Techniques to Extend Canary-Based Standby V_{DD} Scaling for SRAMs to 45 nm and Beyond

Jiajing Wang, *Student Member, IEEE*, and Benton Highsmith Calhoun, *Member, IEEE*

***Abstract*—V_{DD} scaling is an efficient technique to reduce SRAM leakage power during standby mode. The data retention voltage (DRV) defines the minimum V_{DD} that can be applied to an SRAM cell without losing data. The conventional worst-case guard-banding approach selects a fixed standby supply voltage at design time to accommodate the variability of DRV, which sacrifices potential power savings for non-worst-case scenarios. We have proposed a canary-based feedback to achieve aggressive power savings by tracking PVT variations through canary cell failures. In this paper, we show new measured silicon results that confirm the ability of the canary scheme to track PVT changes. We thoroughly analyze the adaptiveness of the canary cells for tracking changes in the SRAM array, including the ability to track PVT fluctuations. We present circuits for robustly building the control logic that implements the feedback mechanism at subthreshold supply voltages, and we derive a new analytical model to help tune the canary cells in the presence of variations. To realistically quantify the potential savings achievable by the canary scheme, we assess the impact of various sources of overhead. Finally, we investigate the performance of the canary based scheme in nanometer technologies, and we show that it promises to provide substantial standby power savings down to the 22 nm node.**

***Index Terms*—Closed loop, DRV, low-power memory, reliability, SRAM, standby V_{DD} scaling, variation.**

I. Introduction

With technology scaling, the leakage power consumed by transistors grows dramatically and becomes the most important challenge for many applications in both active and standby mode. For battery-constrained devices, the reduction of standby leakage power is especially important for longer battery life. Since SRAM/Cache is the largest component in many digital systems or SOCs, its leakage power during standby mode usually dominates the overall standby leakage power. Therefore, it is important to reduce SRAM standby leakage power. There are several techniques that have been used to reduce SRAM leakage power, such as V_{DD} scaling [1], source biasing (sleep transistor) [2], body biasing [3] and multiple V_T FETs. V_{DD} scaling stands out among these techniques because it can reduce the gate leakage current more efficiently, which has become relatively more significant in deep submicron technologies.

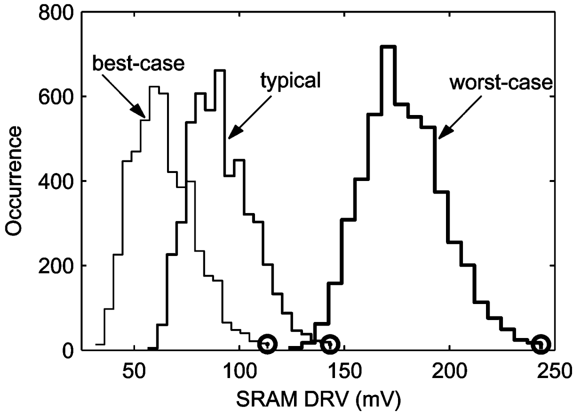


Fig. 1. DRV distribution of a 5 Kb SRAM array with global PVT variations and local V_T variations. Three PVT variations (typical, best-case, and worst-case) are shown.

While V_{DD} scaling offers the benefit of leakage power savings for SRAM, it degrades the cell hold stability simultaneously. There is a minimum V_{DD}, called the data retention voltage (DRV) [4], 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. Given the randomness of the physical parameters, Monte Carlo simulation can be used to get the DRV statistics. Fig. 1 shows the DRV distribution of a 5 Kb SRAM array at three global cases (typical/best-case/worst-case). Each case shows a 5K-point Monte Carlo simulation with within-die threshold voltage (V_T) variation plus certain global variations, including process variation, temperature change and voltage fluctuation (PVT variations). It is obvious that for each global scenario, local variations spread the DRV of the cells across the same array, and the cell with the highest DRV (the tail marked with circle in Fig. 1) actually determines the minimum V_{DD} (Vmin) that can be applied to the whole SRAM array under the current global variations. In contrast, the global variations predominantly move the entire DRV distribution (and the tail) around, so the Vmin for the whole SRAM shifts with global effects.

To account for this variability, existing V_{DD} scaling approaches add a safety margin to the worst scenario to prevent the loss of data. Many previous works select a standby V_{DD} at design time that maintains sufficient margin to protect data in the cells (e.g., the drowsy cache in [5] and the microprocessor with a drowsy mode in [6]). 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

Manuscript received March 19, 2008; revised August 18, 2008. Current version published November 19, 2008. This work was supported by the MARCO/DARPA Focus Research Center for Circuit and System Solutions (C2S2).

The authors are with the Electrical and Computer Engineering Department, University of Virginia, Charlottesville, VA 22904-4743 USA (e-mail: jjwang@virginia.edu; bcalhoun@virginia.edu).

Digital Object Identifier 10.1109/JSSC.2008.2005814

0018-9200/$25.00 © 2008 IEEE

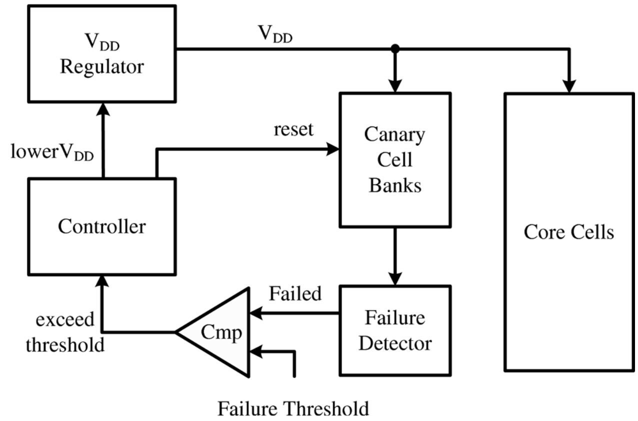


Fig. 2. Architecture of the canary-based feedback loop for SRAM standby V_{DD} scaling [7].

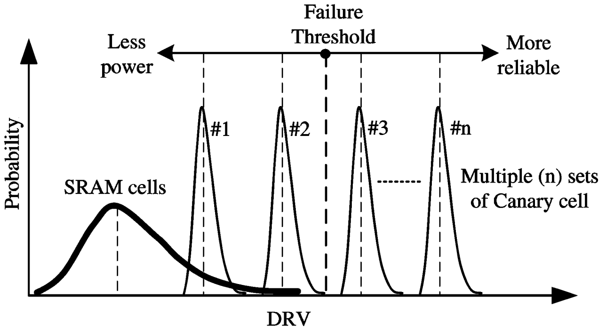


Fig. 3. Mechanism of the canary scheme. Multiple canary sets are designed to fail at regular intervals above the average of SRAM cells and maintain this behavior despite PVT variations. Failure threshold can be set to tradeoff data reliability with power savings [7].

worst case. 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 standby V_{DD} offers an appealing alternative for conditions that allow extra power savings without data loss. We have previously proposed a feedback architecture using a canary replica structure for SRAM bitcells [7]. This 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 on the DRV, and provides a mechanism to tradeoff the reliability of stored data with leakage power savings. In this paper, we will examine the effectiveness of our canary-based scheme in cutting edge technologies at the 45 nm node and beyond. Specifically, we examine practical implementation issues including mechanisms for adapting to environmental changes, sources of overhead and their impact on the canary approach, and the projected effectiveness of applying the scheme in future nanometer SRAMs. We first briefly introduce our previous work in Section II. In Section III, we examine the adaptiveness of the scheme to PVT variations. Then we evaluate overhead sources in Section IV. In Section V, we use simulations with predictive technology models (PTMs) to assess the effectiveness of using closed-loop feedback for scaling V_{DD} to nodes under 65 nm. Finally, we draw our conclusions in Section VI.

II. Canary Feedback Scheme Overview

A. Principle

Fig. 2 shows an example architecture for the canary-based 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 [7]. A voltage regulator supplies V_{DD} to the core cells and to the canary replicas. When entering the standby mode, the controller starts lowering V_{DD}. Several banks of canary cells are designed to fail across a range of voltages above the actual DRV of the tail of the core bitcells. Canary cell failures are monitored by the online failure detector. If a failure is detected, then the controller raises V_{DD} to the last working value, resets the failed canary cells, and continues to monitor.

The big advantage of this feedback scheme will be the improvement of power savings. The online monitoring provided by the canary cells can track any global variations or environmental changes because they are affected by these changes in the same ways that the core cells are affected. Under PVT variations, the failure voltage of the canary replica changes by the same amount as the typical SRAM cell. Using feedback based on the canary failures, V_{DD} can be adjusted to approach the real SRAM failure point for the current scenario. Therefore, we can effectively remove the need to guard for the worst-case scenario and achieve more power savings for non-worst-case scenarios.

In addition, we have proposed a canary cell bank structure with a programmable failure threshold for trading off SRAM data reliability with power savings. Fig. 3 illustrates the basic mechanism that provides this tradeoff. The canary cell bank contains multiple (n) canary sets, which fail at a regular intervals above the average DRV of the core cells and maintain this behavior despite changes in global variations and environmental conditions so that V_{DD} can adjust with those changes. Local variation smears the distribution of canary DRVs in each set, but the canary distributions are not good indicators of the core cell distribution because there are too few canary cells. We will emphasize that the purpose of the canary categories is not to estimate the full distribution of the core SRAM cells, but instead to sense the proximity of the currently applied V_{DD} to the DRV of the average SRAM cell. To assess the proximity of failure to the tail of the distribution, we must model or measure that tail and relate its location to the canary behavior, as we will discuss in Section III. 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. When power saving is the major concern and losing a few bits of this data is acceptable (e.g., when using an ECC method), a failure threshold may be quite near (or below) the real tail of SRAM 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 array-wide DRV.

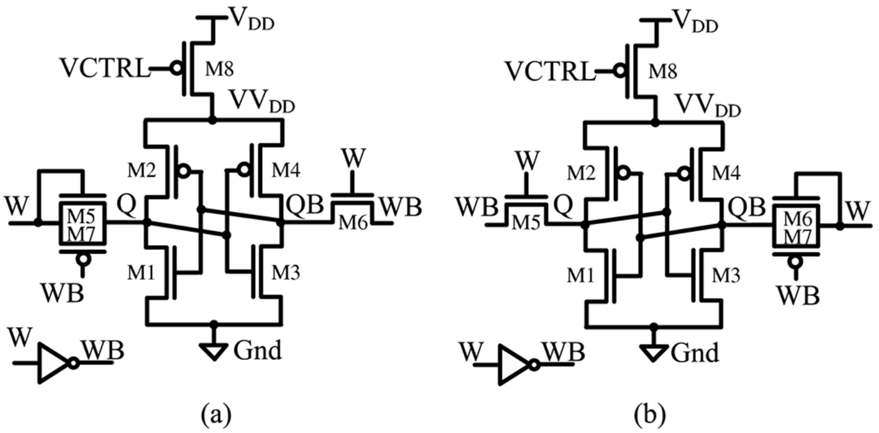


Fig. 4. (a) Canary cell “1” schematic. (b) Canary cell “0” schematic [7].

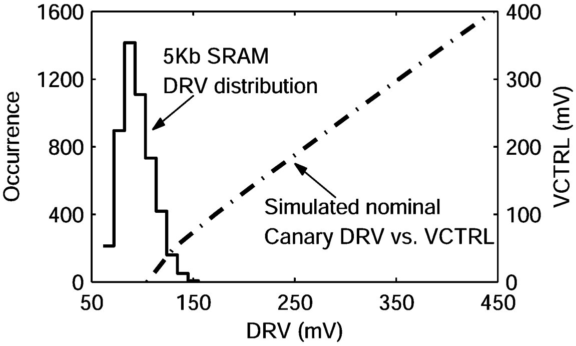


Fig. 5. Simulated nominal canary cell DRV versus VCTRL relative to a 5 Kb SRAM DRV distribution.

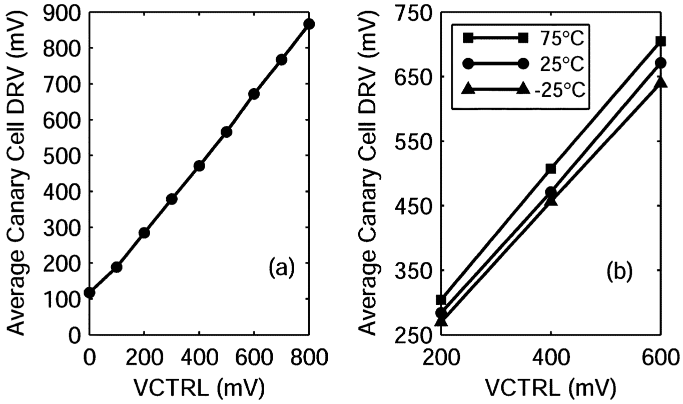


Fig. 6. Measured average canary DRV versus VCTRL at (a) room temperature and at (b) different temperatures [7].

B. Canary Cell

The canary cell must duplicate the impact of global changes on the core SRAM cell stability. Also, the canary cell must fail before the SRAM cells to prevent the loss of data in SRAM, which means it must use a design that makes it more sensitive to V_{DD} than it would be simply due to the impact of local variation. We proposed the circuit in Fig. 4(a) and (b) as canary cells to hold “1” and “0”, respectively [7]. The canary cell “0”/“1” contains the same 6T transistors (M1–M6) as any SRAM cell. Q and QB are the internal storage nodes. To enhance the write capability at subthreshold supply voltages (e.g., for canary reset), another PMOS pass transistor (M7) is added to the side of the cell that stores a “1”. The input signal, W, and its inversion, WB, act as the bitlines and wordline for writing data to the cells during a reset. When W is high, the canary cell resets its data; when it is low, the canary cell enters the standby mode. A PMOS header (M8) is inserted between the supply voltage of the canary cell and V_{DD}, and another input signal VCTRL drives its gate.

These small circuit modifications, especially the addition of the PMOS header, contribute to a higher DRV (failure voltage) for the canary cell. A larger VCTRL value increases the resistance of the header, and the actual supply voltage of the canary cell, VV_{DD}, drops lower than V_{DD}. This powerful knob essentially moves the mean of the DRV distribution for each canary cell across a wide range (as desired in Fig. 3). Fig. 5 shows the simulated DRV of the nominal canary cell versus VCTRL relative to the core cell DRV distribution. It is clear that the control of the header allows us to provide the desired continuum of failure voltages for the canary cells. It also illustrates the approximately linear relationship between the canary DRV and VCTRL, so canary DRVs can be placed at regular intervals above the core DRV using evenly spaced VCTRLs.

C. Silicon Results

A 90 nm CMOS bulk test chip implements all of the circuits that we have described except the V_{DD} regulator [7]. Fig. 6(a) shows the measured average DRV of canary cells versus VCTRL at room temperature. The canary cells exhibit the desired linear dependency on VCTRL. We also measured canary DRV with VCTRL at different temperatures as shown in Fig. 6(b), which demonstrates that the canary cells successfully track temperature changes.

In this paper, we will present additional measured results from the 90 nm test chip. Fig. 7 is the measured DRV histogram of one 8 Kb SRAM array and the measured DRV histogram of five canary categories (with VCTRL values ranging from 0 to 800 mV with a step of 200 mV). For testing the DRV distribution of one canary category (e.g., Canary #3), we use a test mode that sets all of the canary cells on the test chip to

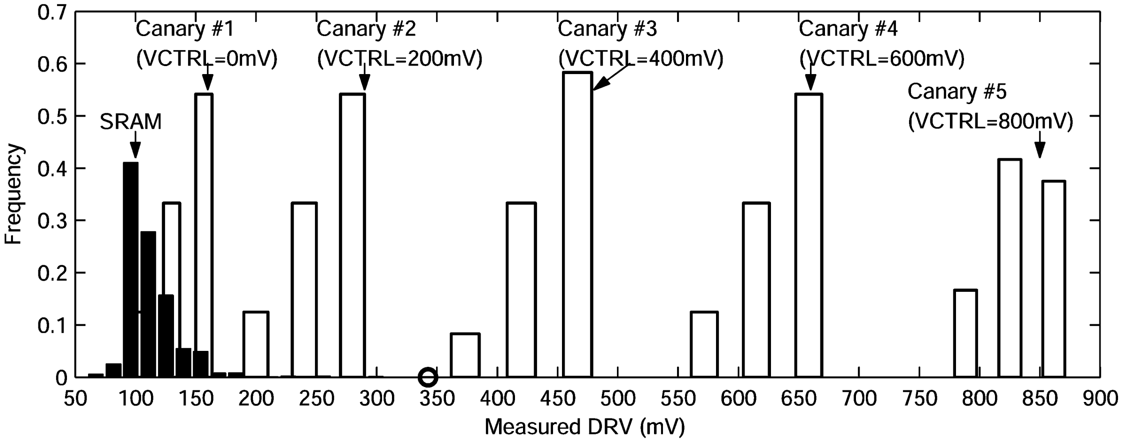


Fig. 7. Measured DRV histogram of one 8 Kb SRAM array and measured DRV histogram of five canary categories. The circle denotes the tail of the measured SRAM DRV distribution.

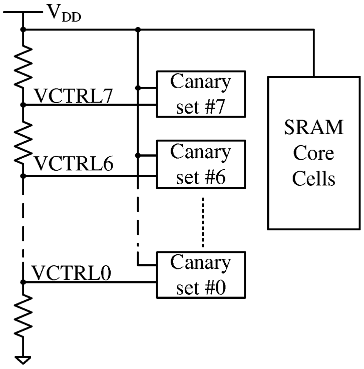


Fig. 8. One closed-loop measurement structure.

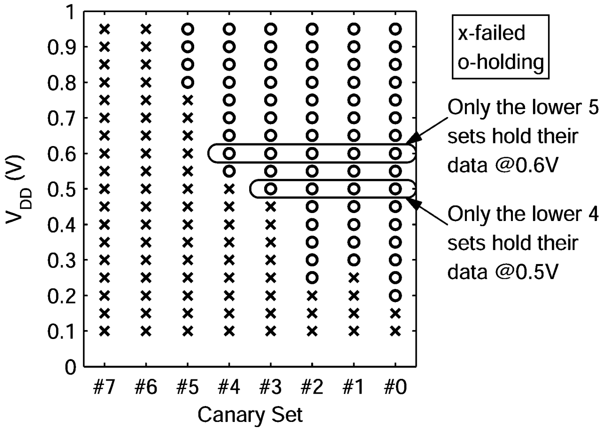


Fig. 9. Measured failure status of each canary set with V_{DD} scaling.

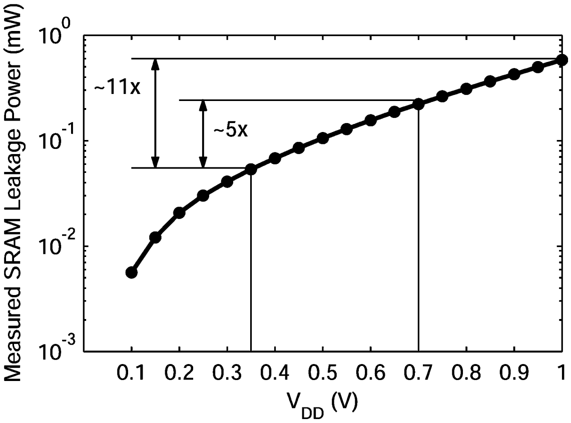


Fig. 10. Measured 128 Kb SRAM leakage power versus V_{DD}.

have the same VCTRL value (e.g., 400 mV) supplied by an external reference source. The measured SRAM array DRV values range from 60 to 350 mV with a mean value of 112 mV and a standard deviation of 22 mV. This wide distribution confirms the expected effects of local mismatch in the chip. Each measured canary category has a relatively narrower spread compared with the SRAM cells, and each one has a similar distribution. By applying different VCTRL values, we locate the failure voltage of different canary categories across a broad range that starts within the failure range of the core SRAM cells and extends to voltages well above the failure range of the core. This measured result proves the feasibility of implementing the tradeoff between SRAM reliability and leakage power, which was illustrated in Fig. 3.

We also tested one closed-loop control method shown in Fig. 8. A VCTRL generator (implemented as a resistor ladder) shares V_{DD} with the SRAM core cells as well as the canary cells. It consists of eight identical resistors so that eight regularly spaced voltage reference values can be generated. These nodes serve as the VCTRL signals and connect to the corresponding canary category (set). Hence in this test mode, the canary sets would fail in a sequence from set #7 to #0 as we continuously scale V_{DD}. Fig. 9 shows the measured results using this method. Here, each column is one canary set and each row shows the status of all the canary sets at one V_{DD} point. The cross symbol means the canary set fails and the circle symbol means it holds its data. For example, at $V_{DD} =$ 0.6 V, the upper three canary sets (with higher VCTRL) have failed, but the lower five sets continue to successfully hold their data. When further reducing V_{DD} to 0.5 V, the canary set #4 fails while the lower four sets keep holding their data. This figure demonstrates that lowering V_{DD} encourages more canary cells to fail, which then implies closer proximity to the failure of the core SRAM cells.

The measured leakage power of the SRAM array on one die with V_{DD} scaling is shown in Fig. 10. Without losing generality,

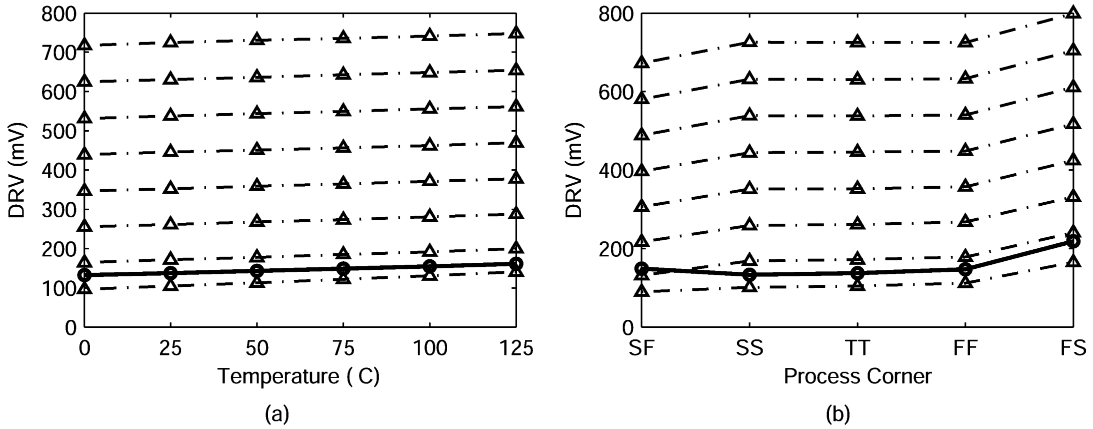


Fig. 11. Simulated DRV of the canary sets (lines with triangles and the upper ones have higher VCTRL) and the worst DRV of a 1 Kb SRAM (the line with circles) change consistantly with (a) temperature and (b) process corner for the 90 nm technology.

we can assume 0.7 V to be the standby V_{DD} for the worst-case scenario among all the PVT variations. Then for the die with the DRV tail at 0.35 V under normal environmental conditions, our canary-based feedback approach can adjust V_{DD} to this value and thus bring $\sim 5\times$ more power savings compared with the conservative worst-case-based approach, and $\sim 11\times$ compared with using the nominal V_{DD}.

III. Adapative Reaction to Canary Environment

In this section, we extend our previous work by taking a detailed look at the ability of the canary cells to track the impact of global effects on the core SRAM cells. We describe the circuits that allow the feedback loop to monitor the canary cells, even at low voltages, and we present a new analytical model that maps the VCTRL voltage to the DRV of the canary cells.

A. PVT Variation Tracking

One of the most important traits of the canary cell is its ability to track global PVT effects on the core SRAM cells. Without this characteristic, the feedback system cannot react properly under global stimuli. So it is necessary to examine the canaries under different PVT variations.

Fig. 11(a) and (b) shows simulated results that compare the canary behavior with an SRAM array across temperature changes and global process corners, respectively. We used a 1 Kb SRAM as an example. In this figure, the curve with circles stands for the worst DRV of the 1 Kb SRAM, and the curves with triangles stand for different canary sets (the upper ones are the sets with higher VCTRL). The upper seven canary sets consistently fail before the SRAM at all the temperatures and all the process corners with the only exception of the SF (Slow-N Fast-P) corner. This indicates that the canaries will successfully track global effects on the SRAM array. The one exception occurs because our technology is a strong-NMOS process (e.g., NMOS is noticably stronger than PMOS in subthreshold). At the SF corner, the impact of the global process variation becomes too weak compared with the local variations, so the whole SRAM DRV distribution (or tail) is not strongly influenced by this process variation for the canary set designed to fail closest to the core cells. If temperature gradients are a concern, then canary cells can be dispersed at different locations in a core array. If voltage fluctuation occurs, the DRV of core cells and canary cells will increase/decrease by the same amount because they are sharing the same power supply. Therefore, the canary cells are able to track PVT variations.

B. Failure Detector and Controller

We have demonstrated above that the canary cell can fail properly under all the PVT variations. However, without a robust failure detector and a good controller, the feedback system cannot successfully adjust V_{DD} even if the canary cells properly send out a failure alarm. So both the failure detector and the controller are critical components for our feedback system.

Fig. 12(a) shows the proposed circuit and structure for these components. Each canary cell connects to its own failure detector through the storage nodes Q and QB. Once Q and QB flip or converge to a single value, the detector should be able to capture that and assert the output "Fail" signal. Since the flipping failure is the major concern for the canary cell due to the asymmetrical bitlines, we propose a static sense amplifier as the failure detector. It shares V_{DD} with the canary cell. The inputs to the differential pair MN1 and MN2 come directly from the canary cell. For the canary cell "1" (Fig. 4(a)), Q connects to MN1 in this example; for the canary cell "0" (Fig. 4(b)), QB should connect to MN1 instead.

For simple illustration, only a single canary cell and its failure detector is shown here to connect with the controller. In fact, the failure signals from all the canary cells will be processed in order to generate a final "Fail" signal for the controller. Since we actually employ three-way redundancy in the banks of canaries in our test chip to reduce the impact of local variation on canary cells, the failure signals from the redundancy banks first go through the majority-3 gates to screen out abnormalities caused by rogue cells with large variations. Then the failure signals from canary bank "0" and canary bank "1" combine through the OR gates before comparing with the failure threshold. The failure threshold is a 8-bit value preloaded before operation. If the generated failures are larger than the failure threshold, a final failure signal will be asserted and sent to the controller. This is the signal that causes the controller to raise V_{DD} slightly and reset the canary cells.

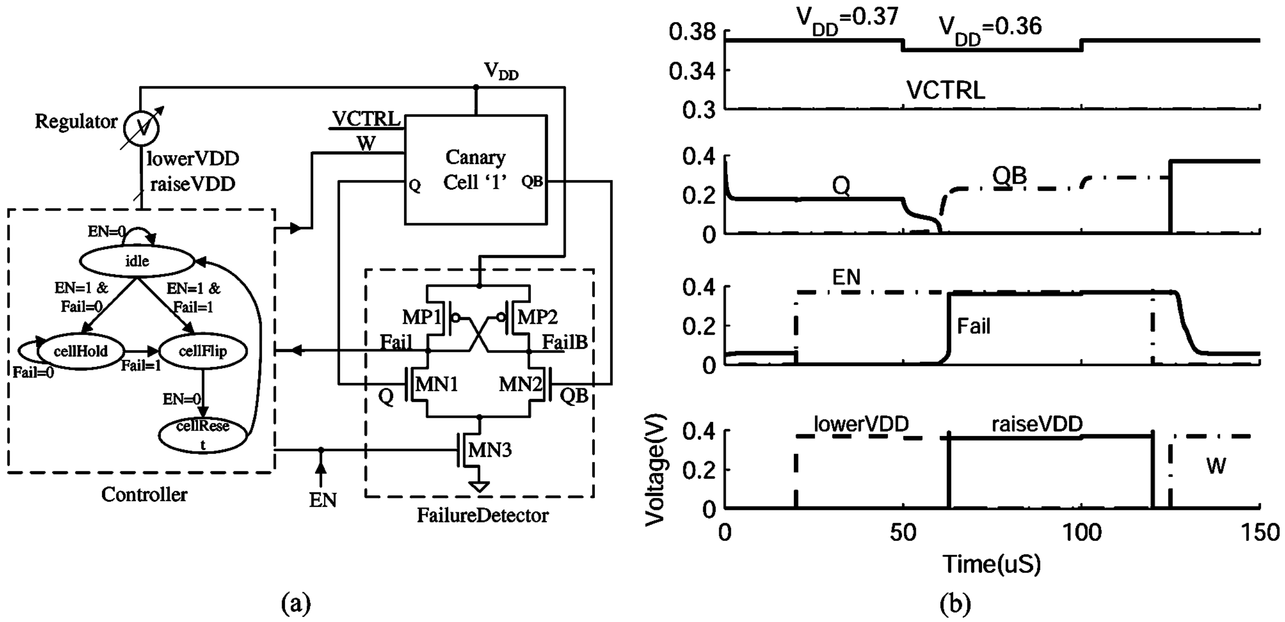


Fig. 12. (a) Canary failure detector and controller. (b) Timing diagram of detecting failure and resetting canary cell "1" when VCTRL is 0.3 V. The DRV of the canary cell is 0.37 V.

Fig. 12(a) also illustrates the state transistions in the controller that we implemented on the 90 nm test chip. There are four states: idle, cellHold, cellFlip and cellReset. The controller receives "Fail" signal from the failure detector, and sends out the two control signals "lowerVDD" and "raiseVDD" to the regulator and "W" signal to the canary cell for resetting.

Fig. 12(b) gives the timing diagram that shows how the states transfer for canary cell "1" with a 0.3 V VCTRL value. When V_{DD} is 0.37 V, Q and QB hold their original value. After we assert the enable signal "EN", the failure detector evaluates Q and QB, and then "Fail" goes to zero, which makes the controller change from the "idle" state to the "cellHold" state, and the signal "lowerVDD" rises up to inform the voltage regulator to decrease V_{DD} by one step of 0.01 V (for example). Once V_{DD} is lowered to 0.36 V, Q and QB flip to the opposite value, and hence "Fail" rises up and the "cellFlip" state becomes valid. This state asserts "raiseVDD" to make the regulator increase V_{DD} by one step and thus go back to the previous value 0.37 V, which is actually the DRV of this canary cell. After V_{DD} has raised up to 0.37 V, "EN" goes low to disable the failure detector and the controller enters the "cellReset" state, which asserts the "W" signal to write the original values into Q and QB.

Both the failure detector and the contoller have been implimented in our test chip and measured to function correctly at low V_{DD}.

C. Models for Adaptive Setting

Previously we have mentioned that our feedback scheme has the flexibility to trade off SRAM reliability with leakage power savings. This ability is dependent on the appropriate setting of the canary cells and failure threshold for a given SRAM for a required constraint on either relibility or power consumption. To make these settings more quickly and precisely, here we present two models to estimate SRAM DRV and canary DRV, and we integrate them into a final model that can directly compute the necessary canary VCTRL values to provide a desired level of SRAM reliability.

We have previously proposed the CDF and inverse CDF models for an SRAM DRV distribution in [8]. Equation (1) is the inverse CDF model of DRV:

$$F_{\text{DRV}}^{-1}(x) = \frac{1}{k}\left[\sqrt{2}\sigma_0 \cdot erfc^{-1}\left(2-2\sqrt{x}\right) - \mu_0\right] + V_0 \qquad (1)$$

where x is the probability that DRV $< F_{\text{DRV}}{}^{-1}(x)$, k is the slope of SNM High (SNM for 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 $erfc^{-1}(\cdot)$ 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 (1.5 K–5 K) Monte Carlo simulation. This model has shown a high accuracy in comparison with Monte Carlo simulation out to 6σ as well as in comparison with the Statistical Blockade tool [9] beyond 6σ, which uses a fast Monte Carlo method to filter the tail samples and fit them to a Generalized Pareto Distribution (GPD) model.

In this paper, we present a new model to estimate the canary DRV. As observed in Fig. 5, the canary DRV changes approximately linearly with VCTRL. This linear dependency can be modeled by analyzing the current through the PMOS header M8 (in Fig. 4(a)). Let us assume the minimum current to hold the canary cell data is $I_{\min}$, which occurs when the actual supply voltage of the canary cell, VV_{DD}, is equal to the cell DRV. Because the cell operates in the subthreshold region during the data retention mode, M8 will also operate in the subthreshold region. So the leakage current through M8, I_8, is

$$I_8 = I_0 \cdot \exp\left[\frac{V_{DD} - \text{VCTRL} - V_{T8} + \eta_8(V_{DD} - \text{DRV})}{n_8 V_{\text{th}}}\right] \times \left[1 - \exp\left(\frac{-V_{DD} + \text{DRV}}{V_{\text{th}}}\right)\right] \qquad (2)$$

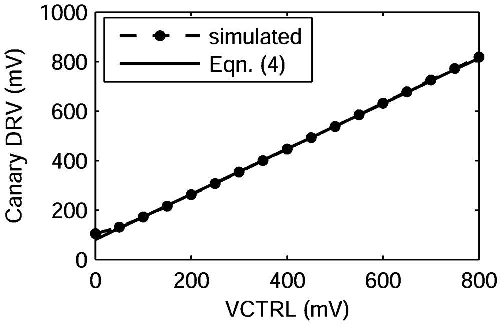


Fig. 13. Estimated canary DRV from (4) versus VCTRL compared with the simulated results.

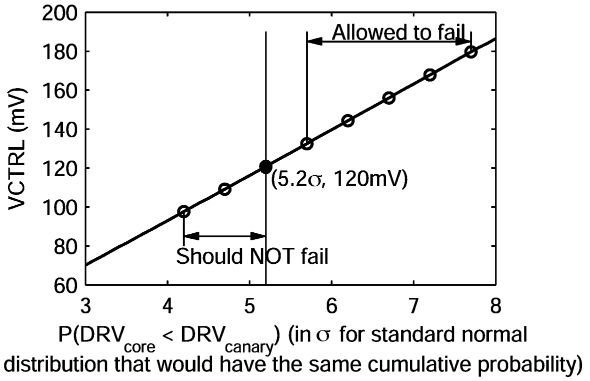


Fig. 14. Estimated VCTRL value versus the probability that $\mathrm{DRV}_{\mathrm{core}} < \mathrm{DRV}_{\mathrm{canary}}$ (in σ). Failure threshold (the vertical line) is set according to the reliability constraint, e.g., 5.2σ. Only the canary sets on the right side of the failure threshold (the upper five sets here) are allowed to fail.

where V_{T8} is M8's threshold voltage, η_8 is its DIBL coefficient, n_8 is its subthreshold swing factor, V_{th} is the thermal voltage, and I_0 is the off current. I_8 is also equal to $I_2 + I_4$, where I_2/I_4 is the leakage current through M2/M4. For a given canary cell, we assume that the DRV remains the same no matter what VCTRL is. This is reasonable because M1–M7 are not changed. Therefore, I_2 and I_4 keep constant and so does I_8. If we define that constant as I_C, (2) can be written as

$$\exp\left[\frac{(1+\eta_8)V_{DD}-\text{VCTRL}}{n_8V_{\text{th}}}\right]\left[1-\exp\left(\frac{-V_{DD}+\text{DRV}}{V_{\text{th}}}\right)\right] = \frac{I_C}{I_0}\exp\left(\frac{V_{T8}+\eta_8\text{DRV}}{n_8V_{\text{th}}}\right). \quad (3)$$

Since the right-hand side of (3) are all constant values, we can simply replace them with a constant K. Furthermore, when V_{DD} is much larger than the DRV, we can ignore the roll-off term. So finally, we can derive that

$$V_{DD} = \frac{\text{VCTRL} + n_8V_{\text{th}}ln(K)}{1+\eta_8} = \frac{\text{VCTRL}}{1+\eta_8} + b \quad (4)$$

which verifies the linear relationship between the canary DRV and VCTRL and implies that the slope is about $1/(1+\eta_8)$. With an initial pair of VCTRL_0 and V_{DD0}, we can obtain the offset value b. Fig. 13 compares the canary DRV values from (4) with the simulated results. This first-order linear model provides a good approximation for most VCTRL values. However, when VCTRL is less than 100 mV, the model is less accurate because V_{DD} is near the actual cell DRV, and the rolling-off term cannot be ignored, in which case (3) is a more accurate equation.

Now combining (1) and (4), we can estimate the VCTRL value necessary to satisfy a given SRAM reliability constraint with (5):

$$\text{VCTRL} = \frac{1+\eta_8}{\text{k}}\left[\sqrt{2}\sigma_0 \cdot \text{erfc}^{-1}\left(2-2\sqrt{\text{x}}\right) - \mu_0\right] + (V_0 - b)\cdot(1+\eta_8) \quad (5)$$

where x is the probability that SRAM DRV $(\mathrm{DRV}_{\mathrm{core}})$ is less than the canary DRV $(\mathrm{DRV}_{\mathrm{canary}})$ with the desired VCTRL value, and all the other parameters are the same as in (1) and (4). Fig. 14 shows the estimated VCTRL values from (5) with the solid curve. In this figure, the probability is expressed by σ, which is the equivalent point for a standard normal distribution that would have the same cumulative probability. For example, if 5.2σ probability is required (for a 100 Kb SRAM with 99% yield), VCTRL is about 120 mV. This implies that canary cells with VCTRL larger than 120 mV have an even higher probability of failing before all of the core cells. Now with the required SRAM reliability constraint, we can set an appropriate canary failure threshold. As in the example in Fig. 14, if at least 5.2σ reliability is needed, we can consider the vertical line at 5.2σ as the failure threshold and select the point (5.2σ, 120 mV) as one of the canary sets. Then we can pick the other seven points along the solid curve for the remaining canary sets so that the entire possible SRAM data reliability range can be covered. Here, we selected five points with VCTRL higher than 120 mV and two lower ones as an example. This configuration means that the feedback loop will allow only the upper five rows of canary sets (corresponding to the upper five points) to fail. We can know the approximate reliability of the core SRAM cells by observing the failure status of the canary sets. If the application changes and needs a higher reliability, we can reset the failure threshold for the current canary configuration or even reconfigure all of the canary sets (by remapping VCTRLs) for better results.

IV. Overhead Analysis

Thus far, we have analyzed the benefits of using canary-based V_{DD} scaling without accounting for overhead. In this section, we quantify the impact sources of overhead on the potential savings achievable by our scheme.

A. Canary Circuit Overhead

Our test chip has shown only about 0.6% area overhead due to the canary circuits. We can expect even smaller area overhead for systems with larger memory blocks.

The canary power overhead includes the power of canary array (48-bit canary cell) and peripheral (including all the failure detectors, controller, and other components) circuits. The dynamic power of the canary circuits is small relative to their leakage power since the canary system works at a very low frequency. We can thus consider the total overhead power to equal the leakage power of the canary circuits. Fig. 15 shows the leakage power of differently sized SRAM macros as well as the simulated canary circuit leakage power for the average (typical) PVT senario. To account for local variation impact on

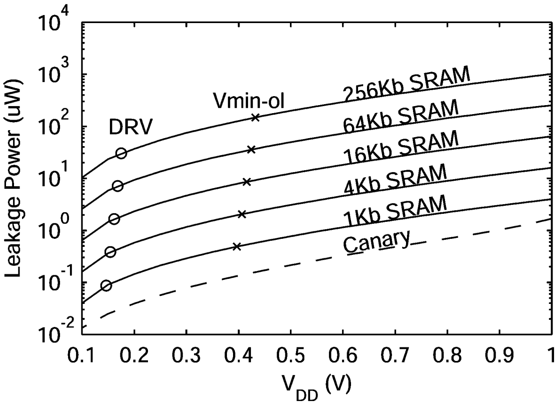


Fig. 15. Leakage power consumption of SRAM array with different size as well as canary power overhead at typical PVT scenario.

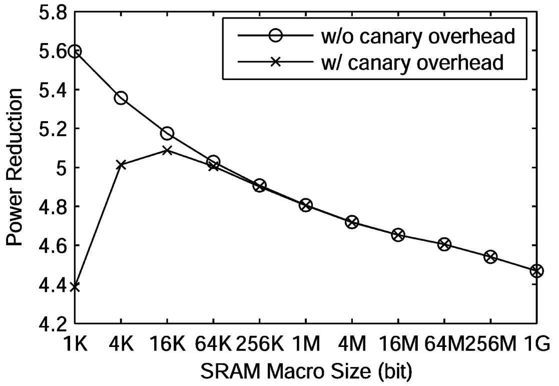


Fig. 16. Power reduction of using canary approach relative to the open-loop approach versus SRAM size (with or without taking account of the canary overhead) at typical PVT scenario.

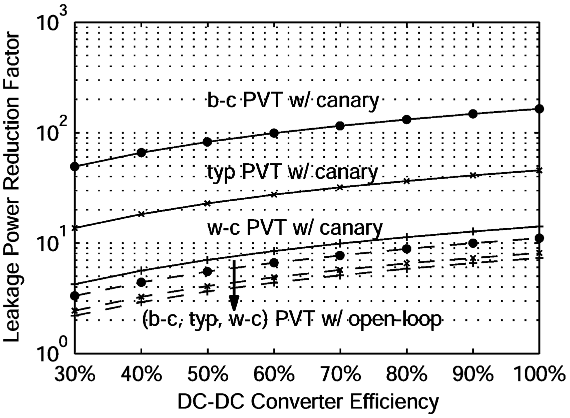


Fig. 17. Power reduction of 1 Kb SRAM using canary or open-loop V_{DD} scaling when DC-DC converter efficiency is considered. Power reduction is relative to the power consumed at the nominal V_{DD} (1.0 V). Best-case (b–c), typical (typ), and worst-case (w-c) PVT scenarios for each approach are shown.

the SRAM array, Monte Carlo simulation with mismatch can be used. However, for big arrays, running Monte Carlo simulation is too expensive. We use an alternative fast way that obtains SRAM leakage power from the statistics of the cell leakage current. We get I_{cell}, the mean of the cell leakage current distribution, from a 5000 iteration Monte Carlo simulation. The leakage current of a N-bit memory can be approximated as a normal random variable with the mean of NI_{cell} by applying the Central Limit Theorem. In Fig. 15, each solid curve denotes the average leakage power of the corresponding SRAM macro. The circle point on the curve denotes the DRV tail of this SRAM macro at typical PVT scenario, while the cross point denotes the open-loop Vmin "Vmin-ol" (i.e., V_{DD} when the worst-case PVT scenario has a 50 mV SNM margin) for this SRAM macro. Both the DRV and Vmin-ol values are obtained from our DRV inverse CDF model (1). With an SRAM larger than 64 Kb, the power overhead of all of the canary circuits is at least two orders of magnitude smaller than SRAM leakage power, so it is negligible.

Fig. 16 shows the power reduction achieved by the canary approach relative to the open-loop approach for different SRAM macro sizes for the typical PVT scenario. For each SRAM size, the canary approach uses the DRV value in Fig. 15 as the standby V_{DD}. While the open-loop approach uses the Vmin-ol value in Fig. 15 as the standby V_{DD}. Without considering the canary overhead, the smallest SRAM has the largest saving because it is possible to lower V_{DD} farther using feedback due to the lower DRV. However, accounting for the canary overhead shows that the effective savings with the smallest SRAM are reduced. It should be noted that all the SRAMs up to 1-Gb have more than 4× of power reduction compared with the open-loop approach. It indicates that our canary scheme is efficient for any size of SRAM in terms of leakage power reduction, even when accounting for canary overhead power.

B. DC-DC Converter Overhead

Our canary-based V_{DD} scaling approach requires a DC-DC converter that can provide a standby supply voltage across a fairly large range of values. Since this low, variable voltage is only supplied during standby, the load current may be relatively low. The DC-DC regulator may be on-chip or off-chip, but either way, we need to account for the impact of its efficiency on the overall power savings from using our approach. [10] has described a switched DC-DC converter that can deliver load voltages ranging from 0.3 V to 1.1 V. That particular converter provided $> 70\%$ efficiency over a wide range above 0.45 V. The minimum efficiency remained larger than 55%. This converter shows the sort of efficiencies that we might expect to see in this space. The recent interest in low-voltage operation is leading to further investigation of regulators with higher efficiencies that are tailored specifically to low supply voltages.

Fig. 17 shows the leakage power reduction factor for a 1 Kb SRAM by using our canary approach relative to using the nominal V_{DD} when accounting for a range of DC-DC converter efficiencies. The power reduction achieved by using the open-loop V_{DD} scaling approach is also shown. Best-case (b–c), typical (typ) and worst-case (w-c) PVT global scenarios for each approach are simulated. The open-loop approach uses the Vmin-ol value as shown in Fig. 15 so that the worst PVT scenario can have a guard-band of 50 mV SNM margin. The canary approach can apply V_{DD} down to the actual DRV of the given PVT global corner without losing data, so additional power savings can be achieved. However, when accounting for a non-ideal

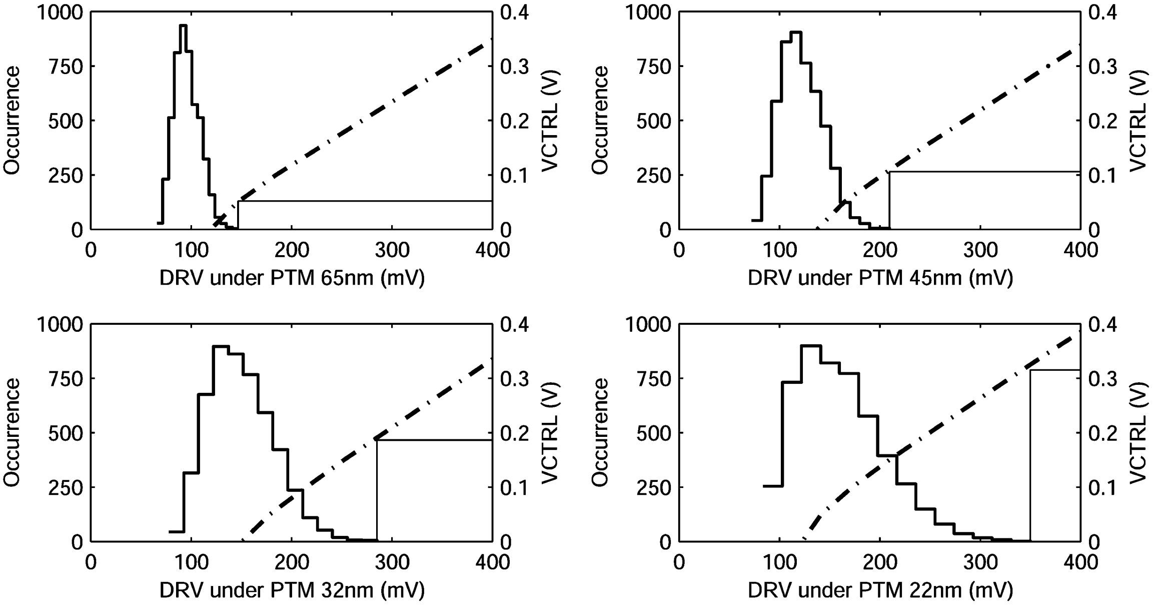


Fig. 18. 5 Kb SRAM DRV distribution (left axis) and canary DRV versus VCTRL (right axis) under PTM 65–22 nm nodes.

DC-DC efficiency, the actual power of both the canary approach and the open-loop approach will increase due to the overhead of the DC-DC converter. Under a given DC-DC efficiency, the power reduction factor is the ratio between the power consumed without V_{DD} scaling (i.e., $V_{DD} = 1.0$ V) and the actual power consumed with V_{DD} scaling. When the converter efficiency is only 30%, the canary approach gains $\sim$ 50× and $\sim$ 14× of power reduction while the open-loop approach gains $\sim$ 3× and $\sim$ 2× of reduction at the best-case and typical PVT scenario, respectively. Therefore, even when using a non-ideal DC-DC converter, it is obvious that V_{DD} scaling (by either open-loop or our canary approach) can bring substantial power reduction. In addition, for both the typical and the best-case PVT scenario, the power reduction from the canary approach with 30% converter efficiency is still higher than that from the open-loop approach with an ideal converter (100% efficiency), which demonstrates that our canary approach can effectively achieve extra power savings over the open-loop approach even under the condition of using a less efficient converter.

V. Scaling Effect

Using Predictive Technology Models (PTMs) for 65 nm, 45 nm, 32 nm, and 22 nm nodes[1] [11], we investigated the scaling of the canary DRV as well as the SRAM DRV for more advanced technologies. The SRAM cell transistor has the length of Lmin and a width of 2*Lmin (Lmin is the minimum length for the technology). The canary cell header transistor has the same sizing as the other 6T transistors. 3σ of the V_T local variation for 65 nm, 45 nm, 32 nm, and 22 nm is 10%, 15%, 20%, and 25% of the normal V_T, respectively.

Fig. 18 shows the SRAM DRV distribution and the canary DRV versus VCTRL for PTM nodes from 65 nm to 22 nm. The canary cell can keep the linearity property with VCTRL changes for all of the smaller technologies. This means that we can still create a continuum of failure voltages above the actual failure point of the SRAM array down to 22 nm. The plots also show that the SRAM DRV distribution has a higher mean and larger tail value as well as a larger standard deviation because of the increased variation at smaller dimensions. Therefore, we will need to use higher VCTRL values when using canary cells for SRAMs in smaller technologies to create failures above the DRV of the array.

Fig. 19 shows that the canary cells can track global process variation for 65 nm, 45 nm, and 32 nm nodes. For the 22 nm node, because of gate leakage, the canary DRV is no longer linear with the VCTRL (header gate voltage) value at some global process corners when VCTRL is high. This could potentially limit the range over which we can trade off power savings with reliability, but there is enough linearity to successfully deploy the canary scheme at 22 nm. If new techniques such as high-k materials provide the anticipated reduction of gate leakage, then the canary scheme will be able to offer a broad range of voltages for this tradeoff. These simulations indicated that our canary scheme can provide effective power reduction for future nodes down to 22 nm.

VI. Conclusion

A feedback scheme using canary replicas provides for aggressive V_{DD} scaling for SRAM standby leakage power reduction without losing data. We have provided new measured results from a 90 nm test chip that show that this feedback approach can provide about 5× reduction in leakage power compared with the conventional guard-banding approach. We have examined the adaptiveness of our canary scheme for tracking global stimuli. Simulation results show that the canary cells reliably fail in a continuum at higher voltages than the average of the core cells across PVT variations. Analysis of the overhead sources and simulation results in PTM technology nodes beyond 65 nm have shown that this is an efficient technique

[1] http://www.eas.asu.edu/~ptm

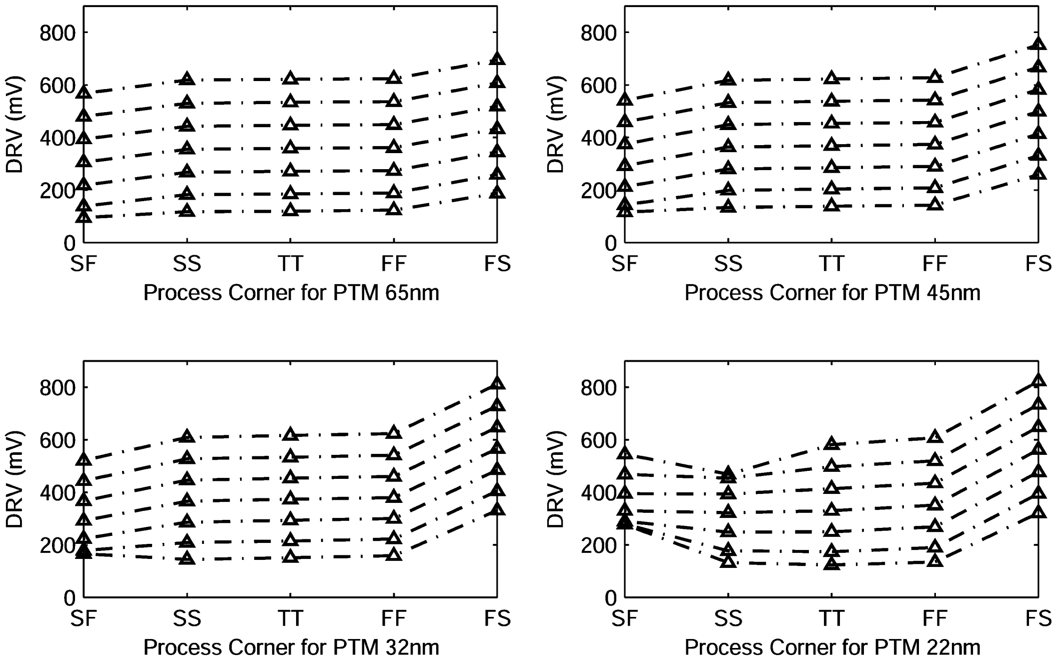


Fig. 19. DRV of canary categories (each line denotes one category and the upper ones have higher VCTRL values) at different process corners under PTM 65–22 nm nodes.

for saving SRAM standby leakage power for future nanometer SRAMs down to 22 nm.

Acknowledgment

The authors would like to thank A. Chandrakasan for his ideas and suggestions supporting this work.

References

[1] R. Krishnamurthy, A. Alvandpour, S. Mathew, M. Anders, V. De, and S. Borkar, "High-performance, low-power, and leakage-tolerance challenges for sub-70 nm microprocessor circuits," in *Proc. Eur. Solid-State Circuit Conf. (ESSCIRC)*, Sep. 2002, pp. 315–321.

[2] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 895–901, Apr. 2005.

[3] A. Keshavarzi, S. Narendra, S. Borkar, C. Hawkins, K. Royi, and V. De, "Technology scaling behavior of optimum reverse body bias for standby leakage power reduction in CMOS IC's," in *Proc. Int. Symp. Low Power Electronics and Design (ISLPED)*, 1999, pp. 252–254.

[4] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *Proc. Int. Symp. Quality Electronic Design (ISQED)*, 2004, pp. 55–60.

[5] N. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Single-Vdd and single-Vt super-drowsy techniques for low-leakage high-performance instruction caches," in *Proc. Int. Symp. Low Power Electronics and Design (ISLPED)*, Aug. 2004, pp. 54–57.

[6] L. Clark, M. Morrow, and W. Brown, "Reverse-body bias and supply collapse for low effective standby power," *IEEE Tran. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 9, pp. 947–956, Sep. 2004.

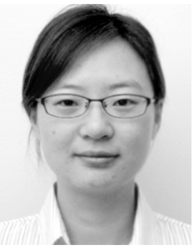
[7] J. Wang and B. Calhoun, "Canary replica feedback for near-DRV standby vdd scaling in a 90 nm SRAM," in *Proc. Custom Integrated Circuit Conf. (CICC '07)*, Sep. 2007, pp. 29–32.

[8] J. Wang, A. Singhee, R. Rutenbar, and B. Calhoun, "Statistical modeling for the minimum standby supply voltage of a full SRAM array," in *Proc. Eur. Solid-State Circuit Conf. (ESSCIRC)*, Sep. 2007, pp. 400–403.

[9] A. Singhee and R. Rutenbar, "Statistical blockade: A novel method for very fast Monte Carlo simulation of rare circuit events, and its application," in *Proc. Design Automation & Test in Europe Conf. & Exhibition (DATE '07)*, Apr. 2007, pp. 1–6.

[10] Y. Ramadass and A. Chandrakasan, "Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip applications," in *IEEE Power Electronics Specialists Conf. (PESC 2007)*, 2007, pp. 2353–2359.

[11] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit simulation," in *Proc. Custom Integrated Circuit Conf. (CICC 2000)*, 2000, pp. 201–204.



Jiajing Wang (S'06) received the B.S. degree in electrical engineering from the East China University of Science and Technology, Shanghai, China, in 2000, and the M.S. degree from Fudan University, Shanghai, China, in 2003. She is currently pursuing the Ph.D. degree at the University of Virginia, Charlottesville, VA.

From 2003 to 2005, she was with Agere Systems, Shanghai, where she worked on IC design and verification for communication systems. In the summer of 2007, she held an internship with Freescale Semiconductor to design low-power SRAMs. Her research interests include SRAM design and modeling for nanometer CMOS technologies, low-power and variation-tolerant digital circuits, and subthreshold digital circuits.

Benton Highsmith Calhoun (M'02) received the B.S. degree in electrical engineering with a concentration in computer science from the University of Virginia, Charlottesville, VA, in 2000. He received the M.S. degree and Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, in 2002 and 2006, respectively.

In January 2006, he joined the faculty at the University of Virginia as an Assistant Professor in the Electrical and Computer Engineering Department. His research interests include low-power digital circuit design, subthreshold digital circuits, SRAM design for end-of-the-roadmap silicon, variation-tolerant circuit design methodologies, and low-energy electronics for medical applications. Dr. Calhoun is a coauthor of *Sub-threshold Design for Ultra Low-Power Systems* (Springer, 2006).