Canary Replica Feedback for Near-DRV Standby V_{DD} Scaling in a 90nm SRAM

Jiajing Wang and Benton H. Calhoun

University of Virginia, 351 McCormick Rd., Charlottesville, VA, 22904

***Abstract*- Canary bitcells act as online monitors in a feedback architecture to sense the proximity to the Data Retention Voltage (DRV) for core SRAM bitcells during standby voltage scaling. This approach implements aggressive standby V_{DD} scaling by tracking PVT variations and gives the flexibility to tradeoff between the safety of data and decreased leakage power. A 90nm 128Kb SRAM test chip confirms that the canary cells track changes in temperature and V_{DD} and that they provide a reliable mechanism for protecting core cells in a closed loop V_{DD} scaling system. Power savings improve by up to 30× compared with the conventional guard-banding approach.**

I. Introduction

SRAM leakage power dominates the overall leakage power of many digital systems. V_{DD} scaling is an effective technique for standby power reduction [1], however, there is a minimum V_{DD}, called the Data Retention Voltage (DRV) [2], below which a bitcell has negative Static Noise Margin (SNM) and will lose its state. The cell DRV can thus be defined as the voltage at which a cell has SNM equal to zero. Global and local variations cause a distribution of the DRV for bitcells across the chip, and the bitcell with the highest DRV actually determines the minimum V_{DD} that can be applied to the whole SRAM. Given the randomness of the physical parameters, Monte-Carlo (M-C) simulation can be used to get the statistical characteristics of the DRV. Fig. 1 shows the distribution of DRV for 90nm and 45nm nodes from a 5k-point M-C simulation of within-die threshold voltage (V_T) variation. The tail of the distribution grows as technology scales due to increased process variation.

Existing V_{DD} scaling approaches add a safety margin to the worst scenario to prevent the loss of data. In other words, the standby supply voltage ($V_{Standby}$) is selected based on worst case PVT variations and local mismatch plus an extra guard-band, which is added for more robustness. Many previous works select a $V_{Standby}$ at design time that maintains sufficient margin to protect data in the cells (e.g., the drowsy cache in [3] and the microprocessor with a drowsy mode in [4]). This open-loop approach can leave substantial power savings on the table because the full range of potential DRVs can be quite large when accounting for the worst-case. For example, in our 90nm technology, assuming ±50mV V_{DD} fluctuation, a 0ºC-100ºC temperature range, 3σ local mismatch, and 50mV of noise margin guard-band, $V_{Standby}$ is about 400mV higher than the real DRV (i.e., the V_{DD} point when SNM=0) for the best case, as Fig. 2a illustrates. Setting $V_{Standby}$ at design time to accommodate this worst case reduces the achievable leakage savings by up to 30× as shown in Fig. 2b. With the scaling of technology, we can expect to sacrifice more leakage power savings by using this conservative worst-case approach due to increased device variability. Closed loop control of $V_{Standby}$ offers an appealing alternative that can take advantage of the extra savings.

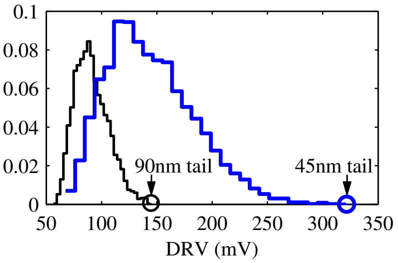


Fig. 1 DRV distribution from a 5k-point Monte-Carlo simulation of within-die variation for 90nm and 45nm nodes. The tail sets the array-wide $V_{Standby}$.

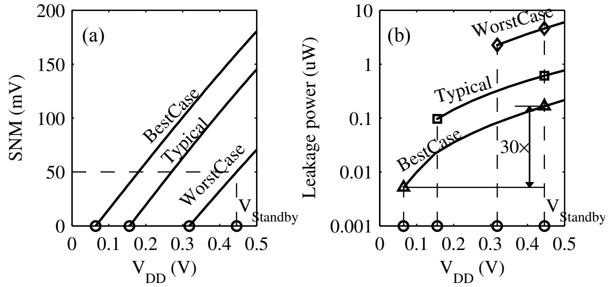


Fig. 2 Simulated worst bitcell SNM (a) and 1kb SRAM leakage power (b) vs. V_{DD} under PVT variations (best-case, typical and worst-case) and 3σ local mismatch.

Canary replica flip-flops were proposed to monitor the proximity to failure for V_{DD} scaling in flip-flops [5]. In this paper, we propose a feedback architecture using a new canary replica structure for SRAM bitcells. Our approach allows aggressive leakage power reduction (up to 30× improvement over the conservative approach) for ultra-low-power applications by tracking the impact of global variation and environmental changes (e.g. temperature, V_{DD} instability) on the DRV, and provides a unique method to tradeoff reliability of stored data with leakage power reduction. We describe the concept of this new scheme in Section II, and provide details of the new circuits to implement this scheme in Section III. Section IV further discusses the main advantages of our scheme. Section V demonstrates the measured results from a 90nm test chip. The conclusions are drawn in Section VI.

II. Canary Feedback Scheme

Fig. 3 shows the feedback loop used to lower V_{DD} for leakage power savings while protecting data by keeping V_{DD} above the DRV for the core cells. A voltage regulator supplies V_{DD} to the core cells and to the canary replicas. This regulator may be on-chip or off-chip, and recent results have

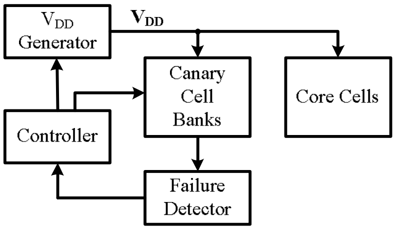


Fig. 3 Architecture for standby V_{DD} scaling using canary cell monitors.

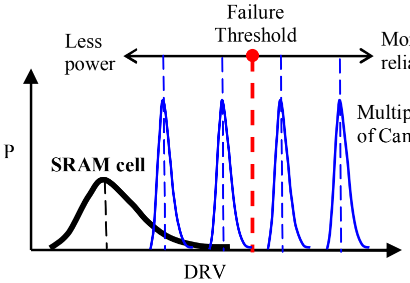


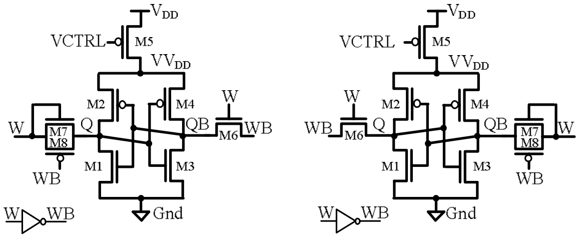
Fig. 4 Different sets of canary cells fail at regular intervals above the average DRV of core cells, allowing a trade-off of reliability with power savings.

demonstrated that dc-dc converters can supply a large range of low voltages with high efficiency [6].

To protect all of the data in the core cells, the loop should maintain V_{DD} above the worst-case DRV of the core array. A combination of global effects (e.g. global variation, V_{DD} variation, temperature) and local variation will set this worst-case DRV. Local variation sets the spread of the DRV distribution, and global effects predominately move its mean value. Since the relatively small number of canary cells cannot replicate the statistics of the large core array, the canaries cannot effectively track local variation. But the canaries will track global effects such as those listed above, so they can effectively remove the need to guard for these conditions.

The canary cell banks are designed for the canaries to fail at a range of higher voltages relative to the average core cell and to maintain this behavior despite changes in global variations and environmental conditions so that V_{DD} can adjust with those changes. This leads to larger power savings allowing lower V_{DD}s for non-worst-case conditions. Fig. 4 illustrates how different sets of canaries are tuned to fail at regular intervals above the average DRV of the core cells. Local variation smears the distribution of canary DRVs in each set. Providing a continuum of canary failures at voltages above the DRV of the average core bitcell allows the designer to set and to alter the tradeoff between storage reliability and power. This architecture allows for a variety of power-saving policies, and we provide a simple one as an example.

Consider a handheld device holding video data during standby. The controller starts lowering the standby V_{DD}. When canary cell failures indicate that V_{DD} reaches a predefined failure threshold, then the controller raises V_{DD} slightly, resets the canary cells, and continues monitoring. Since losing a few bits of this data is acceptable, the reset point may be quite near (or below) the predicted array-wide DRV. When the application changes and data are more important, the failure threshold can be reset to a higher value. This makes the controller raise V_{DD} until meeting the new failure threshold to provide a larger margin of protection above the predicted array-wide DRV. Clearly, accurate prediction of the DRV spread due to local variation improves our ability to set this failure threshold. Section IV-A describes the methods we have proposed to predict the array-wide DRV under mismatch.



(a) Canary cell ‘1’ (b) Canary cell ‘0’

Fig. 5 Canary cell schematic

III. Canary Circuits Design

Since the DRV usually occurs in the sub-threshold region, new low voltage circuits are necessary for implementing the canary cells, failure detection, and controller.

A. Canary Cells

The canary cells must exhibit the same DRV dependencies on global conditions as core cells and should allow DRV tuning to much higher than the core cell DRV. An un-tuned canary should ideally have a DRV at or above the average DRV of the core cells. These features provide the functionality shown in Fig. 4. Fig. 5 shows the schematic of the proposed canary cell, which includes several features to increase its DRV relative to the average core cell.

First, the reset circuit that writes the cell applies the worst-case vector to the bitlines during hold. For example, the canary cell ‘1’ (Fig. 5a) has nodes Q and WB at ‘1’ while QB and W remain ‘0’ during standby. This creates the worst-case leakage through the access transistors to encourage the cell to flip, increasing the mean of DRV by 10.65%. Second, we can use data dependencies caused by asymmetric local variations. We define DRV0 as the DRV for data ‘0’, and DRV1 as the DRV for data ‘1’. For a matched cell, DRV0 equals DRV1. However, any asymmetry in variations causes one lobe of the cell’s butterfly curve to be smaller than the other, making its DRV higher for a specific data value. In other words, DRV0 and DRV1 are negatively correlated (one increases while the other decreases). Therefore, if at test time, we identify the worst-case data for each canary cell and use it henceforth, then the probability that the canary cell fails at a supply voltage x, $P_{canaryfail}=1-F_{X0,X1}(x, x)$, where $F_{X0,X1}$ is the joint CDF of DRV0 and DRV1 and is smaller than $F_{X0}(x)\cdot F_{X1}(x)$ because of their negative correlation. This is higher than the failure probability of a single cell with random data, which is equal to $1-F_{X0}(x)$ when assuming DRV0 and DRV1 are identically distributed. If we wish to avoid reconfiguration at test time, we can allocate two separate canary cells, one for holding ‘1’ (Fig. 5 a) and the other for ‘0’ (Fig. 5b), and then use the OR operation to set the failure status when either of them fails. In this case, $P_{canaryfail}=1-F_{X0}(x)\cdot F_{X1}(x)$, which is also higher than the probability of a single cell failure but lower than the probability of the programmed cell failure. For simplicity, we

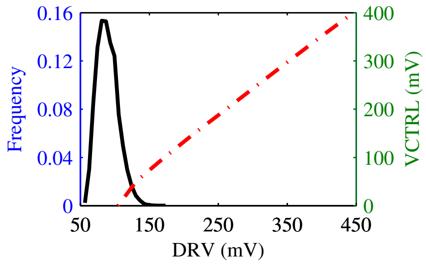


Fig. 6 Simulated Core-cell DRV distribution and average canary-cell DRV varying with VCTRL.

use separate canary cells for ‘0’ and ‘1’ in our test chip.

To provide for tuning the DRV of canary sets to be higher than the average core DRV, a PMOS header connects each canary to V_{DD}. The gate voltage VCTRL of the header sets the supply voltage VV_{DD} of the canary cell to a value lower than V_{DD}. This powerful knob essentially moves the mean of the DRV distribution for each canary cell and thus moves it across a wide range (as in Fig. 4). Fig. 6 shows the simulated DRV of the average canary cell vs. VCTRL relative to the core cell DRV distribution. This figure shows that the proposed cells allow the desired continuum of failure voltages for canaries. It also implies the approximate linear relationship between canary DRV and VCTRL, so canary DRVs can be placed at identical intervals with evenly spaced VCTRLs. To reduce the spread of canary cell DRV distribution relative to the core, larger sizes are used for canary transistors.

B. *Canary Bank Arrangement*

For the test chip, we arranged the canary cells into a bank structure (Fig. 7). The canary bank contains multiple sets (rows) of canary cells (e.g. 1-cell/row), and each set shares a distinct VCTRL. A programmable failure threshold allows a range of policies for trading off power and reliability. We employ 3-way redundancy of the banks with majority-3 gates to screen out abnormalities caused by rogue cells with large variation. The VCTRL values are set off-chip or by an on-chip resistor ladder that generates evenly spaced VCTRL values between the voltage rails.

C. *Failure Detector*

The amplifier circuit in Fig. 8 detects the failure of a canary cell. ‘Qin’ and ‘QBin’ are directly connected with the canary cell nodes ‘Q’ and ‘QB’. When Q and QB flip, the output signal ‘Fail’ rises. This amplifier works in the sub-threshold region and functions well even when the input signals are extremely small. Other circuits in the control block (not shown) operate in sub-threshold for robust low V_{DD} operation and implementation of the desired power saving policy.

IV. Near-DRV Standby V_{DD} Scaling

A. *Obtaining Worst-Case Array-Wide DRV*

As described before, our canary-based feedback structure removes the impact of global stimuli on DRV, but the specific value of the failure threshold for the loop depends on the tail of the DRV distribution caused by local mismatch. For large memories, this tail extends beyond 5σ or 6σ. Here we propose three methods to set the failure threshold based on this tail.

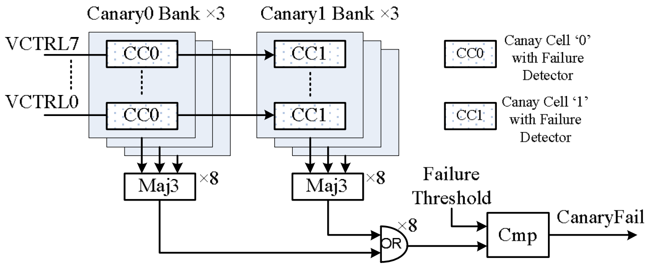


Fig. 7 Canary bank structure.

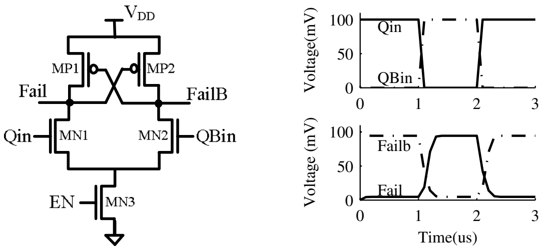


Fig. 8 Sub-threshold canary failure detection circuit schematic and simulated waveforms when V_{DD}=100mV.

First, we can characterize the DRV distribution for each die at test time under normal environmental conditions. Since the DRV of each cell is fairly independent of data in adjoining cells, we can measure the DRV for all-ones (e.g. write all 1’s, lower V_{DD} and settle, raise V_{DD} and read; repeat with slightly lower V_{DD}) and then for all-zeros test patterns. If the first failure occurs at the current iteration, the worst-case DRV is the previous holding V_{DD}.

If test time characterization is undesirable, then we can model the tail of the array-wide DRV. Monte-Carlo simulation is costly for large memories with the tail out to 5~6σ. We propose two fast methods for providing an accurate estimation of the DRV at the tails [7]. One method applies the Statistical Blockade tool [8] for DRV tail estimation. The second method uses a statistical model based on the connection between DRV and SNM [7]. Equation (1) provides the model [7]:

$$\max(DRV_{core})=\frac{1}{k}\left(\sqrt{2}\sigma_0\cdot erfc^{-1}\left(2-2\sqrt[m]{\frac{m-1}{m}}\right)-\mu_0\right)+V_0\,, \quad (1)$$

where m is the memory size in bits, k is the slope of SNM High (SNM holding ‘0’) versus V_{DD}, μ_0 and σ_0 are the mean and standard deviation of SNM High at V_{DD}=V_0, and $\mathrm{erfc}^{-1}(\cdot)$ is the inverse complementary error function. k, μ_0 and σ_0 are fitting coefficients; k can be extracted from a DC sweep simulation, and μ_0 and σ_0 can be extracted from a small-scale (1.5k~5k) Monte-Carlo simulation. Both methods provide a speedup of more than 4 orders of magnitude for a 1G-b memory and an average error of less than 2% relative to Monte-Carlo simulation [7]. Any of these three approaches can help to fine-tune the failure threshold for different power savings policies that utilize the canary-based architecture.

B. *Trading off reliability with leakage power saving*

Using the linear relationship of VCTRL to the canary DRV (Fig. 6) and the inverse CDF model for array-wide DRV from

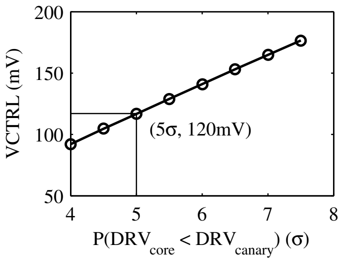


Fig. 9 Required VCTRL voltage for the probability that core DRV is less than canary DRV (in sigmas).

[7], we can estimate the relationship between VCTRL and SRAM reliability. Fig. 9 shows the simulated VCTRL value necessary to satisfy a given SRAM reliability constraint, i.e., a probability that $DRV_{core}<DRV_{canary}$, accounting for the tails. For example, if 5σ probability is required, VCTRL is 120mV. This implies that canary cells with VCTRL larger than 120mV have an even higher probability of failing before all of the core cells. By setting VCTRL of each canary set to the value for the desired σ, we represent the reliability of the SRAM using the canary cells stability. Then with the required SRAM reliability constraints, we can set an appropriate canary failure threshold, which finally sets the achievable power savings by the loop. Therefore, the known amount of stability margin can be traded off to achieve more power savings.

C. Tracking environmental changes

For a given failure threshold, the closed-loop structure in Fig. 3 adjusts V_{DD} according to the feedback of the canary cell failures under environment changes or for dies with differing global variation. If temperature gradients are a concern, then canary cells can be dispersed at different locations in a core array. Fig. 10 shows that the simulated canary cells in the upper rows (with higher VCTRL) consistently fail before core cell at all temperatures. If voltage fluctuation occurs, the DRV of core cells and canary cells will increase/decrease by the same amount. Therefore, the canary cells continue to fail before the core cells across environmental changes.

V. Test Chip Implementation and Measurement

A 90nm bulk CMOS test chip implements a 128Kb SRAM with adjacent canary circuits causing 0.6% area overhead (Fig. 11). Fig. 12a shows the measured average failure V_{DD} for canary cells versus VCTRL under different temperatures. The canary cells exhibit the desired dependency on VCTRL and successfully track the temperature changes. Fig. 12b shows the function of one example closed-loop control method. When the reference voltage of the resistor ladder is connected to V_{DD}, lowering V_{DD} encourages more canary cells to fail.

VI. Conclusion

We propose a feedback scheme using canary replicas to implement aggressive V_{DD} scaling for SRAM standby leakage power reduction without losing data. Simulation and measurement of a 90nm test chip confirm that the canary cells reliably fail in a continuum at higher voltages than the average core cell and that this relationship holds across environmental changes. The proposed mechanism allows data stability to be traded-off for increased power savings. The power savings from our scheme is about 30× higher than the conventional guard-band approach.

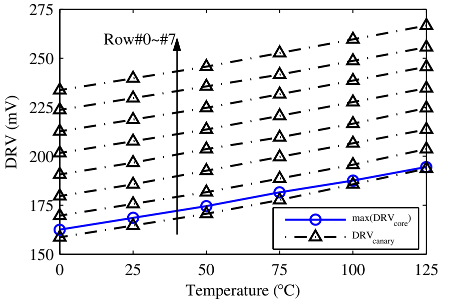


Fig. 10 Simulated DRV of canary cell (DRV_{canary}) at upper 7 rows (larger VCTRL) are always higher than the worst DRV of core cells, $\max(DRV_{core})$, under all the temperature fluctuations.

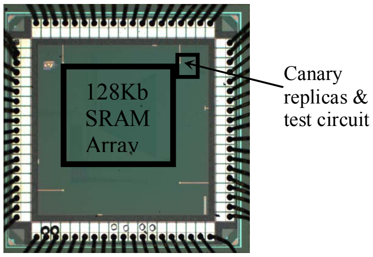


Fig. 11 Chip die photograph.

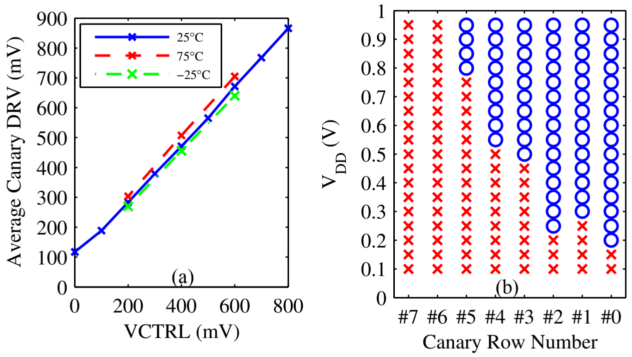


Fig. 12 Measured results: (a) Canary DRV vs. VCTRL under different temperatures and (b) More canary cells failed with V_{DD} scaling ('x'-failed; 'o'-passed).

Acknowledgment

This work was supported by the MARCO/DARPA Focus Research Center for Circuit and System Solutions (C2S2). The authors thank Anantha Chandrakasan for his ideas and suggestions supporting this work.

References

[1] R. Krishnamurthy, et al., *CICC*, pp. 125−128, 2002.

[2] H. Qin, et al., *ISQED*, pp. 55-60, 2004.

[3] N. Kim, K. Flautner, D. Blaauw and T. Mudge, *ISLPED*, pp.54-57, 2004.

[4] L. T. Clark, et al., *IEEE VLSI Systems*, Vol. 12, pp. 947-956, 2004.

[5] B. Calhoun and A. Chandrakasan, *JSSC*, Vol. 39, No. 9, pp. 1504-1511, September 2004.

[6] Y. Ramadass and A. Chandrakasan, *ISSCC,* 2007.

[7] 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.

[8] A. Singhee and R. A. Rutenbar, *DATE*, 2007.