An Improved Canary-Based System with BIST for SRAM Standby Power Reduction

B. RAMESH
M. Tech Dept. of VLSI System Design
Sri Krishna Devaraya Engineering College Gooty, Anantapur

T. MAHABOOB DOULA
Assistant Professor,
Dept. of VLSI System Design
Sri Krishna Devaraya Engineering College Gooty, Anantapur

Abstract:
This paper presents a new technique to accurately measure the achieve aggressive standby power reduction for static random access memory (SRAM), we have previously proposed a closed-loop V_{DD} scaling system with canary replicas that can track global variations. In this paper, we propose several techniques to enhance the efficiency of this system for more advanced technologies. Adding dummy cells around the canary cell improves the tracking of systematic variations. A new canary circuit avoids the possibility that a canary cell may never fail because it resets into its more stable data pattern. A built-in self-test (BIST) block incorporates self-calibration of SRAM minimum standby V_{DD} and the initial failure threshold due to intrinsic mismatch. Measurements from a new 45 nm test chip further demonstrate the function of the canary cells in smaller technology and show that adding dummy cells reduces the variation of the canary cell.

Keywords: *Built-in self test (BIST), Data retention voltage (DRV), Standby power, Static random access memory (SRAM), Variation*

1. Introduction

Built-In Self-test (BIST) is the ability of an integrated circuit (IC) to examine its own functional health, in order to detect and report faults that may jeopardize the reliability of the application wherein it is deployed. The test time of a chip depends on the types of tests conducted [1]. These may include parametric tests (leakage, contact, voltage levels, etc.) applied at a slow speed, and vector tests (also called "functional tests" in the ATE environment) applied at high speed. The time of parametric tests is proportional to the number of pins since these tests must be applied to all active pins of the chip. The vector test time depends on the number of vectors and the clock rate. The total test time for digital chips ranges between 3 to 8 seconds. The vectors may not cover all possible functions and data patterns but must have a high coverage of modeled faults. The main driver is cost, since every device must be tested. Test time (and therefore cost) must be absolutely minimized. During Test mode, power consumption will double than normal mode [2].This Priority Test Pattern method saves significant test time and power consumption by shortening the pattern sequence.

Block Diagram of BIST is shown in figure. BI is enable pin for BIST operation and BO is output of BIST operation, based on BO only can say whether given CUT is working properly are not. When BI = 0, Test Pattern Generator and Ideal Response Block are in OFF state. MUX will accept Input and applied to CUT and outputs are taken at Output pin. When BI = 1, Test Pattern Generator and Ideal Response Block are in ON state. MUX will accept Test Pattern Generator output and applied to CUT and outputs are taken at BO pin. When BO is 1, indicates IC is

working properly otherwise malfunction is there in IC.

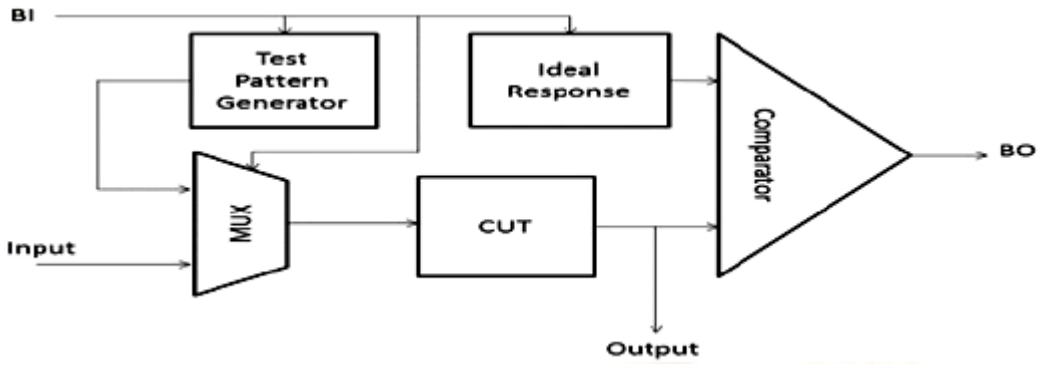
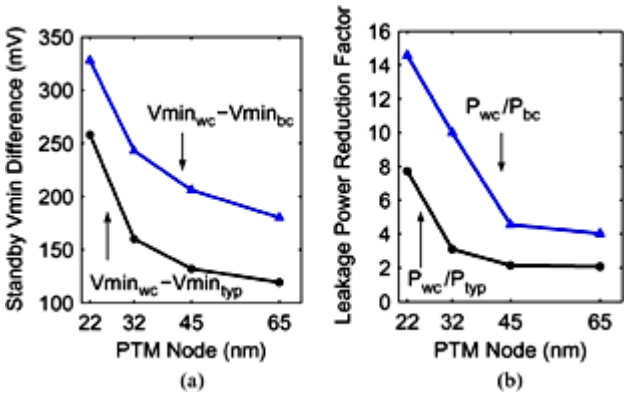


Fig 1.Block Diagram of BIST

Since SRAM/Cache continues to be the largest and most dense component in many digital systems or system-on-chips (SoCs), its leakage power dominates the overall leakage power of the system. One of the most effective leakage reduction techniques is supply voltage scaling. All the leakage current components, including sub-threshold leakage, gate leakage, and junction leakage current, decrease dramatically with a smaller. Leakage power decreases even more rapidly only reduces cell stability itself but also heightens the sensitivity of cell stability to mismatch. The data retention voltage (DRV) is the minimum for the cell to preserve its data [3]. Local variation spreads the DRV of the cells across the chip. To preserve all the data in an SRAM, must be above the DRV of the worst cell within the SRAM array, which we call standby Vmin in this paper. Standby Vmin varies with process variations, voltage fluctuations, and temperature changes (PVT variations). Thus we must address this Vmin variability when choosing standby.

The most straightforward solution is the worst-case based open-loop approach, in which the standby voltage is picked based on the DRV for the worst scenario at design time and maintains unchanged for all the scenarios. Although it is robust, substantial power and energy are wasted because of two reasons. First, the worst PVT scenario only occurs in extreme conditions like extremely high temperature, which is rare for most applications. Second, the margin for the worst PVT protection can be quite large, and it even becomes larger as CMOS technology continuously scales.



Margin between the standby Vmin under the worst-case PVT variation $_{wc}$ and that under the

best/typical case $_{bct}$ and(b) leakage power reduction by using the true Vmin at the best/typical case in-stead of $_{wc}$. A 1-kb SRAM is simulated across the PTM bulk technologies from 65 to 22 nm. the standby Vmin margin between the worst-case PVT variation and the best-case/typical variation increases as technology scales for a 1-kb SRAM array using predictive technology models (PTMs) [5] from 65 to 22 nm. Hat up to 42 leakage power reduction can be achieved if the margin is removed for the 65 nm node. For the 22 nm node, the best-case leakage power reduction increases to 142 and savings for typical silicon increase to 82. Thus using the optimum Vmin instead of the worst Vmin becomes more appealing in smaller technologies. We have proposed an adaptive approach that can tune closer to the optimum Vmin point for each global PVT condition during standby operation. It scales in a closed-loop fashion based on the feedback from canary replicas, which can track the impact of PVT variations on SRAM DRV [6], [7].

2. The General Framework

To measure the DRV of an SRAM, the voltage at which the SRAM data is compromised must be determined. Thus, the DRV computing circuit must check the integrity of the SRAM data at the different voltages. The last voltage which retains the SRAM data is saved as the DRV. To accomplish this, the DRV computing circuit consists of a DC-DC converter to change the supply voltage, a BIST unit that could test data retention and the TCU responsible for controlling the circuit. The DRV is determined and saved once when the chip is in test mode. However, the DRV of the array might change after the chip heats up or due to transistor aging. To account for these cases, the DRV test can be done after the chip was left to operate and heat-up and once the DRV is determined, it is raised one step up. After exiting the test mode, the DRV is retrieved from the TCU and supplied to the SRAM array in standby mode. below shows the general setup of the DRV computing circuit during test mode.

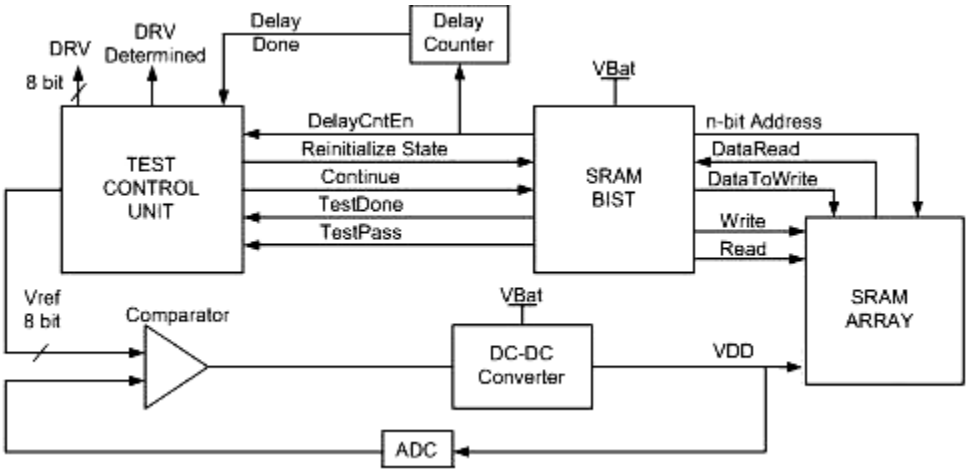


Fig. 3. DRV computing circuit in test mode

A low power DC-DC converter is used to dynamically step down the supply voltage. Along with the comparator and analog-to-digital converter, the DC-DC converter ensures that the supply voltage fed to the SRAM array remains stable and equal to the reference voltage supplied by the TCU. The DC-DC converter must be able to efficiently supply voltages within the range of variations of the DRV. Under process variations, Monte-Carlo simulations of 100k samples showed that the DRV voltage ranges between 20mV and 150mV. To ensure that such voltages are attainable, the pulse frequency modulation (PFM) buck converter proposed in [13] is used. When the chip is in test mode, the DC-DC converter tracks the voltage reference (Vref) provided by the TCU. When the chip exits the test mode, Vref is chosen according to the SRAM operating mode (standby or normal operation). During normal SRAM operation (read/write operations),

Vref is chosen to be the nominal supply voltage. Whereas, during retention mode, Vref is chosen to be the DRV voltage saved within the TCU.

The SRAM BIST unit implements a testing algorithm that determines if the array fails at a specific retention voltage. The BIST unit will be discussed in more details in section III. The SRAM TCU is responsible for controlling the DRV computing circuit: it provides the converter with the reference voltage provides the SRAM BIST unit with the appropriate control signals and saves the DRV once it is determined. Section IV describes the TCU in more details.

3. Canary Scheme Review

The example architecture of our canary scheme [6]. An on-chip or off-chip voltage regulator supplies to the SRAM array and to the canary banks. Several banks of canary cells are de-signed to fail across a range of voltages above the DRV of the SRAM cells as illustrated, and their failures are monitored by the online failure detectors. A programmable failure threshold determines

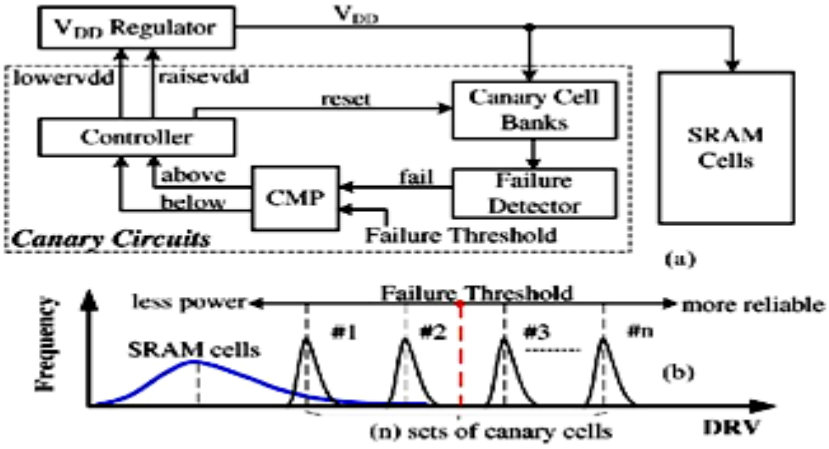


Fig. 4 (a) Example of architecture and (b) principle of canary-based closed-loop $_{DD}$ scaling approach

The proximity of the applied standby to the tail of the SRAM DRVs, and it enables the tradeoff between power saving and data re-liability. When entering the standby mode, the controller starts lowering until the canary failures meet the failure threshold. Once the global stimuli occur, the canary failures may exceed or drop below the failure threshold, which triggers the controller to raise or lower accordingly. The canary system was first successfully implemented on a 90 nm bulk test chip, and the measurement results from that chip showed that it offered 2 power reductions over the worst-case approach for the typical operating condition [7].

4. Canary Cell Improvement

4.1 New Canary Cell Structure

The most critical component in our system is the canary cell. It must duplicate the impact of global stimuli on SRAM cell stability. In addition, it must fail ahead of all the SRAM cells to prevent the loss of data in SRAM. Hence, we have proposed to add a pMOS header on a standalone 6T cell as the canary cell [6]. By tuning the gate voltage of the header (VCTRL), the canary DRV can be altered in a wide range. To improve the correlation of global effects on canary cells and SRAM cells, here we further propose to add dummy 6T cells around the functional 6T cell in the canary cell to mimic the real physical environment of an SRAM cell (). To reduce area cost, we use a 3 23 SRAM mini-array for each canary. A failure detector

monitors the active cell in the center. To ensure the canary cell behaves more like SRAM cells in the presence of systematic variations, we use the same layout pattern as the SRAM array except for minor changes on metal wires for pulling out the storage nodes of the central cell. Both SRAM cells and canary cells use logic rules in our test chip. The actual power supply of the mini array is connected with the pMOS header. As before, when we tune VCTRL to a higher value, the pMOS header is partially turned on, which causes the canary cell to operate at a lower effective than that seen by the core cells?

4.2 New Circuit for Canary Cell Reset

1) ***Issue****:* Since one cell can either hold a "0" or "1", we previously built each canary set with two separate cells for storing "0" and "1". The canary set fails when either the canary cell "0" or the canary cell "1" fails. Although this method is simple and easy to implement, it has one drawback. Mismatch causes a cell to be more stable at one data value than the other, and it is uncertain which data value is more stable due to randomness of local variation (e.g., from dopant fluctuation). For one canary set, if both the canary cell "0" and the canary cell "1"

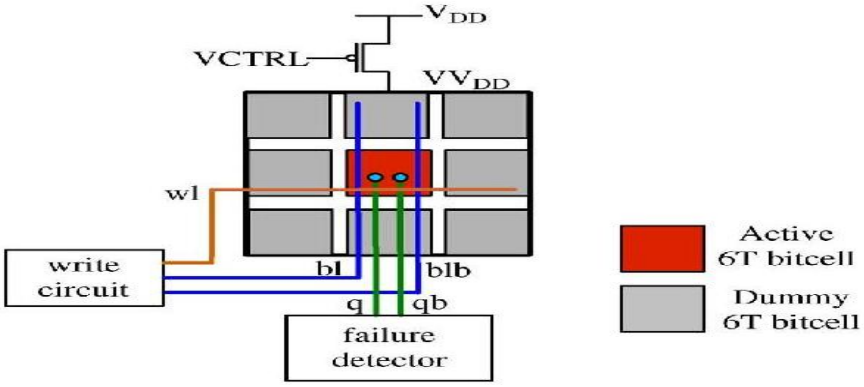


Fig. 5 New canary cell structure with dummy cells. Only modification to active 6T layout is connecting to the internal nodes q and qb

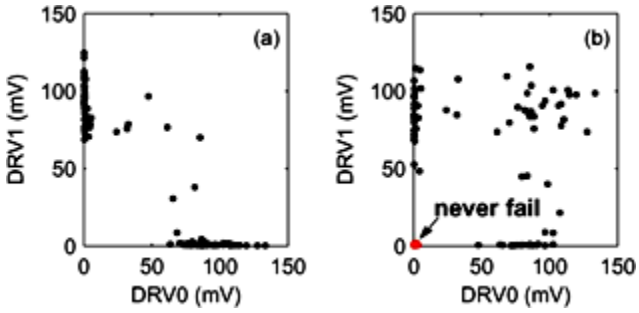


Fig. 6. Correlation between DRV0 and DRV1 (a) when they come from the same cell and (b) when they come from separate cells. 100 samples are plotted

Happen to be more stable at the value that they are holding, this canary set will never fail or fail at a very low supply voltage regardless of the VCTRL value. This can be better explained with the help of DRV. We denote DRV0 and DRV1 as the DRV for holding "0" and "1", respectively. The correlation between DRV0 and DRV1 when both come from the same cell with 100-point Monte Carlo simulations. Most of the samples have one DRV value near or equal to 0 and the other much greater than 0 because device mismatch causes the cell to be unbalanced.

A few samples have two values close to each other because they are more balanced. However, DRV0 and DRV1 never simultaneously equal to 0. Now if the DRV1 comes from a separate cell, the correlation map between DRV0 and DRV1 changes Roughly 20% of samples have both DRV0 and DRV1 near or equal to 0, which means both cells can hold their respective specific data at any voltage. We observed this issue in our first test chip. Although we can use more redundant canary sets to mitigate this issue, it degrades the accuracy of the tracking performance as well as the area efficiency.

2) Solution: To eliminate this issue, we propose a new circuit shown in that automatically stores the least stable data value in each canary cell. Besides the mini-array (simplified as a 6T cell for illustration here), the circuit includes a latching voltage-mode sense amplifier (SA), a D-latch, and two MUXs. the timing waveforms. There are three phases: the restoring, latching, and writing phase. In the restoring phase, "VCTRL" first rises to a high value. This turns off the pMOS header and leaves the actual power of the 6T cell floating. We boost "VCTRL" to so that the cell leakage drops below the DRV to reset the cell. After VCTRL returns back to 0, the storage nodes (q, qb) restore the cell's more-stable state [e.g., (0, 1) Then the latching phase starts with the rise of "saen", which enables the SA so the stable state can be passed to the SA

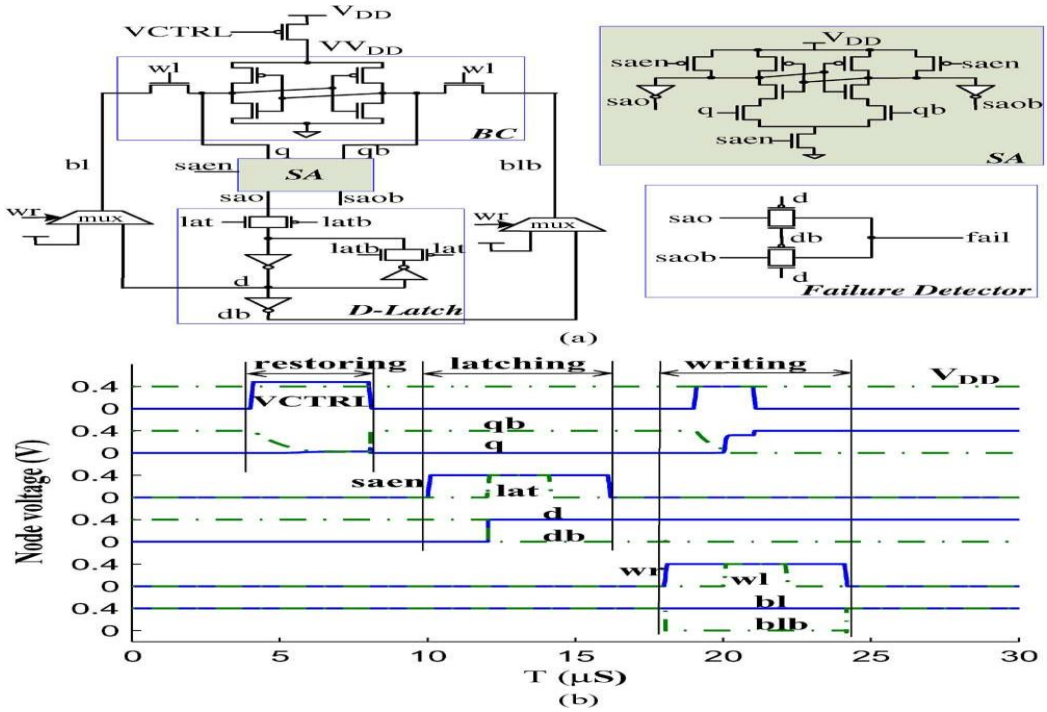


Fig. 7. (a) Circuits and (b) waveforms for canary cell self-loading its less-stable state.

outputs (sao, saob). After some delay time, a pulse on the "lat" signal allows the D-latch to capture the inverted "sao" value into its output "d". So (d, db) are driven to the values of the cell's less-stable state (1, 0). "saen" falls back to 0 at the end of the latching phase to disable SA. In the last writing phase, "wr" rises first so that the MUX can select the value of (d, db) for the bit lines. Then a pulse of "wl" writes the less-stable state (1, 0) into the canary cell, which ensures the canary cell will flip to its more-stable state at its increased DRV. To enhance writ ability, we also design the option to raise "VCTRL" to and float the supply of the cell during write.

5. Built In Self Test (BIST)

Simulation and measurement from a 90 nm test chip have demonstrated that the canary cells can successfully track global variation. However, the canary cells cannot directly track local variation (i.e., mismatch) without a large population of instances. Thus we have to deal with local variation separately. We have previously proposed a fast and accurate model to estimate SRAM DRV tail under local random variation [7]. In this paper, we propose an alternative method to modeling. We incorporate a BIST to detect the initial SRAM Vmin due to intrinsic local mismatch after manufacture at one global condition. We use this value to set the initial failure threshold for the canary cells, which then can track global PVT changes during operation.

5.1 Measuring the SRAM DRV Tail

Based on the direction of searching, there are three methods to measure the standby SRAM Vmin using the BIST: the downward, upward, and binary searching. Among them, the binary searching is the fastest one, but its circuit implementation is most complicated. To reduce circuit complexity, we choose either the downward or upward searching. From our simulation and measurement results, standby Vmin is typically below half of the nominal for a moderate-scale SRAM (e.g., 256 kb) under normal condition. Thus the upward searching requires less iteration. In addition, the upward searching stops checking the remaining cells and increases by one step once the number of failures exceeds the tolerable error limit. In contrast, the downward searching must check all the SRAM cells to ensure that the total number of errors is within the tolerable limit before decreasing by one step. Therefore, we choose the upward searching method to save more test time. Simulation results for a 256-kb SRAM show that upward searching is about 3 times faster than downward searching when standby Vmin is 0.5 V. For each iteration of upward searching, the BIST first checks failures for holding "0" and then for holding "1". This process is repeated after increasing by one step until checking both "0" and "1" complete successfully. Row/column redundancy and ECC are conventionally used for reducing the yield loss due to manufacturing defects and soft errors. For low standby power operation, they can also be used to tolerate data-retention errors so that the minimum standby voltage can be less than the worst DRV in the SRAM [8]. The detailed flow for checking hold failures is illustrated in. First, in active mode the nominal value), "0"/"1" is written into each address. Then the SRAM enters standby mode the standby value) and maintains standby for a period of time. After the standby operation, data is read out and checked in active mode. If the number of failed bits is larger than the number of correctable bits with ECC and all the redundant rows have been used, the checking process is terminated with, which means the current standby voltage is too low to retain data and hence must be increased. Note that the standby time should be sufficiently long to ensure the occurrence of the worst static scenario. An example of one SRAM cell in the 45 nm technology we use. The DRV value decreases for less standby time, which means SRAM cells can tolerate more dynamic noise when the duration of the noise is shorter. This similar behavior of larger dynamic noise tolerance has been observed in logic gates [9]. After the standby time exceeds a threshold point, its DRV reaches the largest value, which equals the one from the static dc simulation.

5.2 Calibrating Initial Failure Threshold

Simulation and measured results have shown that the DRV of the canary cell is approximately linear with the VCTRL value [6]. By analyzing the leakage current through the header when reaches the true DRV of the cell, we derived that the linearity can be approximated with, where

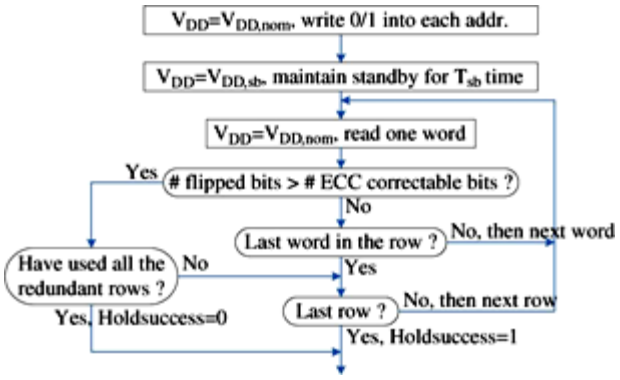


Fig. 8. Flow for hold failure check with BIST

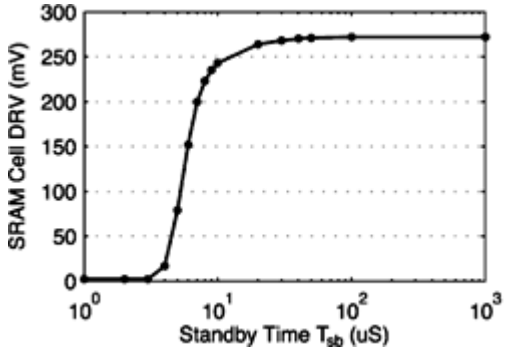


Fig. 9. DRV of an SRAM cell changes with the standby time

is the DIBL coefficient of the header [7]. Hence we can generate a series of VCTRL values (e.g., with a resistor ladder) to create a group of canary categories that fail at regular intervals across a wide range. During self-calibration,

6. 45 nm Test Chip Implementation and Measurement

Our first prototype has been implemented and measured in a bulk 90 nm test chip [6], [7]. To verify the effectiveness of our scheme in scaled technologies, we implemented the canary circuits in a bulk 45 nm test chip. Fig 10 shows its die photo. On each die, there are two canary

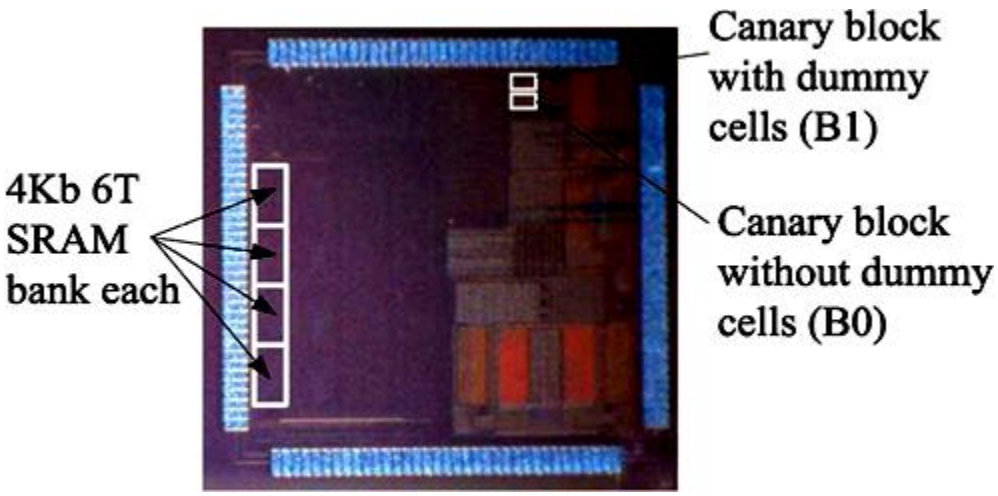
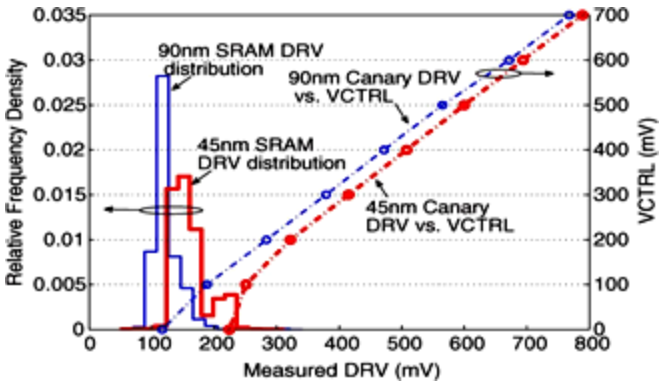


Fig. 10.: 45 nm test chip die photo



Measured canary DRV against VCTRL and measured frequency density of the SRAM DRV from both the new 45 nm chip and the previous 90 nm chip. 16 and 8 kb SRAM cells are measured for 45 and 90 nm, respectively blocks. Each canary block contains all the canary circuits plus test circuits. The canary bank consists of eight canary sets and each canary set employs three-way redundancy. All the canary cells in the first canary block use the standalone cell structure, and those in the second canary block use the improved structure with the dummy cells as shown. We also implemented four 4 kb banks of SRAM on the die.

The measured canary cell DRV against VCTRL and the frequency density of the measured SRAM DRV from the new 45 nm test chip are plotted . For comparison, we also plot the measured DRV results from the previous 90 nm test chip. Both the 90 and 45 nm canary DRV measurements maintain excellent first-order linearity with VCTRL values above 100 mV. The nonlinearity for VCTRLs below 100 mV is due to the rolling off term in the sub-threshold current equation [7]. Note that for the same amount of VCTRL increment (e.g., 100 mV), the 45 nm canary DRV has smaller increase than the 90 nm counterpart because the sensitivity of the canary DRV to VCTRL is inversely proportional to the header's DIBL coefficient, which increases with technology scaling. Fig. shows that the 45 nm SRAM DRV spreads wider than the 90 nm counterpart due to device variability increasing with technology scaling. Although the variance of the SRAM DRV distribution grows in 45 nm, demonstrates that tuning VCTRL can still provide a sufficiently large range of canary DRV above the tail of the SRAM DRV in 45 nm just as in 90 nm. This ensures that the canary scheme maintains functional in 45 nm. We further compare the results from the two different canary blocks to examine the effect of dummy cells. Fig. shows the measured results from 85 dies on one wafer. For each die, the VCTRL value of each canary set is generated by an on-die resistor ladder. The canary set with the higher index number connects to a higher VCTRL value. The variation of the canary DRV is computed as the ratio of the sigma to

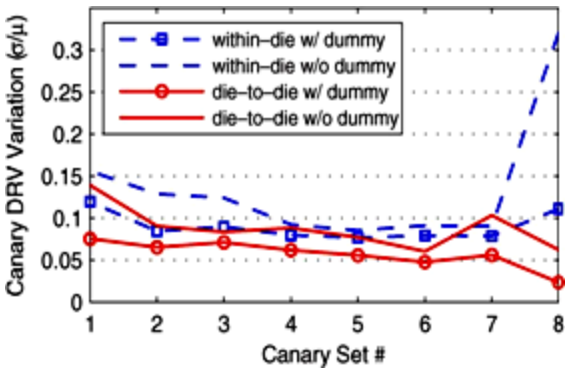


Fig. 11. With dummy cells, both within-die and die-2-die variations of the ca-nary DRV are reduced

the mean. A smaller ratio value means less variation occurred on the canary. We first compare the within-die variation, i.e., the variation of the three redundant copies of each canary set on each die. The average result from 85 dies is plotted with dashed curves. The block with the dummy cells has less within-die variation, especially for the ca-nary set #8 that is configured to have the largest DRV. We also plot the die-to-die variation with the solid curves. In this case, the canary DRV value of each die is obtained through the majority-3 voting among the redundancies on the same die. The block with dummy cells also has less die-to-die variations. Therefore, the use of dummy cells inside the canary cell can effectively reduce both with-in-die and die-to-die variations of the canary cell.

7. Conclusion

SRAM standby Vmin, i.e., the DRV of the worst SRAM cell, shifts with global PVT variations. The traditional worst-case open-loop approach prevents the potential power savings for non-worst-case dies and scenarios. We have proposed a feedback scheme using canary replicas for aggressive standby scaling while maintaining sufficient data reliability. In this paper, we propose several enhancements to this scheme. Dummy cells is added in the canary cell to improve the correlation between the canary cell and SRAM cells under systematic variation. A new resetting circuit ensures that the canary cell holds the less-stable state so that it can flip at a higher voltage. We also propose a BIST to self-calibrate the SRAM standby Vmin and the initial failure threshold due to intrinsic mismatch after manufacture. Measurement results from a 45 nm test chip demonstrate that the canary cells can fail at regular intervals across a wide range above the SRAM DRV tail in smaller technology. In addition, measurements confirm that using dummy cells can reduce the variation of the canary cell and thus improve the accuracy of the tracking behavior.

References

1. Cao, Y. T. Sato, M. Orshansky, D. Sylvester, and C. Hu, (2000).“New par-adigm of predictive mosfet and interconnect modeling for early cir-cuit simulation,” in Proc. IEEE Custom Integr. Circuits Conf. (CICC). pp. 201–204.
2. Flautner, K. N. S., Kim, S., Martin, D. Blaauw, and T. Mudge, (2002) “Drowsy caches: Simple techniques for reducing leakage power,” in Proc. Int. Symp. Comput. Arch., May 25–29, pp. 148–157.
3. Kim, N. S. K., Flautner, D. Blaauw, and T. Mudge, (2004). “Single-Vdd and single-Vt super-drowsy techniques for low-leakage high-performance instruction caches,” in Proc. Int. Symp. Low Power Electron. Des. (ISLPED), 2004, pp. 54–57.
4. Qin, H. A., Kumar, K.. Ramchandran, J.. Rabaey, and P. Ishwar, “Error-tolerant SRAM design for ultra-low power standby.
5. Qin, H. Y., Cao, D. Markovic, A. Vladimirescu, and J. Rabaey (2004). “SRAM leakage suppression by minimizing standby supply voltage,” in Proc. Int. Symp. Quality Electron. Des. (ISQED) pp. 55–60.
6. Wang, J. and B. Calhoun, (2007). “Canary replica feedback for near-DRV standby $_{DD}$ scaling in a 90 nm SRAM,” in Proc. CICC, 2007, pp. 29–32.
7. Wang, J. and B. H. Calhoun, (2008).“Techniques to extend canary-based standby $_{DD}$ scaling for SRAMs to 45 nm and beyond,” IEEE J. Solid-State Circuits, vol. 43, no. 11, pp. 2514–2523, Nov.
8. Wang, Y. H.. Ahn, U. Bhattacharya, T.. Coan, F.. Hamzaoglu, W.. Hafez, C.-H.. Jan, R. Kolar, S. Kulkarni, J. Lin, Y. Ng, I. Post, L. Wel, Y. Zhang, K. Zhang, and M. Bohr, (2007). “A 1.1 GHz 12 -leakage SRAM design in 65 nm ultra-low-power CMOS with integrated leakage reduction for mobile applications,” in Proc. ISSCC, Feb. pp. 324–606.