Transaction Briefs

Minimum Supply Voltage and Yield Estimation for Large SRAMs Under Parametric Variations

Jiajing Wang and Benton H. Calhoun

***Abstract*—SRAM cell minimum operation voltage (Vmin) exhibits a skewed distribution in the presence of random parametric variations. Standard Monte Carlo (MC) simulation is prohibitively expensive to estimate the tail of the Vmin distribution for large SRAMs. We propose a fast and accurate method to estimate Vmin based on the statistical trend of static noise margin with V_{DD} scaling. Our preliminary work has shown its efficiency for standby Vmin estimation. In this work, we extend the method to estimate read and write Vmin and yield. We also generalize it for both symmetric and asymmetric types of cells. With comparable accuracy, the proposed model offers a huge speedup over standard MC. Compared with an alternative fast MC method, importance sampling, it shows a good agreement with less complexity.**

***Index Terms*—Minimum operation voltage (Vmin), Monte Carlo (MC), SRAM, static noise margin (SNM), variation, yield.**

I. Introduction

SRAM supply voltage (V_{DD}) has been lowered to achieve more power savings in dynamic voltage scaling (DVS) environments. SRAMs operating with V_{DD} near or below the threshold voltage (V_T) are also proposed for energy-efficient applications. In the presence of random variations such as dopant fluctuation, the minimum supply voltage (Vmin) is determined by the worst mismatched cell across the whole array. Since large SRAMs often contain millions of cells, the limiting event occurs only once out of millions of simulations. For such a rare event, the standard Monte Carlo (MC) method is prohibitively expensive. One way to reduce MC run time is to hasten the generation of the rare events. Interesting techniques include importance sampling (IS) [1], [2] and the statistical blockade (SB) tool [3]. However, the efficiency of IS and the SB tool relies on the quality of the sampling distribution and the tail filter, respectively. We can also shorten MC runs by using extrapolation when the target distribution is predetermined empirically or theoretically. Unfortunately, commonly used distributions (e.g., normal, log-normal, weibull, and gamma) either overestimate or underestimate the tail of the Vmin distribution (see Fig. 1). Therefore, we propose a method to estimate Vmin from another figure of merit, static noise margin (SNM), which is easier to model statistically.

Our preliminary work was focused on Vmin during standby mode for leakage power minimization. We derived a statistical model based on the sensitivity of the hold SNM distribution to V_{DD} [4]. In this paper, we make the following contributions.

Manuscript received October 26, 2009; revised February 22, 2010 and July 06, 2010; accepted August 07, 2010. Date of publication September 30, 2010; date of current version September 14, 2011. This work was supported by the MARCO/DARPA Focus Research Center for Circuit & System Solutions (C2S2).

J. Wang is with Intel Corporation, Hillsboro, OR 97124 USA (e-mail: jiajing.wang@intel.com).

B. H. Calhoun is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: bcalhoun@virginia.edu).

Digital Object Identifier 10.1109/TVLSI.2010.2071890

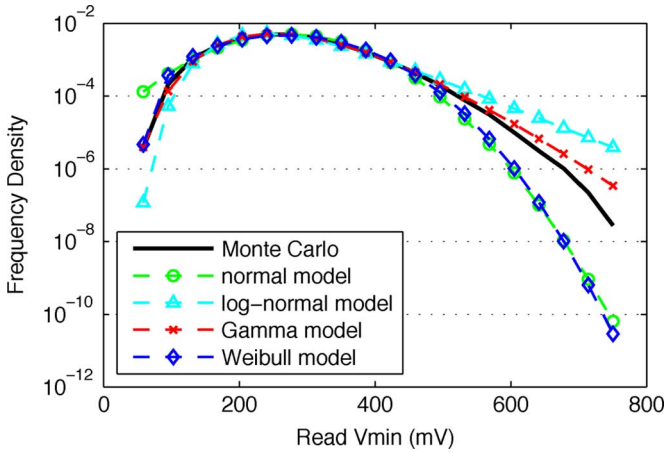

Fig. 1. Commonly used distribution models either underestimate or overestimate the worst tail of the read Vmin.

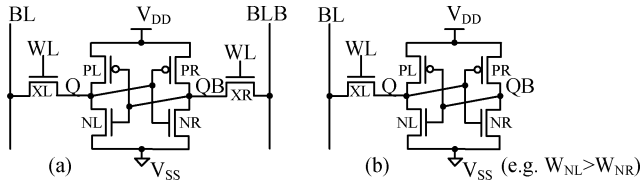

Fig. 2. Schematic of (a) a symmetric 6T cell and (b) an asymmetric 5T cell.

1) We further investigate Vmin for read and write operations, which are essential for active power reduction and even more susceptible to variations under a lower voltage. We discover that cell read stability and write ability also behave regularly as V_{DD} scales, so we generalize the Vmin model for all the operations.
2) Since the conventional symmetric 6T cell [see Fig. 2(a)] becomes more vulnerable to variations as technology scales, asymmetric cells such as the 8T cell [5] and the 5T cell [see Fig. 2(b)] [6] have been proposed for better stability. We hence generalize the model to support both symmetric and asymmetric cells.
3) The proposed model is compared with both standard MC and IS in a commercial 45 nm process for both the 6T and 5T cells.

Note that we use SNM in our work. Dynamic noise margin (DNM) is an alternative metric for cell stability. Since the read/write operation is performed in a dynamic fashion (i.e., with a pulsed WL), DNM is more accurate than SNM. Recently, statistical analysis of SRAM yield based on DNM has been proposed (e.g., [7]). However, DNM requires more complicated transient simulations that often cost longer run-time compared with simple dc simulations used in SNM. While SNM only measures the dc noise, it is strongly correlated with DNM in the tail region where the worst cells occur [8]. As a result, the SNM-based Vmin model can provide a good approximation of the tail of the Vmin distribution. Fast static analysis like ours can quickly screen out improper cell designs in the early design phase, and more accurate while more time-consuming dynamic analysis can be performed to optimize good design candidates.

II. RSNM and WSNM Statistics

We denote SNM0 and SNM1 as the SNM when the cell stores "0" and "1". The true SNM is the minimum of SNM0 and SNM1. We

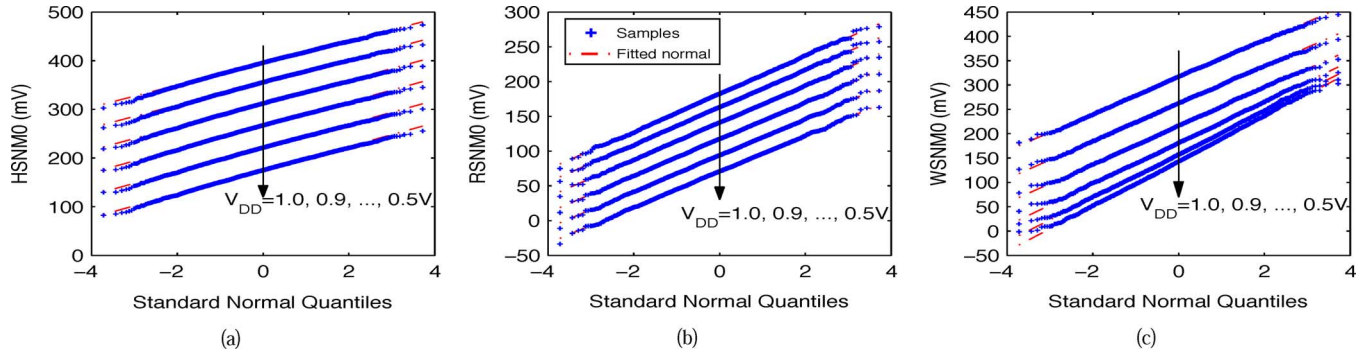

Fig. 3. Quantiles of (a) HSNM0, (b) RSNM0, and (c) WSNM0 against the standard normal quantile remain excellent linearity when $V_{\rm DD}$ scales.

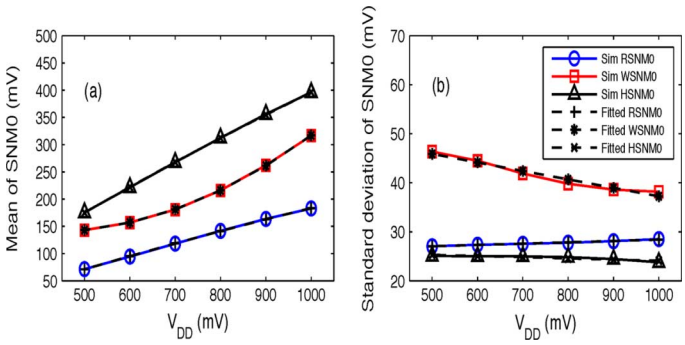

Fig. 4. (a) Mean and (b) standard deviation of the simulated RSNM0, WSNM0, and HSNM0 against $V_{\rm DD}$ are fitted to the polynomial model.

also call the SNM for hold, read and write operation HSNM, RSNM, and WSNM, respectively. In [4], we have observed that HSNM0 or HSNM1 remains normally distributed when $V_{\rm DD}$ scales [see Fig. 3(a)]. In this work, we further examine the statistics of RSNM and WSNM.

Similar to HSNM, butterfly curve based RSNM has been widely used as the measure of SRAM read stability. The RSNM1 (RSNM0) is the length of the maximum square that can be embedded inside the upper (lower) lobe of the butterfly curves. We use the WL-sweep approach to measure static write margin [9]. WSNM1 (WSNM0) is defined as the margin between $V_{\rm DD}$ and the WL voltage at which the nodes flip. Under normally distributed random parametric variation, Fig. 3(b) and (c) show the quantile-quantile (Q-Q) plot of RSNM0 and WSNM0 versus a standard normal variable at different $V_{\rm DD}$ values. The nice linearity of each curve implies that RSNM0 and WSNM0 can be approximated as a normal distribution at each $V_{\rm DD}$. Moreover, Fig. 4 shows that the sensitivity of the mean (μ) and the standard deviation (σ) of each SNM0 distribution to $V_{\rm DD}$ actually exhibits a nice trend, which can be fitted with the polynomial models:

$$\frac{\partial \mu}{\partial \mathrm{V_{DD}}} \approx a \cdot \mathrm{V_{DD}} + b, \quad \frac{\partial \sigma}{\partial \mathrm{V_{DD}}} \approx c. \tag{1}$$

Here, a, b, and c are fitting coefficients. If we know the estimate of μ and σ of SNM0 at one initial supply voltage v_0 are μ_0 and σ_0, then we can compute μ and σ at any new $V_{\rm DD}, v$, as

$$\mu = \mu_0 + a(v^2 - v_0^2) + b(v - v_0), \quad \sigma = \sigma_0 + c(v - v_0). \tag{2}$$

For symmetric cells such as the traditional 6T cell, the μ and σ of SNM1 are equal to those of SNM0; for asymmetric cells, their values might differ. Though the 8T cell [5] has the asymmetric structure (i.e., two additional nMOS devices create the buffer for reading data from one storage node), its SNM0 and SNM1 distribution are identical because it still uses the symmetric 6T cell as the latch. For asymmetric 5T cell in Fig. 2(b), they are different. In Section V-B, we will show their difference for the 5T cell. Since read and write SNM have similar statistical trends with $V_{\rm DD}$ scaling as hold SNM, we extend the statistical model in [4] to estimate Vmin for both standby and active operations. We also generalize our model to support both symmetric and asymmetric types of cells.

III. Vmin and Yield Model

Cell failure occurs when its SNM at current voltage $v(\mathrm{SNM}_v)$ is less than the acceptable noise margin s. We can compute the cell failure probability p_f as

$$\begin{aligned} p_f &= \mathrm{P}(\mathrm{SNM}_v < s) \\ &= P(\min(\mathrm{SNM0}_v, \mathrm{SNM1}_v) < s) \\ &= P(\mathrm{SNM0}_v < s) + P(\mathrm{SNM1}_v < s) \\ &\quad - P(\mathrm{SNM0}_v < s, \mathrm{SNM1}_v < s). \end{aligned} \tag{3}$$

From observation, SNM0 and SNM1 are more negatively correlated in the main body of their distributions. For a higher accuracy, we can approximate them with a bivariate Gaussian distribution. However, this will make (3) much more complicated and make it impossible to find a closed form solution for Vmin estimation. In addition, prior work in [10] has shown that the analytical model based on the assumption of independence of SNM0 and SNM1 matches very well with the tail of the SNM distribution. Therefore, here we also assume SNM0 and SNM1 are independent when computing the joint probability, i.e.,

$$P(\mathrm{SNM0}_v < s, \mathrm{SNM1}_v < s) = P(\mathrm{SNM1}_v < s) \cdot P(\mathrm{SNM0}_v < s). \tag{4}$$

Since SNM0 and SNM1 are normal random variables, we can express

$$\begin{aligned} P(\mathrm{SNM1}_v < s) &= \frac{1}{2}\mathrm{erfc}\left(\frac{\mu_h - s}{\sqrt{2}\sigma_h}\right) \\ P(\mathrm{SNM0}_v < s) &= \frac{1}{2}\mathrm{erfc}\left(\frac{\mu_l - s}{\sqrt{2}\sigma_l}\right) \end{aligned} \tag{5}$$

where $\mu_h(\mu_l)$ and $\sigma_h(\sigma_l)$ are the mean and standard deviation of $\mathrm{SNM1}_v(\mathrm{SNM0}_v)$, and can be individually substituted by (2). $\mathrm{erfc}(\cdot)$ is the complementary error function, which can be computed numerically. We finally obtain

$$\begin{aligned} p_f = &\frac{1}{2}\mathrm{erfc}\left(\frac{\mu_h - s}{\sqrt{2}\sigma_h}\right) + \frac{1}{2}\mathrm{erfc}\left(\frac{\mu_l - s}{\sqrt{2}\sigma_l}\right) - \frac{1}{4}\mathrm{erfc}\left(\frac{\mu_h - s}{\sqrt{2}\sigma_h}\right) \\ &\cdot \mathrm{erfc}\left(\frac{\mu_l - s}{\sqrt{2}\sigma_l}\right). \end{aligned} \tag{6}$$

When the cell is symmetric (like traditional 6T cell), $\mu_h = \mu_l$ and $\sigma_h = \sigma_l$. Equation (6) allows us to quickly estimate the cell failure probability at any $V_{\rm DD}$ without rerunning simulations at the new $V_{\rm DD}$ condition. Given a cell yield or failure probability, we can estimate the Vmin by numerically solving (6) and (2). Especially, for any symmetric

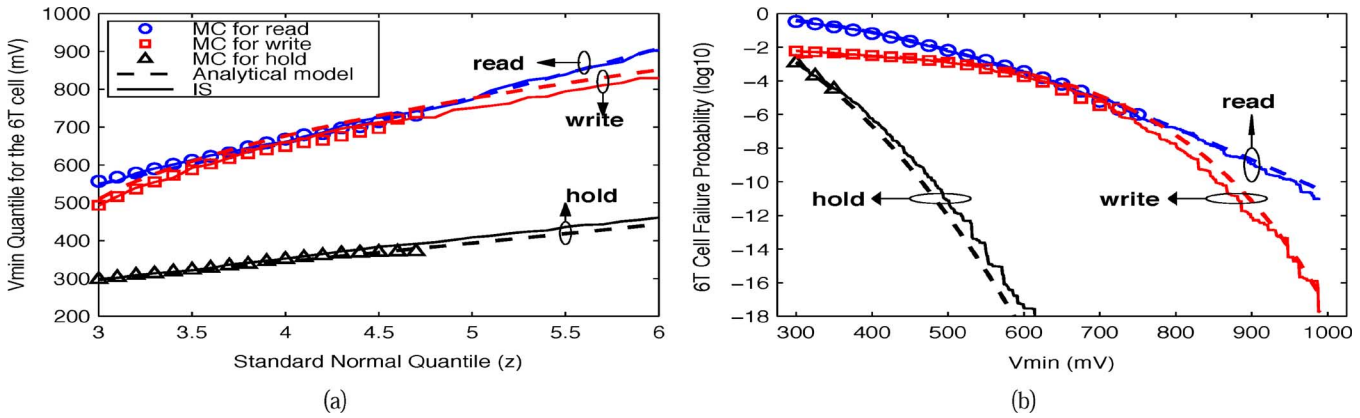

Fig. 5. Estimates of (a) Vmin and (b) cell failure probability of the 6T cell for read/write/hold from three methods.

cell, we can directly solve (7) to obtain the required Vmin value, v, for a cell failure probability p

$$av^2 + (b - c\lambda)v = av_0^2 + (b - c\lambda)v_0 + \lambda\sigma_0 + s - \mu_0 \quad (7)$$

where

$$\lambda = \sqrt{2} \cdot \mathrm{erfc}^{-1}(2 - 2\sqrt{1-p})$$

here $\mathrm{erfc}^{-1}(\cdot)$ is the inverse function of $\mathrm{erfc}(\cdot)$.

IV. Obtaining Vmin With MC and Importance Sampling

We compare the proposed method with standard MC and one fast MC method, importance sampling. In this section, we describe a fast simulation method to obtain read or write Vmin value for each MC sample. We also briefly introduce how we estimate Vmin with IS.

A. Vmin Simulation Methods

Since Vmin is the minimum V_{DD} for a non-negative SNM, we can search one sample cell's Vmin value through iterations of SNM simulations. For each iteration, we can simulate SNM with the V_{DD} value decreased by one step until the SNM drops below 0. The drawback of this simulation method is, it costs many dc simulations for each Vmin sample. To reduce simulation time, we use an alternative method to simulate Vmin with a single dc run.

For a read operation, we connect the cell supply voltage, WL, and BL/BLB to V_{DD}. Then we run a dc simulation by sweeping V_{DD} from high to low. The read Vmin is the V_{DD} value before Q and QB flip to the opposite state. For a write "1" ("0") operation, we tie the cell supply voltage, WL and BL (BLB) to V_{DD} and BLB (BL) to ground; Q and QB initially are holding "0" ("1") and "1" ("0"). Then we sweep V_{DD} from low to high. The write Vmin is the V_{DD} point where Q and QB start to flip. Thus, a single dc sweep replaces multiple SNM simulations to identify read or write Vmin.

B. Importance Sampling (IS)

IS is a widely used technique to reduce the variance of MC simulation. Suppose parameter X has the original density $f(x)$ and the sampling density $g(x)$, and Y is the output of an unknown function of X. Then the probability $p = P(Y > y)$ for some threshold y is estimated as [11]

$$\hat{p}(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} D(Y_i),$$

where

$$D(Y_i) = \begin{cases} 1, & Y_i > y \\ 0, & Y_i \le y \end{cases} \quad (8)$$

and n is the total number of samples. $(f(X_i))/(g(X_i))$ is also called a weight function $w(X_i)$. When $g(x) = f(x)$, (8) actually gives the estimator from standard MC. An estimator of the empirical quantile ξ for $\theta = P(Y \le \xi)$ is computed as

$$\hat{\xi} = (\max\{y : \hat{p}(y) > 1 - \theta\} + \min\{y : \hat{p}(y) \le 1 - \theta\})/2. \quad (9)$$

For the application of SRAM, since V_{T} variation has the biggest impact on cell stability, we only modify the density of V_{T} for each transistor and assume they are independent random variables $X_i, i \in [1, 6]$. Originally, X_i is a random normal variable $X_i \sim N(\mu_i, \sigma_i^2)$. The key of IS is to choose a good sampling distribution that can efficiently generate rare events. [1] used a complex adaptive form with the mixture of shifted and ratioed distributions. Recently, [2] proposed to use a widened distribution and showed that this simpler form of IS can accurately estimate failure probability of $p_f \le 10^{-10}$ for SNM. In this paper, we also use the simpler IS with the widened distribution $X_i \sim N(\mu_i, (\beta_i\sigma_i)^2)$, $\beta_i = 3$ for Vmin estimation.

V. Results

We test our method in a commercial 45-nm CMOS technology. Prelayout simulations are run under the typical process corner and room temperature. Without loss of generality, we choose 0 as the acceptable noise margin (i.e., $s = 0$).

A. Symmetric Case: 6T Cell

We first test our model with the conventional symmetric 6T cell [see Fig. 2(a)]. Fig. 5 plot the estimated Vmin and cell failure probability for each operation with three methods.

1) *Analytical Model*: The estimated Vmin from (7) and the cell failure probability from (6) are plotted as the dashed curves. We sample 6 V_{DD} points from 0.5 V to 1.0 V with the step of 0.1 V. At each V_{DD} point, 5000 MC simulations are run to obtain μ and σ of SNM. a, b, and c are extracted by fitting the curves of SNM0 μ and σ against V_{DD} (see Fig. 4). Note that we plot the calculated result with $v_0 = 800$ mV here. We will compare the results with different v_0 and discuss the sensitivity of the accuracy to v_0 in Section V-C.
2) *Standard MC*: The estimates from 1-million standard MC simulations with the methods described in Section IV-A are plotted with markers.
3) *IS*: Since the proposed analytical method uses totally 30 000 MC simulations, for fair comparison 30 000 MC samples with the new sampling distribution described in Section IV-B are simulated for IS. The Vmin estimates from IS are computed with (9) and the

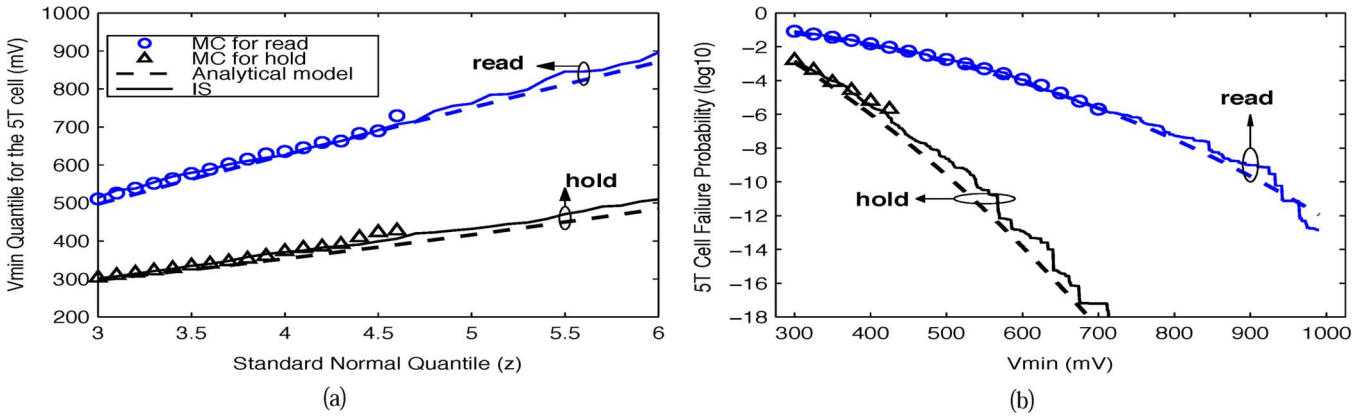

Fig. 6. Estimates of (a) Vmin and (b) cell failure probability of the 5T cell for read and hold from three methods.

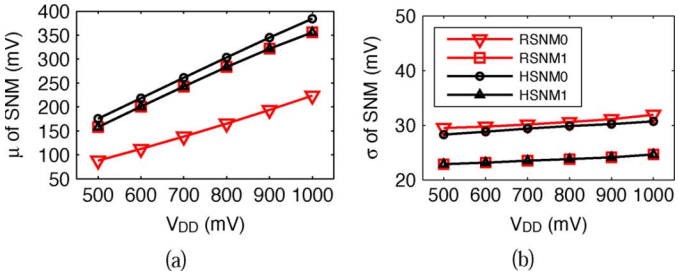

Fig. 7. (a) μ and (b) σ of the 5T SNM1 and SNM0 for read and hold against $V_{\rm DD}$ from MC simulations.

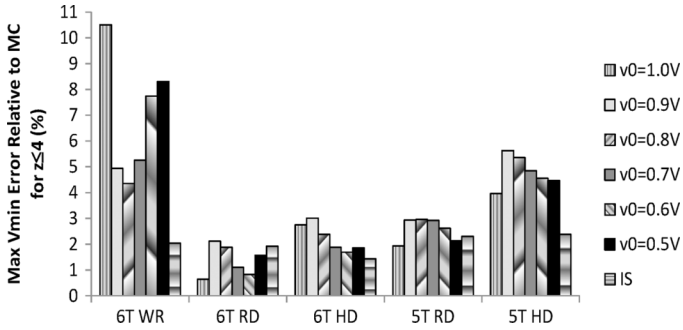

Fig. 8. Maximum error of the Vmin estimates from the proposed model when using different v_0 values and the maximum error of IS relative to standard MC when $z \leq 4$.

cell failure probability is computed with (8). They are plotted as solid curves.

Fig. 5(a) plots the 6T Vmin quantiles against the quantiles of a theoretical standard normal variable Z for read, write, and hold operation from each method. For the point at the (z, y) coordinates of the figure, $P(\text{Vmin} \leq y) = P(Z \leq z) = \Phi(z)$. Since SRAM arrays usually have at least 1000 bitcells, we are only interested in the quantiles larger than the 99.9th percentile, which is equivalent to $z = 3.09$ for a standard normal distribution. With 1-million MC samples, the maximum quantile we can estimate with MC is equivalent to $z \approx 4.7$. For tails at $z \leq 4$, the maximum error of the proposed model relative to the standard MC is 4.3%, 1.9%, and 2.4% for write, read, and hold; the maximum error of IS is 2.0%, 1.9%, and 1.4% for write, read, and hold. A relatively larger error occurs when $z > 4$. In fact, the MC result itself lacks confidence at this region because of fewer occurrences of rare events. With more MC samples, the difference in this region might be smaller. For $z > 4.7$, there is no MC sample available since the full MC is too costly. However, the good agreement between the model and the IS in the high sigma region enhances the confidence of their accuracy. Fig. 5(b) plots the estimated 6T cell failure probability from three methods for each operation. Similar to the Vmin estimation, the results of cell failure probability from our model show an excellent agreement with those from standard MC and IS.

B. Asymmetric Case: 5T Cell

We also test our model with an asymmetric 5T cell [see Fig. 2(b)]. NL is stronger than NR so that the butterfly curve can be skewed to achieve higher read stability [6]. The 5T cell has the major benefit of better read stability compared with the 6T cell and its write ability must be improved with write assist methods such as cell $V_{\rm DD}$ collapse. We mainly test the proposed model for the read and hold stability of the 5T cell. The extension of the model for cells with write and read assists will be explored in the future.

Fig. 7(a) and (b) show that both μ and σ of SNM1 and SNM0 for read and hold operation change linearly with $V_{\rm DD}$, though μ and σ of SNM1 are different with those of SNM0 due to asymmetry. Note that $\mu_{\rm HSNM0}$ becomes larger than $\mu_{\rm HSNM1}$ since $W_{\rm NL} > W_{\rm RL}$; $\mu_{\rm RSNM1}$ becomes larger than $\mu_{\rm RSNM0}$ and equal to $\mu_{\rm HSNM1}$ because of the absence of the access transistor on the right side. Unlike the 6T cell, a, b, c as well as the initial μ and σ of SNM1 are different with those of SNM0 in the 5T cell. We then estimate the 5T cell Vmin by numerically solving v from (6) and (2), and estimate the 5T cell failure probability by directly computing with (6) and (2). For comparison, we also run a 500 000-point standard MC simulation and a 30 000-point IS simulation. Fig. 6(a) and (b) plot the estimated 5T Vmin and cell failure probability from the analytical model with $v_0 = 800$ mV as well as the other two methods. For the asymmetric 5T case, our model offers 2.9% and 5.3% error for read and hold Vmin estimation relative to standard MC when $z \leq 4$. It also shows a good agreement with IS in the high sigma tail region where it is impractical to obtain standard MC result. In the next section, we will also present the accuracy of the proposed model at different v_0 for both the 6T cell and 5T cell.

C. Discussion

Fig. 8 plots the error of the analytical model under different v_0 values as well as the error of the IS method for $z \leq 4$ relative to standard MC. Since the proposed model is derived based on the approximation that both SNM1 and SNM0 are normal distributions at an arbitrary $V_{\rm DD}$, its accuracy is mainly dependent on how truly this approximation is. From a more careful examination of the 6T SNM quantiles at the tail region shown in Fig. 3, we observe that read SNM has the least error while write SNM has the largest error when fitting to the normal distribution. The tail of the write SNM distribution is particularly more skewed at

lower $V_{\rm DD}$. This larger deviation of write SNM distribution from the normal distribution leads to more errors as well as more sensitivity to v_0 for write Vmin estimation than for read and hold Vmin estimation. However, a careful selection of v_0 can reduce the error of write Vmin. Since the write SNM distribution is more skewed at lower $V_{\rm DD}$ and the estimated μ and σ are less accurate, we should avoid to pick v_0 too low ($\sim$ 0.5 V). Note that the true write Vmin for $z \leq 4$ is in the range of 0.5–0.7 V. Consequently, choosing v_0 too farther away from that range also leads to larger errors when fitting the write SNM mean and sigma against $V_{\rm DD}$. Hence, a moderate v_0 value (e.g., 0.7 or 0.8 V) can result in a smaller error of 6T write Vmin for the analytical model as shown in Fig. 8. The error of read and hold Vmin for 5T is relatively larger than the counterpart for 6T but is still within 3.0% for 5T read and 5.6% for 5T hold under all the v_0 conditions.

Compared with IS, the proposed method shows a similar accuracy for 6T read, 6T hold, and 5T read but a slightly lower accuracy for 6T write and 5T hold for $z \leq 4$. However, the variance of IS at the tail points beyond 5σ (i.e., $z > 5$) might become larger when using the simply scaled sampling distribution with only 30 000 samples. To reduce the variance of IS further out in the tail, we have to either run more simulations or choose a more proper sampling distribution for a specific tail point and rerun simulations. A serious overhead to IS is the time taken to devise and program the technique and to derive the desired weight function [12]. Though recently some new approaches are proposed to find a better sampling distribution when using IS for SRAM application (e.g., [13]), the extra time and efforts taken to analyze and obtain the desired sampling distribution cannot be neglected when evaluating the efficiency of IS. By contrast, the Vmin estimates from the proposed method with 30 000 samples show relatively small variance (e.g., 6T hold Vmin has $< 4\%$ error for 95% confidence interval in the tail region up to $z = 8$). In terms of additional cost other than simulation, the proposed method only needs the time to fit data to a normal distribution or a linear curve for extracting coefficients and then directly compute with (7) or numerically solve (6) and (2). This amount of time is negligible compared with the simulation time.

With comparable accuracy, our model offers a huge speed-up relative to standard MC. For instance, if we want to design a 1M-b SRAM with 99% yield, the cell failure probability must be smaller than 1e-08 (i.e., $z = 5.6$). For standard MC, this requires at least 100-million runs. But our method only requires a small number of MC runs (e.g., 1–5 k) for SNM at several typical $\rm V_{DD}$ points (e.g., 5–6). For example, we use 5 k $\times 6 = 30$ k MC simulations in the test. This amount of samples can provide a good accuracy for the estimation of SNM μ and σ as well as their sensitivity to $V_{\rm DD}$, especially for more regularly behaved read and hold SNM distributions. For write, more MC runs and/or $V_{\rm DD}$ points might be chosen in order to reduce error. In our test case, the number of simulations is reduced by 3300 times relative to the standard MC method. A modern computer might run more than 3 years to complete the required simulations for standard MC. With our method, the execution times can be reduced to only about 8 hours. Similarly, the IS method also gains a huge speedup over standard MC.

With comparable accuracy and significant speedup, both the proposed method and IS method can help SRAM designers to more efficiently improve 6T cell design or replace 6T cell with alternative cells such as the 5T cell.

1) For 6T Cell Improvement: From Fig. 5, the designer can quickly tell that the yield of this 6T cell is more limited by read operation, and hence read stability should be improved in the early design phase. Fig. 5(b) also informs the designer that the difference between read failure probability (p_{rf}) and write failure probability (p_{wf}) varies with the operation voltage. For $V_{\rm DD} > 880$ mV, p_{wf} is at least two orders of magnitude smaller than p_{rf}. Thus, it is safe to only enable a read assist technique but to disable the write assist technique for saving extra power and/or performance overhead from write assist. When $V_{\rm DD} \in [650, 800]$ mV, both the read and write operation are likely to fail at a moderate rate (<1e-4). Thereby, we should turn on both the read and write assist features. However, when $V_{\rm DD} < 650$ mV, both p_{rf} and p_{wf} are higher than 1e-4, which requires more efforts to assist these operations. So either more voltage bias should be applied in the assist methods or other redundancy and/or repair techniques such as ECC and row/column replacement should be activated. A quick and accurate estimation of the cell failures across all the possible voltages can help designers quickly find the best solution to improve yield with the minimum cost.

2) 5T Versus 6T: By using our model, we can quickly compare the 5T cell with the 6T cell in terms of the Read/hold Vmin and cell failure probability. Using the 5T cell, the read Vmin is reduced by 53.9 mV at $z = 5.6$ and the read cell failure probability is reduced by 4.02 times at $V_{\rm DD} = 800$ mV. However, the hold Vmin at $z = 5.6$ is increased by 35.7 mV and the hold cell failure probability at $V_{\rm DD} = 400$ mV is increased by 5.12 times. Therefore, the 5T cell improves read stability with the slight degradation on hold stability.

VI. Conclusion and Future Work

Standard MC is too expensive to estimate the extreme Vmin values of large SRAMs. Instead of directly tackling the skewed Vmin distribution, we obtain the Vmin values from the well-behaved SNM distributions, which change regularly with $V_{\rm DD}$ scaling. Our previous work has applied this method for standby Vmin estimation. In this paper, we discover similar properties for read and write SNM, and thereby derive a generic analytical model to estimate Vmin and yield for all the operations. The model is also extended to support both symmetric and asymmetric cells. Our model offers comparable accuracy and a significant speedup ($> 10^3\times$) over standard MC. It also shows an excellent agreement with a more generic fast MC method, Importance Sampling, but with less complexity and smaller estimate variance further out in the tail. One direction of our future work is to further extend the model to estimate Vmin and yield when read and write assist methods are used. The use of our model under the aging effects such as NBTI can also be explored.

References

[1] R. Kanj, R. Joshi, and S. Nassif, "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events," in *Proc. DAC*, 2006, pp. 69–72.

[2] T. S. Doorn, E. ter Maten, J. Croon, A. Di Bucchianico, and O. Wittich, "Importance sampling monte carlo simulations for accurate estimation of SRAM yield," in *Proc. ESSCIRC*, 2008, pp. 230–233.

[3] 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. DATE*, 2007, pp. 1–6.

[4] J. Wang, A. Singhee, R. A. Rutenbar, and B. H. Calhoun, "Statistical modeling for the minimum standby supply voltage of a full SRAM array," in *Proc. ESSCIRC*, 2007, pp. 400–403.

[5] L. Chang, D. Fried, J. Hergenrother, J. Sleight, R. Dennard, R. Montoye, L. Sekaric, S. McNab, A. Topol, C. Adams, K. Guarini, and W. Haensch, "Stable SRAM cell design for the 32 nm node and beyond," in *Proc. Symp. VLSI Technol.*, 2005, pp. 128–129.

[6] S. Nalam and B. H. Calhoun, "Asymmetric sizing in a 45 nm 5T SRAM to improve read stability over 6T," in *Proc. CICC*, Sep. 2009, pp. 709–712.

[7] D. E. Khalil, M. Khellah, N.-S. Kim, Y. Ismail, T. Karnik, and V. De, "Accurate estimation of SRAM dynamic stability," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 16, no. 12, pp. 1639–1647, 2008.

[8] J. Wang, S. Nalam, and B. H. Calhoun, "Analyzing static and dynamic write margin for nanometer SRAMs," in *Proc. ISLPED*, 2008, pp. 129–134.

[9] Z. Guo, A. Carlson, L.-T. Pang, K. Duong, T.-J. K. Liu, and B. Nikolic, "Large-scale read/write margin measurement in 45 nm CMOS SRAM arrays," in *Proc. Symp. VLSI Circuits*, 2008, pp. 42–43.

[10] B. Calhoun and A. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 41, no. 7, pp. 1673–1679, Jul. 2006.
[11] T. C. Hesterberg, "Advances in importance sampling," Ph.D. dissertation, Dept. Statistics, Stanford Univ., Stanford, CA, 1988.
[12] P. Smith, M. Shafi, and H. Gao, "Quick simulation: A review of importance sampling techniques in communications systems," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 4, pp. 597–613, May 1997.
[13] J. Wang, S. Yaldiz, X. Li, and L. Pileggi, "SRAM parametric failure analysis," in *Proc. DAC*, 2009, pp. 496–501.

Digit-Level Semi-Systolic and Systolic Structures for the Shifted Polynomial Basis Multiplication Over Binary Extension Fields

Arash Hariri and Arash Reyhani-Masoleh

***Abstract*—Finite field multiplication is one of the most important operations in the finite field arithmetic. In this paper, we study semi-systolic and systolic implementations of the shifted polynomial basis multiplication and propose low time complexity semi-systolic and systolic array structures. We show that our proposed semi-systolic multiplier is faster than its existing counterparts available in the literature. Our application-specified integrated circuit (ASIC) implementation of the proposed semi-systolic multiplier demonstrates that reduction in time complexity is achieved without imposing hardware overhead. Furthermore, our proposed systolic array shifted polynomial basis (SPB) multiplier has a low time complexity for general irreducible polynomials.**

***Index Terms*—Binary extension fields, digit-level, multiplication, semi-systolic, shifted polynomial basis, systolic.**

I. Introduction

Cryptographic algorithms such as elliptic curve cryptography (ECC) require different finite field arithmetic operations. Efficient design and implementation of these operations affects the performance of cryptosystems and consequently, has gained lots of interest in the literature, e.g., [1]–[3], and [4]. One of the main finite field arithmetic operations is the multiplication. The shifted polynomial basis (SPB), proposed in [5], is a variation of the polynomial basis (PB). The available works in the literature show that using the SPB results in efficient arithmetic units, e.g., [1], [6]–[9], and [10]. In [1], bit-parallel multipliers are designed for irreducible trinomials and type-II pentanomials, which are faster than the best known polynomial basis and dual basis multipliers. Similarly, it is shown in [6] that the SPB squarers are faster than their PB counterparts. Using the SPB, a new approach for designing subquadratic area complexity parallel multipliers is outlined in [7], where the reported multipliers are better than the other similar ones in terms of area and time complexities. Also using the SPB, different bit-parallel multipliers are designed for irreducible pentanomials and trinomials in [8] and [9], respectively. A parallel digit-serial SPB multiplication algorithm is proposed in [10] which has lower time complexity than the PB and Montgomery multiplication (MM) algorithms.

Manuscript received September 24, 2009; revised February 17, 2010 and July 01, 2010; accepted July 21, 2010. Date of publication September 13, 2010; date of current version September 14, 2011.The work of Arash Reyhani-Masoleh was supported in part by an NSERC Discovery grant.

The authors are with the Department of Electrical and Computer Engineering, The University of Western Ontario, London, ON N6A 5B9, Canada (e-mail: hariri@ieee.org, areyhani@uwo.ca).

Digital Object Identifier 10.1109/TVLSI.2010.2066994

A straightforward implementation of the projective Montgomery scalar multiplication requires up to $(m-1)(6M+3A+5S)+(10M+7A+4S+I)$ clock cycles, where M, A, S, and I represent the number of clock cycles for multiplication, addition, squaring, and inversion, respectively [11]. The inversion using Itoh-Tsujii algorithm requires $\lfloor \log_2(m-1) \rfloor + H(m-1) - 1$ multiplications and $m-1$ squarings, where $H(m-1)$ denotes the Hamming weight of $(m-1)$ [11]. As a result, accelerating multiplication significantly affects the performance of an elliptic-curve based crypto-system.

Semi-systolic array structures provide low latency in comparison to systolic array implementations and require fewer latches. Also, they can be pipelined to increase the throughput of the system. In the literature, semi-systolic array implementations have been presented for the finite field multiplication, see for example [12]–[15], and [16]. In the case of the PB, a classic multiplication structure is proposed in [12] which is studied in [13] comprehensively. For the Montgomery multiplication, [14] introduces a semi-systolic structure. Also, [15] and [16] introduce low-latency semi-systolic Montgomery multipliers.

In systolic array structures, the global lines are avoided and the connections are limited to local ones. This results in more efficient VLSI implementations. In case of the PB multiplication, [3] and [17] outline two structures for general irreducible polynomials, respectively. In [18] and [19], optimized structures are proposed for the PB multiplication using general irreducible polynomials and irreducible trinomials. A low latency systolic structure is proposed in [20] for all-one and equally spaced polynomials. Moreover, digit-serial systolic PB multipliers are proposed in [21], [22], and [23] for general irreducible polynomials. A systolic implementation of the PB multiplication is proposed in [24] for irreducible trinomials with a low latency. In case of the Montgomery multiplication, [25] proposes very low latency systolic multipliers for special irreducible polynomials. Also, two scalable structures are proposed in [26] and [27].

The two contributions of this paper are stated as follows. The first contribution of this paper is introducing a new low time-complexity digit-level semi-systolic array structure for the SPB multiplication. The proposed structure is based on a similar technique used in [15], [16], [28], and [10]. In our proposed structure, the parallel operations are balanced and have the same critical path delay. The Montgomery multipliers presented in [29] include non-pipelined structures for special cases of irreducible polynomials. The semi-systolic structure presented in this paper is a low-latency pipelined multiplier with low critical path delay. We show that our proposed semi-systolic multiplier has the least time complexity among the existing ones available in the literature including [12]–[16], and [30]. The second contribution is to propose a digit-level systolic array SPB multiplier which offers a better time complexity, in terms of the combination of the critical path delay and latency, than the existing counterparts for general irreducible polynomials, such as [3], [17], [19], [23], [26], and [27].

The rest of this paper is organized as follows. In Section II, we present our semi-systolic array implementation of the SPB multiplication. In Section III, we propose a digit-level systolic array structure for the SPB multiplication. In Section IV, we provide our implementation results and comparisons. Finally, we conclude this paper in Section V.

II. Semi-Systolic SPB Multiplication

The binary extension field $GF(2^m)$ includes 2^m elements and is associated with an irreducible polynomial defined as

$$F(z) = z^m + f_{m-1}z^{m-1} + \cdots + f_1 z + 1, f_i \in \{0,1\} \quad (1)$$

for $i = 1$ to $m-1$. If x is a root of $F(z)$, i.e., $F(x) = 0$, the set $\{1, x, x^2, \ldots, x^{m-1}\}$ is known as the polynomial basis (PB). Assuming $0 < v \leq m$ is an integer, the Shifted Polynomial Basis (SPB)