{MIN}$ tracking, and b) measured leakage reduction from $V_{DD}$ scaling.

The combined assist overhead in the SRAM is less than 2.8%. Fig 5 (a) shows the power savings from the combined approach, which extends the operating range down to 0.38V, and gives a 12.4X lower (Fig 5 (b)) leakage power (9.5pW/bit) than at 1.2V. If canary tracking were not available, process variation would require $V_{DD}$ scaling to stop at 0.47V to ensure all chips work (achieves 337X active power reduction using CPA for SRAM and BIST), but $V_{MIN}$ tracking allows an extra 4.3X power reduction by removing the
guard-band for those chips that can function at lower V<sub>DD</sub> (up to 1444X active power savings (Fig 5 (a)). Table I includes the power breakup of the SRAM and BIST in the chip. It shows that these techniques reduce the SRAM power from 14.4mW to 11.4µW (1263X power reduction). Table II compares this work to recent wide voltage range SRAMs for low power applications. Fig 6 (a) and (b) show that the canary based V<sub>MIN</sub> tracking is scalable to 45nm and 32nm technologies for a wide range of voltages and frequencies.

### TABLE I. POWER BREAKUP FOR THE SRAM AND BIST.

<table>
<thead>
<tr>
<th>Supply (V)</th>
<th>SRAM and BIST Power</th>
<th>SRAM Power</th>
<th>BIST Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>18.3mW</td>
<td>14.4mW</td>
<td>3.9mW</td>
</tr>
<tr>
<td>0.47</td>
<td>54.3µW</td>
<td>49.7µW</td>
<td>4.6µW</td>
</tr>
<tr>
<td>0.38</td>
<td>12.5µW</td>
<td>11.4µW</td>
<td>1.2µW</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

This chip extends the 6T SRAM operating range by over 67% (from 1.2V-0.71V=0.49V to 1.2V-0.38V=0.82V, in sub-threshold) using three combined read/write assists and canary-based V<sub>MIN</sub> tracking. The SRAM self-tunes to the V<sub>MIN</sub> across process and temperature for a target frequency. This adaptive solution enables a range of IoT applications and achieves up to 1444X active power reduction. Our canary based V<sub>MIN</sub> tracking technique is scalable to 45nm and 32nm technologies.

ACKNOWLEDGMENT

This work was funded in part by NVIDIA through the DARPA PERFECT program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES


