Dynamic Write V_{MIN} and Yield Estimation for Nanoscale SRAMs

Shourya Gupta, *Graduate Student Member, IEEE*, and Benton H. Calhoun, *Fellow, IEEE*

***Abstract*— The dynamic write-access operation in an SRAM has a long-tailed distribution that sets the threshold for write-access failures. The distribution takes a long time to evaluate since rare failure events lie in its long and heavily skewed tail. Moreover, advanced FinFET technologies are becoming increasingly reliant on assist techniques to resolve these rare failures in the tail for improved dynamic performance and stability. In this work, we present various analytical approaches that work well in both super-threshold and subthreshold regions of operation to quickly determine the write-access failure probability. For FinFET based SRAMs, we present a modified sensitivity analysis-based method to evaluate the write-access operation distribution and discuss the evaluation of contention-limited write-access failures. The impact of various write-assist techniques on the performance and stability of FinFET SRAMs is also discussed. All simulations are performed using commercial 65nm bulk planar and 12nm FinFET technologies.**

***Index Terms*— Bit-cell, failure probability, noise margin, write-access, SRAM, subthreshold, V_{MIN}, yield.**

I. Introduction

RANDOM variations in nano-scale Static Random Access Memories (SRAM) pose a major challenge to achieving design robustness due to their large effect on bit-cell and array characteristics. These variations include device threshold voltage (V_T) mismatch due to random dopant fluctuations (RDF) and line edge roughness (LER) [1]. The device V_T mismatch in deep sub-micrometer technologies is greatest in minimum sized devices, which are often used in SRAMs [2]. The worst-case V_T mismatch, combined with the increased sensitivity of current in the subthreshold region, greatly affects the minimum operating voltage (V_{MIN}) and yield of the memory. This makes it hard to design low power SRAMs or meet frequency and yield constraints. Monte-Carlo (MC) simulation is a well-known approach to determine the worst-case V_{MIN} for a given memory. However, memory arrays can require millions of MC simulations, which is prohibitively expensive since most of the samples do not lie in the tail [3]. For especially long tailed distributions such as the dynamic write-access operation, accurate determination of the tail is crucial to determining the failure point. A few analytical approaches have previously been proposed to determine the dynamic write-access distribution. However, they have limited success in doing so, especially in sub-threshold region of operation, due to the approximations considered in fitting the distribution which we discuss in Section II.

This work provides an in-depth analysis of analytical methods to evaluate the write-access distribution in both bulk planar and advanced FinFET technologies. In this work, we discuss:

1. An analytical transformation model to determine the failure probability of the write-access operation, that works well in both super-threshold and sub-threshold regions of operation.
2. Since the write-access operation distribution resembles the noncentral *F* distribution, we also discuss an analytical solution to determine its tail.
3. In more recent advanced FinFET technologies, the gate work function variations impact the V_T variations significantly, which is why we present a modified sensitivity analysis-based method to determine the distribution of the write-access operation in advanced technologies.
4. A methodology to determine the contention-limited write-access failures which occur irrespective of pulse width.
5. Since in advanced FinFET technologies, assist circuits play a crucial role in ensuring the continued scaling of SRAMs, we also compare various write assist techniques based on their effect on performance and stability across different bit-cell fin ratios and regions of operation.

These proposed alternative analytical methods offer a fast approach with reasonably low error to determine the write-access operation failure threshold and yield in a given SRAM design process in both sub-threshold and super-threshold regions of operation.

The paper has been constructed as follows. Section II briefly discusses the prior analytical methods used to determine the write-access distribution and then describes the proposed transformation-based methodology and the noncentral *F* distribution solution. It then describes the sensitivity analysis-based method for determining the write-access distribution for FinFET based SRAMs. Furthermore, contention-limited write failures are also discussed. Section III discusses various write-assist techniques and their impact on the write-ability and performance in advanced technologies and Section IV summarizes and concludes the paper.

II. SRAM Write-Access Modelling

Write failure is caused when an SRAM cell is unable to reach desired value in the time duration of the clock pulse width. Therefore, the write failure probability can be

Manuscript received March 14, 2021; revised June 20, 2021 and August 31, 2021; accepted September 2, 2021. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Agreement FA8650-18-2-7844. This article was recommended by Associate Editor M.-F. Chang. *(Corresponding author: Shourya Gupta.)*

The authors are with Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: shourya.gupta94@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2021.3110484.

Digital Object Identifier 10.1109/TCSI.2021.3110484

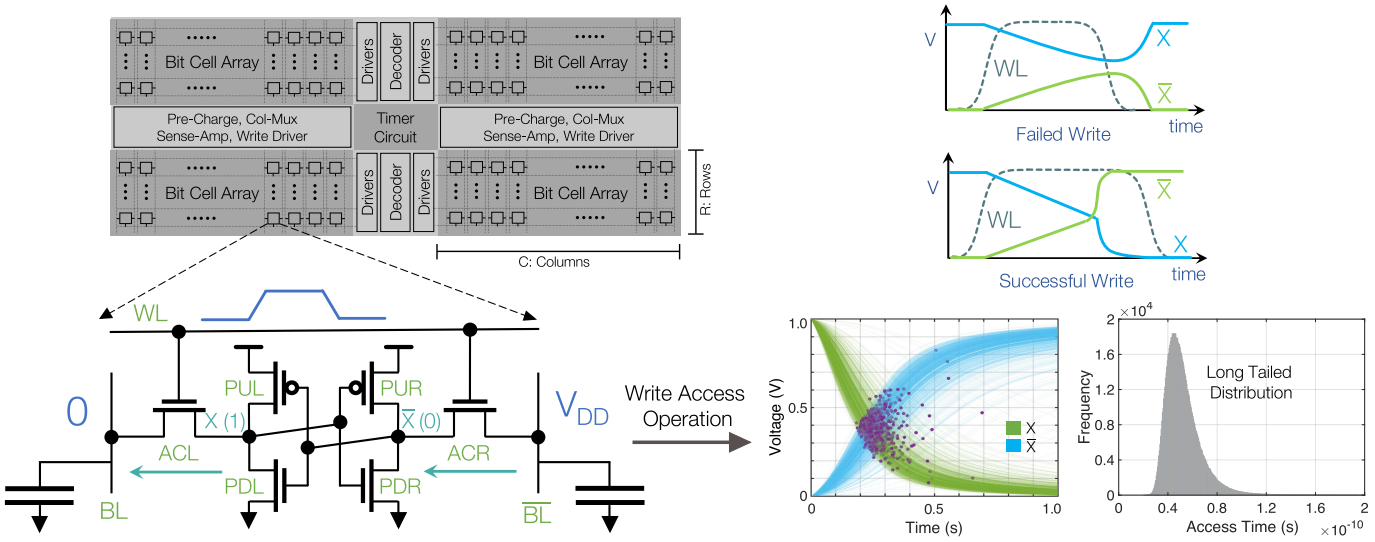

Fig. 1. Schematic of an SRAM bit-cell in write mode, depicting the corresponding dynamic write-access operation. Its distribution has a long tail which requires a large Monte-Carlo simulation for accurate determination and is difficult to model analytically.

expressed as

$$P_W = Prob\left(T_W < T_{WL}^W\right) \tag{1}$$

where TW is the time required to pull down the node storing 'one' and T_{WL}^W is the write word line (WL) pulse width [4], [5]. TW cannot be easily approximated because its distribution has a tail as shown in Fig. 1. In the following section, we briefly discuss previous analytical approaches for determining the distribution of the write access operation.

A. Existing Write-Access Analytical Modelling Approaches

The work presented in [5] shows that the tail of write-access operation closely resembles the noncentral F distribution. The authors in [6] use sensitivity analysis to estimate the tail for write-access, but report an error of 6.83%. The authors in [7] show that by performing a linear inverse transformation on the write-access distribution, it can be transformed into a Gaussian distribution, which can then be used to easily estimate the failure probability by calculating the mean ($\mu_{f(x)}$) and standard deviation ($\sigma_{f(x)}$) of the distribution as

$$\mu_{f(x)} \approx f(\mu_x) \tag{2}$$

$$\sigma_{f(x)}^2 \approx \sum_{i=1}^{n}\left[\left(\frac{\partial f\left(\mu_x\right)}{\partial x_i}\right)\sigma_{x_i}\right]^2 \tag{3}$$

$$P_{FAIL} = Prob\left(\frac{1}{T_W} < \frac{1}{T_{WL}^W}\right) = \Phi\left(\frac{\left(\frac{1}{T_W}\right)_{nom} - \frac{1}{T_{WL}^W}}{\sigma_W}\right) \tag{4}$$

Here Φ represents the standard normal cumulative density function, T_W is the write-access time and, T_{WL}^W is the word-line pulse width. The write-access operation distribution and its long tail nature is shown in Fig. 1. The inverse transformation of the same distribution is shown in Fig. 2 (a). As seen in Fig. 2 (a), the inverse of access time approximates the Gaussian distribution. This is because the linear inverse transformation of delay makes it vary as $\propto (V_{DD} - V_T)$, i.e., linearly with change in V_T. However, this is only true for super-threshold region of operation. In sub-threshold region, the delay has an exponential dependence and a linear transformation does not yield a Gaussian distribution as shown in Fig. 2 (b).

B. Proposed Write-Access V_{MIN} Modelling

To transform the delay values (D) in the write-access distribution such that we can obtain a Gaussian distribution in the transformed domain, we apply a transformation T as

$$T : D \rightarrow \frac{1}{K}\left|ln\left(|D|\right)^{\zeta}\right| \tag{5}$$

where K is an integer constant and ζ is a fit parameter such that the skewness of the transformed distribution approximates to zero. This is because the third moment of a Gaussian distribution is zero.

$$skewness\left[(T : f\left(x\right))\right] \approx 0 \tag{6}$$

An initial MC simulation of 100K runs is performed and then used to estimate the value of ζ using (5) and (6). This process is repeated for any circuit or technology each time during the evaluation of the transformation and corresponding distribution. The value of ζ as a function of V_{DD} is shown in Fig. 2 (i).

The write-access delay values of 100K runs are then transformed using (5), after which the failure probability is quickly calculated as

$$P_{FAIL} = Prob\left(\frac{1}{K}\left|ln\left(|T_W|\right)^{\zeta}\right| < \frac{1}{K}\left|ln\left(\left|T_{WL}^W\right|\right)^{\zeta}\right|\right)$$

$$= \Phi\left(\frac{\frac{1}{K}\left|ln\left(|T_W|\right)^{\zeta}\right|_{\mu} - \frac{1}{K}\left|ln\left(\left|T_{WL}^W\right|\right)^{\zeta}\right|}{\frac{1}{K}\left|ln\left(|T_W|\right)^{\zeta}\right|_{\sigma}}\right) \tag{7}$$

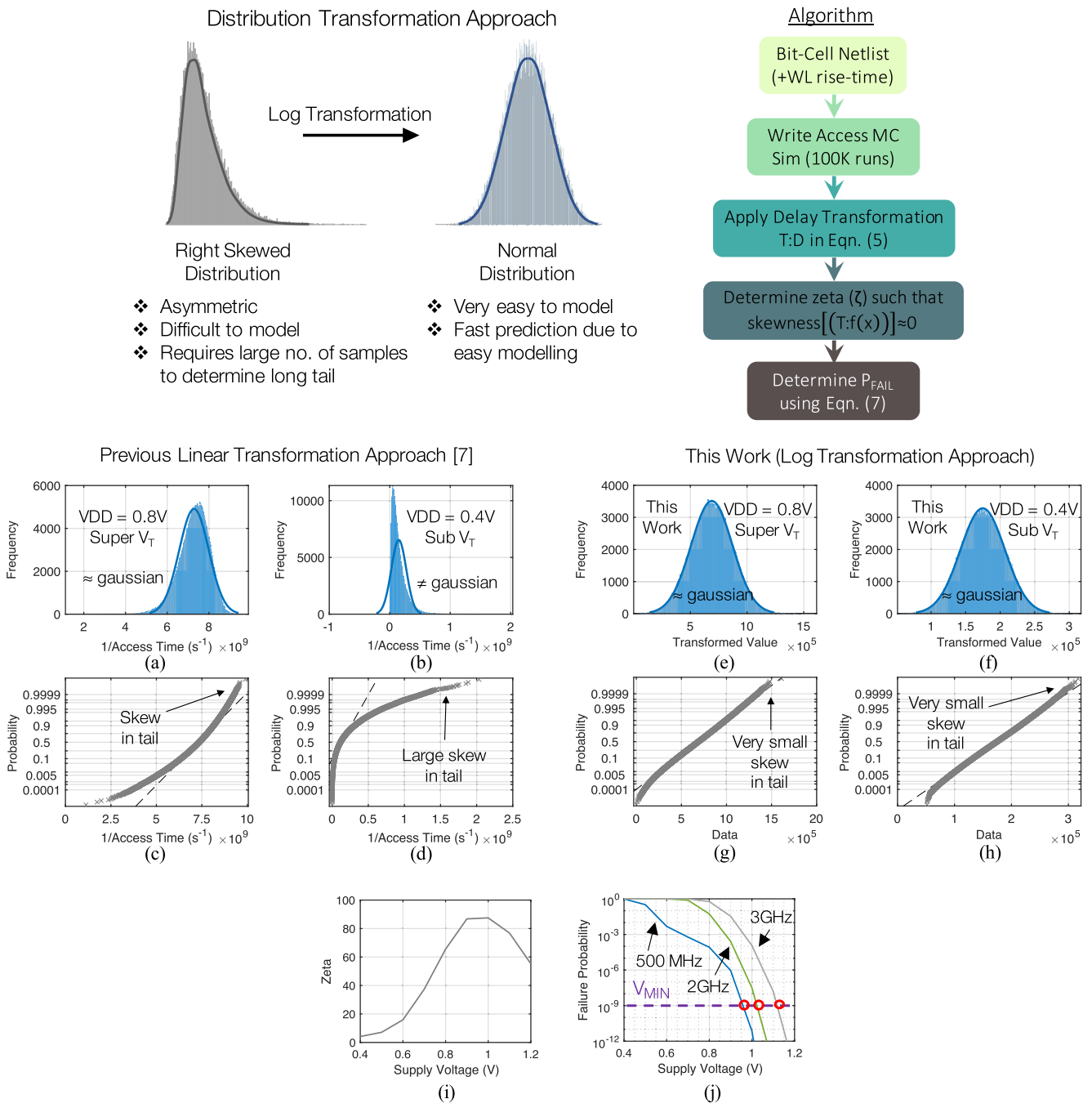

Fig. 2. The log-transformation approach can be used to transform the skewed distribution of the write-access operation into a normal distribution to easily calculate the failure probability. The histograms of the linear inverse transformation [7] distribution are shown in (a) super-threshold & (b) sub-threshold region of operation, with the corresponding probability plots in (c) & (d). The histogram of the proposed transformation model is shown in (e) & (f), with the corresponding probability plots in (g) & (h). (i) Value of zeta variable computed using 100K MC sims. (j) Trend of write-access failure probability and corresponding V_{MIN} with varying supply voltage.

The transformed distribution, along with the linear transformation is shown in Fig. 2 (e). While the inverse transformation is a good fit for super-threshold region, it fails to work in the subthreshold region. Whereas the proposed transformation modelling allows linearity (i.e., Gaussian nature) and very little skew in both regions of operation. The failure probability was calculated using (7) for different access frequencies and is shown in Fig. 2 (j). The write-access V_{MIN} is calculated when the failure probability reaches 10E-9.

Table I shows the comparison of transformation moments for various analytical methods and compares them with the ideal case. The closer the transformation moments are to the ideal case, the lower we can expect the error to be. The error in calculation and time to compute are summarized in Table II, with the process steps for evaluation in Fig. 2.

TABLE I
COMPARISON OF TRANSFORMATIONS

Region	Metric	Ideal	[7]	This Work
Super-V_T (0.8V)	Skewness	0	-0.57	-3.6E-7
	Kurtosis	3	3.62	2.84
Sub-V_T (0.4V)	Skewness	0	2.02	-3.4E-7
	Kurtosis	3	9.77	2.84

C. Write-Access Noncentral F Distribution Analytical Tail Modelling

This section describes how to analytically estimate the tail of the distribution of the write-access operation in an SRAM by fitting a noncentral F distribution to it. This is accomplished by first estimating the moments of the distribution assuming V_{TH} as the variable and then mapping them on to a noncentral F distribution. The variation in threshold voltage due to Random Dopant Fluctuations ($\sigma_{VT,RDF}$), transistor length variations ($\sigma_{VT,L}$), Random Telegraphic Noise ($\sigma_{VT,RTN}$), and other sources of variability ($\sigma_{VT,Other}$), which affect stability and performance of the cell can be modelled as [8]

$$\sigma_{V_T} = \sqrt{\sigma^2_{VT,RDF} + \sigma^2_{VT,L} + \sigma^2_{VT,RTN} + \sigma^2_{VT,Other}} \quad (8)$$

For a given technology with given minimum transistor sizing W_{MIN} and L_{MIN}, the deviation in threshold voltage ($\sigma_{V_{T_i}}$) for any transistor can be calculated by using Pelgrom's Law [9]. Advanced technologies exhibit deviation from Pelgrom's Law [10], [11], which are modelled as

$$\sigma_{V_{T_i}} = \sigma_{V_T} \times \sqrt{\frac{W_{MIN} L_{MIN}}{(W)^{\alpha} (L)^{\beta}}} \quad (9)$$

where α and β are technology constants. The moments associated with V_T can then be represented as

$$\mu_{V_T}\big|_1 = \mu_{V_T} = E(V_T) \quad (10)$$

$$\mu_{V_T}\big|_2 = \sigma^2_{V_T} = E\left(V_T - \mu_{V_T}\right)^2 \quad (11)$$

$$\mu_{V_T}\big|_3 = E\left(V_T - \mu_{V_T}\right)^3 \quad (12)$$

$$\mu_{V_T}\big|_4 = E\left(V_T - \mu_{V_T}\right)^4 \quad (13)$$

If $x_1, x_2, \ldots x_n$ are independent random variables with mean η_i and variance σ_i, then the function $y = f(x_1, x_2, \ldots x_n)$ can be expressed using the multivariable Taylor series expansion as

$$f(x_1, \ldots x_n) = f(\eta_1, \ldots \eta_n) + \sum_{i=1}^{n} \left.\frac{\partial f}{\partial x_i}\right|_{\eta_1, \ldots \eta_n} (x_i - \eta_i) + r(x_1, \ldots x_n) \quad (14)$$

where $r(x_1, x_2 \ldots x_n)$ is the higher order term. Applying this result to V_T and including the first few terms in $r(x_1, x_2 \ldots x_n)$, $f(x_1, x_2, \ldots x_n)$ can be approximated to

$$f(V_T) \cong f\left(\mu_{V_T}\right) + \left(V_T - \mu_{V_T}\right) \frac{\partial f}{\partial V_T} + \frac{1}{2}\left(V_T - \mu_{V_T}\right) \frac{\partial^2 f}{\partial V_T^2} \quad (15)$$

$$\mu_f\big|_1 = E[f(V_T)] \cong f\left(\mu_{V_T}\right) + E\left(V_T - \mu_{V_T}\right)^2 \frac{\partial^2 f}{\partial V_T^2} \quad (16)$$

Considering the third and fourth order moments of V_T, a few more terms in $r(x_1, x_2 \ldots x_n)$ can be included to evaluate the variance of f as

$$\mu_f\big|_2 = \sigma^2_{V_T} = Var(f(V_T)) = E\left[f(V_T) - \mu_f\right]^2 \cong \sigma^2_{V_T}\left(\frac{\partial f}{\partial V_T}\right)^2 - \frac{1}{4}\sigma^2_{V_T}\left(\frac{\partial^2 f}{\partial V_T^2}\right)^2 + E\left(V_T - \mu_{V_T}\right)^3 \frac{\partial f}{\partial V_T}\frac{\partial^2 f}{\partial V_T^2} \quad (17)$$

$$+ \frac{1}{4}E\left(V_T - \mu_{V_T}\right)^4 \left(\frac{\partial^2 f}{\partial V_T^2}\right)^2 \quad (18)$$

For calculating the third moment, we first calculate $E\left[f(V_T)^2\right]$ and $E\left[f(V_T)^3\right]$ as

$$E\left[f(V_T)^2\right] = E\left[\left\{f\left(\mu_{V_T}\right) + \left(V_T - \mu_{V_T}\right)\frac{\partial f}{\partial V_T}\right\}^2\right] \quad (19)$$

$$= f\left(\mu_{V_T}\right)^2 + E\left(V_T - \mu_{V_T}\right)^2 \left(\frac{\partial f}{\partial V_T}\right)^2 \quad (20)$$

$$E\left[f(V_T)^3\right] = E\left[\left\{f\left(\mu_{V_T}\right) + \left(V_T - \mu_{V_T}\right)\frac{\partial f}{\partial V_T}\right\}^3\right] \quad (21)$$

$$= f\left(\mu_{V_T}\right)^3 + E\left(V_T - \mu_{V_T}\right)^3 \left(\frac{\partial f}{\partial V_T}\right)^3 + 3f\left(\mu_{V_T}\right) E\left(V_T - \mu_{V_T}\right)^2 \left(\frac{\partial f}{\partial V_T}\right)^2 \quad (22)$$

Using eqn. (19) and (21), the third moment can be calculated as

$$\mu_f\big|_3 = E\left[f(V_T) - \mu_f\right]^3 = E\left[f(V_T)^3\right] - 3E[f(V_T)]\,E\left[f(V_T)^2\right] + 2E[f(V_T)]^3 \quad (23)$$

$$\mu_f\big|_3 = 3f\left(\mu_{V_T}\right)^2 E\left(V_T - \mu_{V_T}\right)^2 \frac{\partial^2 f}{\partial V_T^2} + E\left(V_T - \mu_{V_T}\right)^3 \left(\frac{\partial f}{\partial V_T}\right)^3 + \left[E\left(V_T - \mu_{V_T}\right)^2\right]^2 \frac{\partial^2 f}{\partial V_T^2} \times \left[6f\left(\mu_{V_T}\right)\frac{\partial^2 f}{\partial V_T^2} + 2E\left(V_T - \mu_{V_T}\right)^2 \left(\frac{\partial^2 f}{\partial V_T^2}\right)^2 - 3\left(\frac{\partial f}{\partial V_T}\right)^2\right] \quad (24)$$

Applying these results to the six transistors of the SRAM cell, we can rewrite the moments as

$$\mu_f\big|_1 = f\left(\mu_{V_T}\right) + \frac{1}{2}\sum_{i=1}^{6}\left(\frac{\partial^2 f}{\partial V_T^2}\right)\sigma^2_{V_{T_i}} \quad (25)$$

TABLE II
SUMMARY OF ERROR IN WRITE-ACCESS V_{MIN} AND EVALUATION TIME

Method	Analytical	Error (Super-V_{TH})	Error (Sub-V_{TH})	Time to Evaluate
MC sim. (6 sigma)	No	-	-	Months*
MC sim. (1M runs)	No	-	-	10 hrs.
Statistical Blockade [12]	No	<1%	<1%	60 hrs.
Sensitivity Analysis (SA) [6]	Yes	4.5%	6.83%	32 min.
Linear Transformation (LT) [7]	Yes	<2%	<15%	1 hr.
This Work (Log Transformation)	Yes	<2%	<4%	1 hr.
This Work (Noncentral F-Dist.)	Yes	<3%	<7%	10 min.
This Work (Sensitivity Analysis)	Yes	<3%	<7%	23 min.

* Estimated from [12]

$$\mu_f|_2 = \sum_{i=1}^{6}\left[\left(\frac{\partial f}{\partial V_{T_i}}\right)\sigma_{V_{T_i}}\right]^2 - \frac{1}{4}\sum_{i=1}^{6}\left[\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)\sigma_{V_{T_i}}\right]^2 + \mu_{V_T}|_3 \sum_{i=1}^{6}\left[\left(\frac{\partial f}{\partial V_{T_i}}\right)\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)\right] + \frac{1}{4}\mu_{V_T}|_4 \sum_{i=1}^{6}\left[\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)^2\right] \quad (26)$$

$$\mu_f|_3 = 3f\left(\mu_{V_T}\right)^2 \sum_{i=1}^{6}\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)\sigma_{V_{T_i}}^2 + \mu_{V_T}|_3 \sum_{i=1}^{6}\left(\frac{\partial f}{\partial V_{T_i}}\right)^3 + 6f\left(\mu_{V_T}\right)\sum_{i=1}^{6}\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)^2\left\{\sigma_{V_{T_i}}^2\right\}^2 + 2\sum_{i=1}^{6}\left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)^3\left\{\sigma_{V_{T_i}}^2\right\}^3 - 3\sum_{i=1}^{6}\left(\frac{\partial f}{\partial V_{T_i}}\right)^2 \times \left(\frac{\partial^2 f}{\partial V_{T_i}^2}\right)\left\{\sigma_{V_{T_i}}^2\right\}^2 \quad (27)$$

The moments are evaluated using the differential coefficients extracted from Fig. 3 (a)-(d) by curve fitting a polynomial whose degree matches the order of the differential coefficient. These are then be used to calculate the parameters of the F distribution (n_1, n_2, λ) by equating (25) to (27) with (A3) to (A5) respectively. Solving these, we get the values of n_1, n_2, λ, which are then used to obtain a noncentral F distribution $ncf\,(n_1, n_2, \lambda)$.

Representing the skewness and kurtosis values of this distribution as $skew\,(ncf)$ and $kurt\,(ncf)$ respectively, the final noncentral F distribution for the write-access operation can be represented by the following moments.

$$Write - Access = \left\{\mu_f|_1,\ \mu_f|_2, skew\,(ncf), kurt\,(ncf)\right\} \quad (28)$$

The distribution has been evaluated using eqn. (28) in both sub-threshold and super-threshold regions of operation using a commercial 65nm bulk technology and shown in Fig. 3 (e) and (f) respectively. The distribution matches well in super-threshold region with very low mismatch in probabilities. The means of the distributions don't match exactly in the subthreshold region, but the tails match approximately, with a difference of ~4E-5 in probabilities as shown in Fig. 3 (d). The error in evaluation and the time to compute are summarized in Table II. The process steps for performing the analysis are shown in Fig. 3.

D. Write-Access Modelling in Advanced Technologies

In comparison to planar bulk technologies, FinFET devices provide higher performance with improved short-channel effects, subthreshold slope, drive current, and mismatch [13]. However, the implementation and development of these advanced devices involves major technical challenges, including an increase in V_T variations in sub-30nm process technologies due to LER, RDF, and work function variations (WFV) [14], [15]. These can seriously degrade the V_T mismatch in various circuit blocks of the Integrated Circuit (IC). Additionally, it has been estimated that the magnitude of WFV-induced V_T variations is larger than that induced by either LER or RDF in sub-30-nm process technology [16], [17].

In this work, to assess the dynamic write-access operation in FinFET based SRAMs, we extend sensitivity analysis in [6] to observe the effect of V_T variations due to WFV. We perform the simulations on a commercial 12nm FinFET technology. The SRAM bit-cell transistor fins are sized as 1:1:2 for pull-up, access, and pull-down respectively. The V_T for both nMOS and pMOS devices is modulated by changing the metal gate work function (PHIG) in the BSIM-CMG model. With change in V_T, the dynamic write-access time is calculated and used to generate access time (T_i) versus V_T curves corresponding to each i^{th} transistor in the bit-cell [6]. A third-degree polynomial is fit to each curve as

$$T_i = aV_T^3 + bV_T^2 + cV_T + d \quad (29)$$

The offset write-access time (T_{OFFSET_i}) for each transistor is then calculated by subtracting the nominal write-access time $(T_{nominal})$ from T_i. As opposed to bulk planar technologies, the V_T variable is not explicitly described in the device model. As such, information about the V_T cannot be directly extracted from the model file and extrapolated using Pelgrom's Law. In this work, an initial simulation of 100K MC runs is performed for all the devices in the bit-cell to generate the V_T distributions. The V_T data samples are then plugged into Eqn. (29), to calculate the dynamic write-access time as

$$T_{WRITE-ACCESS} = T_{nominal} + \sum_{i=1}^{n} T_{OFFSET_i} \quad (30)$$

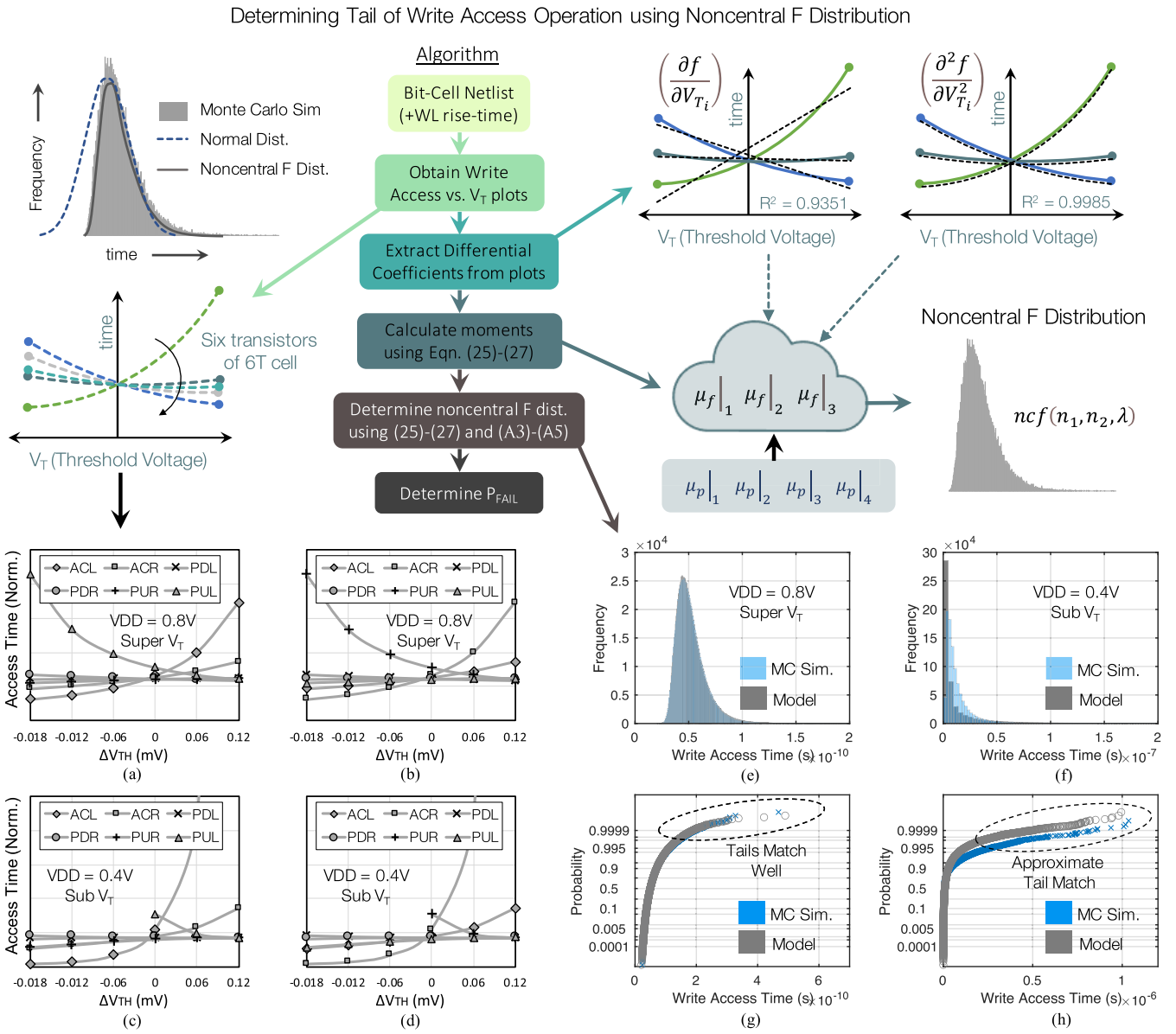

Fig. 3. Determining write-access failure probability by modelling it using noncentral F distribution. Normalized Write Access-Time variation with respect to change in V_T for each transistor in super-threshold region (a) when $X = 1$ (b) $X = 0$ and in subthreshold region (c) when $X = 1$ (d) $X = 0$. Fig. 6. Comparison of write-access operation distributions using proposed modelling technique and MC sims in (e) Super-threshold region (f) Sub-threshold region of operation, with the corresponding probability plots in (g) & (h).

where n is the number of transistors in the bit-cell. This calculation in (30) is repeated N times depending on the desired sample size to generate the final dynamic write-access operation distribution. The distribution has been evaluated (1M iterations) using Eqn. (30) and compared with MC sims in Fig. 4. As seen in Fig. 4, the proposed analysis matches well with MC sims, including in the tail. The process steps for performing the analysis are shown in Fig. 4, with the runtime and error in evaluation summarized in Table II.

E. Contention Limited Write-Access Failures

In this section, the method to calculate the contention-limited write-access failures is presented. These failures are static in nature and occur irrespective of word-line pulse width. Since these failures occur due to insufficiency of write margin (WM) in the bit-cell, they can be modelled using the static write margin equations as

$$\mu_{WM} = WM + \sum_{i=1}^{n}\left(\frac{1}{2}\frac{\partial^2 WM}{\partial V_{t_i}^2}\right)\sigma_{V_{t_i}}^2 + \sum_{k=1}^{n}\sum_{\substack{i=1 \\ i\neq k}}^{n}\frac{\partial^2 WM}{\partial V_{t_i}\partial V_{t_k}} r\,(i,k)\,\sigma_{V_{t_i}}\sigma_{V_{t_k}} \tag{31}$$

$$\sigma_{WM}^2 = \sum_{i=1}^{n}\left[\left(\frac{\partial WM}{\partial V_{t_i}}\right)\sigma_{V_{t_i}}\right]^2$$

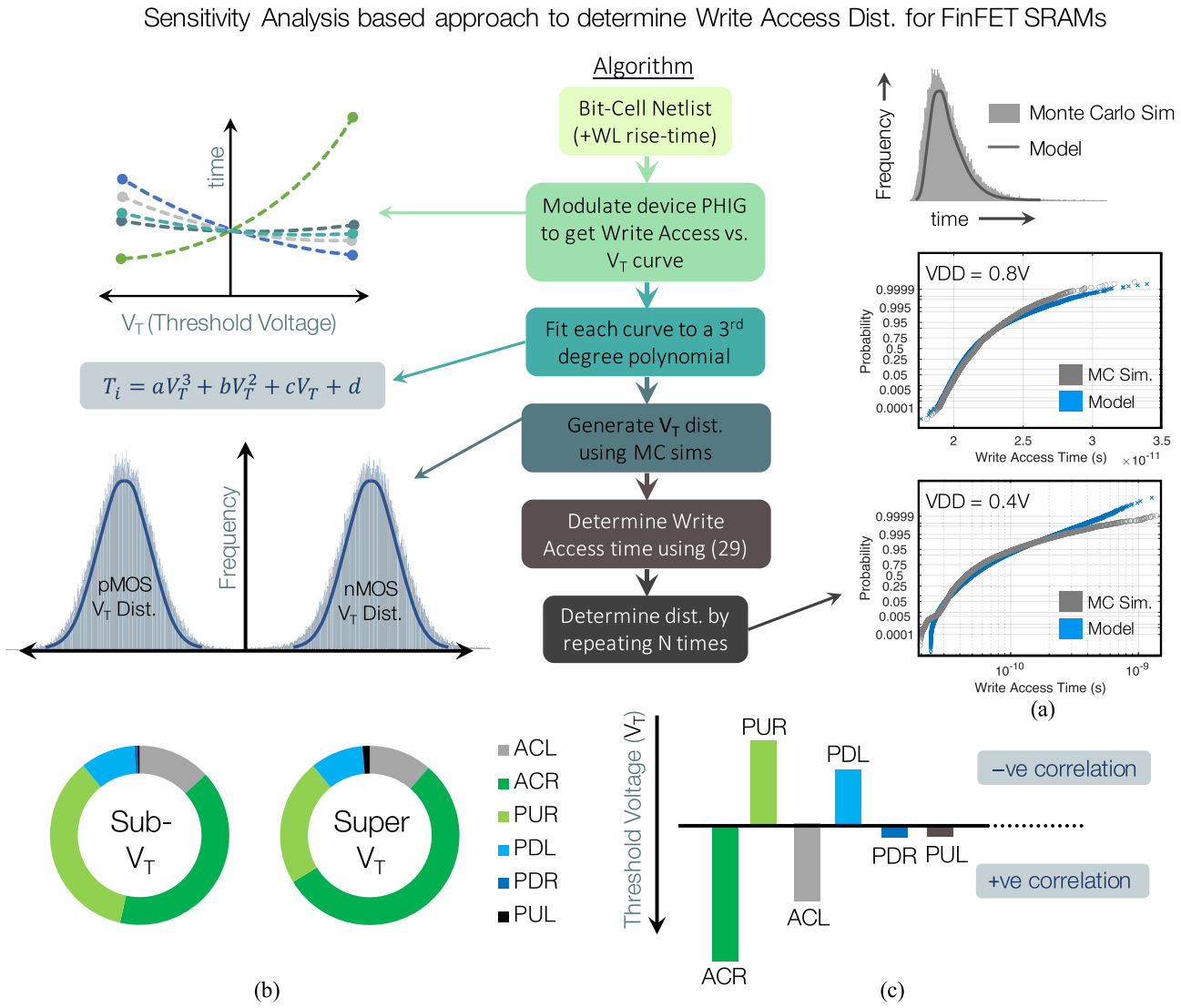

Fig. 4. (a) Comparison of write-access operation distributions using sensitivity analysis and MC sims in 12nm FinFET process (1M iterations) (b) The relative impact and contribution of each transistor in the 6T bit-cell on the write access operation performance in super-threshold and subthreshold region of operation. The contribution of the pull-up transistor on the hold 'one' (and write 'zero') side of the bit-cell increases in subthreshold region. (c) Correlation between threshold voltage of transistor and write-access time for each transistor in the 6T bit-cell.

$$+2\sum_{k=1}^{n}\sum_{\substack{i=1\\ i\neq k}}^{n}\left(\frac{\partial WM}{\partial V_{t_i}}\right)\left(\frac{\partial WM}{\partial V_{t_k}}\right) r\,(i,k)\,\sigma_{V_{t_i}}\sigma_{V_{t_k}} \quad (32)$$

Here, $r\,(i,k)$ is the correlation coefficient and V_{t_i} represents the threshold voltage of the i^{th} transistor. The total number of contention-limited write-access failures are then calculated using the cumulative distribution function as

$$\begin{aligned} Total\ Failures &= \frac{N}{2}\left[1+erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \\ &= \frac{N}{2}\left[1+erf\left(\frac{-\mu}{\sigma\sqrt{2}}\right)\right] \end{aligned} \quad (33)$$

where N is the total number of samples. The write margin distribution was computed for a 6T bit-cell with fin ratio 1:1:2 at 0.4V using both MC simulations and model equations and plotted in Fig. 5 (a). As seen in Fig. (5), the distribution fits the Gaussian curve and both MC and model distributions match well. The area under the curve from negative infinity to zero write margin represents the total number of samples which will fail statically. In Fig. 5 (b), the write-access distribution was plotted for the same bit-cell using the model presented in Fig. 4 and compared against MC simulations. The contention-limited write-access failures calculated using (33) were also added to the same plot to represent the entirety of the write-access distribution. As seen in Fig. 5 (b), the model and MC simulations match well even at low operating voltages (0.4V), with about 1.2% error in the number of contention-limited write-access failures.

III. Write Assist Circuits in Advanced Technologies

As SRAMs have continued to scale, they have required the adoption of assist circuits to ensure scaling. This adoption

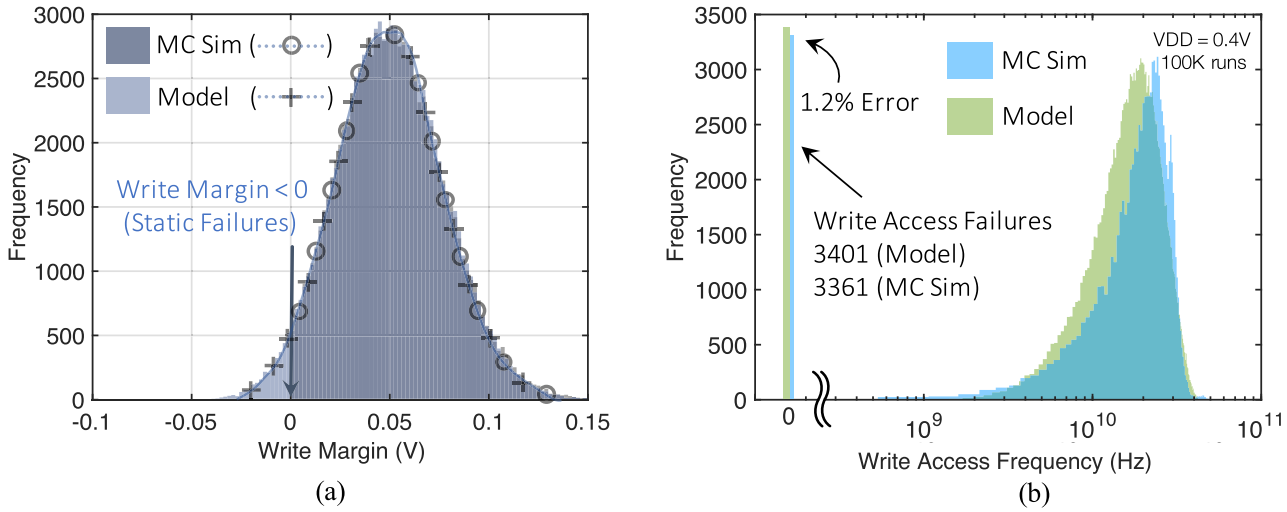

Fig. 5. Comparison of (a) Static write margin distribution using model and MC sims. at $V_{DD} = 0.4V$ (100K iterations) (b) write-access distribution with respect to frequency at $V_{DD} = 0.4V$. Results show about 1.2% error in computation of contention-limited write-access failures.

TABLE III
SUMMARY OF 6σ WRITE-MARGIN V_{MIN} USING VARIOUS ASSIST TECHNIQUES

Bit-Cell Fin Ratio	No Assist (mV)		V_{DD} Collapse (mV)		Negative BL (mV)		WL Boost (mV)	
	MC	Analytical	MC	Analytical	MC	Analytical	MC	Analytical
1:1:1	794.5	808.8	676.6	687.4	690.0	700.3	664.2	674.9
1:1:2	829.1	838.2	699.1	706.1	716.8	724.7	692.3	701.3
1:2:2	674.4	680.5	593.1	599.0	550.6	557.1	565.5	571.6

MC simulation (100K runs) is performed with 20% assists.

becomes necessary because the quantized nature of transistor fins leaves little room for V_{MIN} or performance adjustment, lest the designer gives up array area efficiency. Therefore, it has become imperative to determine which assist technique is most suited for the maximal improvement in V_{MIN} or performance depending on the application use case. In this section, we investigate the effect of various write-assist techniques on the dynamic write performance of 6T bit-cells (1:1:2 fin ratio) from $V_{DD} = 0.4V$ to $V_{DD} = 0.8V$ using the 12nm technology. The nominal supply voltage is 0.8V and the threshold voltages for nFET and pFET are about 350mV and -350mV respectively. The analyses are performed using MC simulations considering assist circuit voltages as a percentage of the supply voltage (10% and 20%) and compared against MC simulations with no assists in Fig. 6. The assist circuits include Word-Line Boosting (WLB), V_{DD} collapse (VDDU), and Negative Bit-Line (NBL) [18]. The Write-Margin [19]–[21] has also been calculated and summarized for varying bit-cell fin ratios and write-assist techniques in Table III.

Simulations show that the WLB assist brings about the most in V_{MIN} reduction in bit-cells with fin ratios 1:1:1 and 1:1:2, whereas the NBL assist is best at reducing the V_{MIN} in the 1:2:2 bit-cell. The WLB assist has the greatest overall impact on write-access performance and the VDDU and NBL assist techniques have the modest impact across operating voltages and amount of assist voltage applied. The distribution is long tailed when no assist is applied, meaning that the outliers take an exceptionally long time to write. The outliers can differ by as much as two orders of magnitude as shown in Fig. 6 (c) and (d). The assist circuits greatly impact these outliers and transform the distribution to more gaussian-like in super-threshold region of operation. In near-threshold region of operation, the tail is impacted even more so, thereby greatly shortening the tail. Even though the overall performance of the WLB technique is the best, it has the worst row half-select stability issue because it exacerbates the static read noise margin problem during a pseudo-read operation in half selected cells in interleaved memories. While a careful circuit implementation can be used to somewhat mitigate this effect, the trade-off between the read-stability and write-ability remains delicate. The NBL technique doesn't impact the row half-select stability, but it impacts the stability of column half selected cells. However, the work in [18] suggests that the probability of half-select stability issues is very unlikely for smaller (<30%) NBL assist values. As such, NBL assist can offer the best overall trade-off between write-access performance improvement and half-select stability issues. The VDDU technique can impact both row and column half-select stability of other cells depending on the designer's implementation. It is up to the designer to decide and trade-off various read and write assist techniques to balance both performance and stability across the entire array of cells.

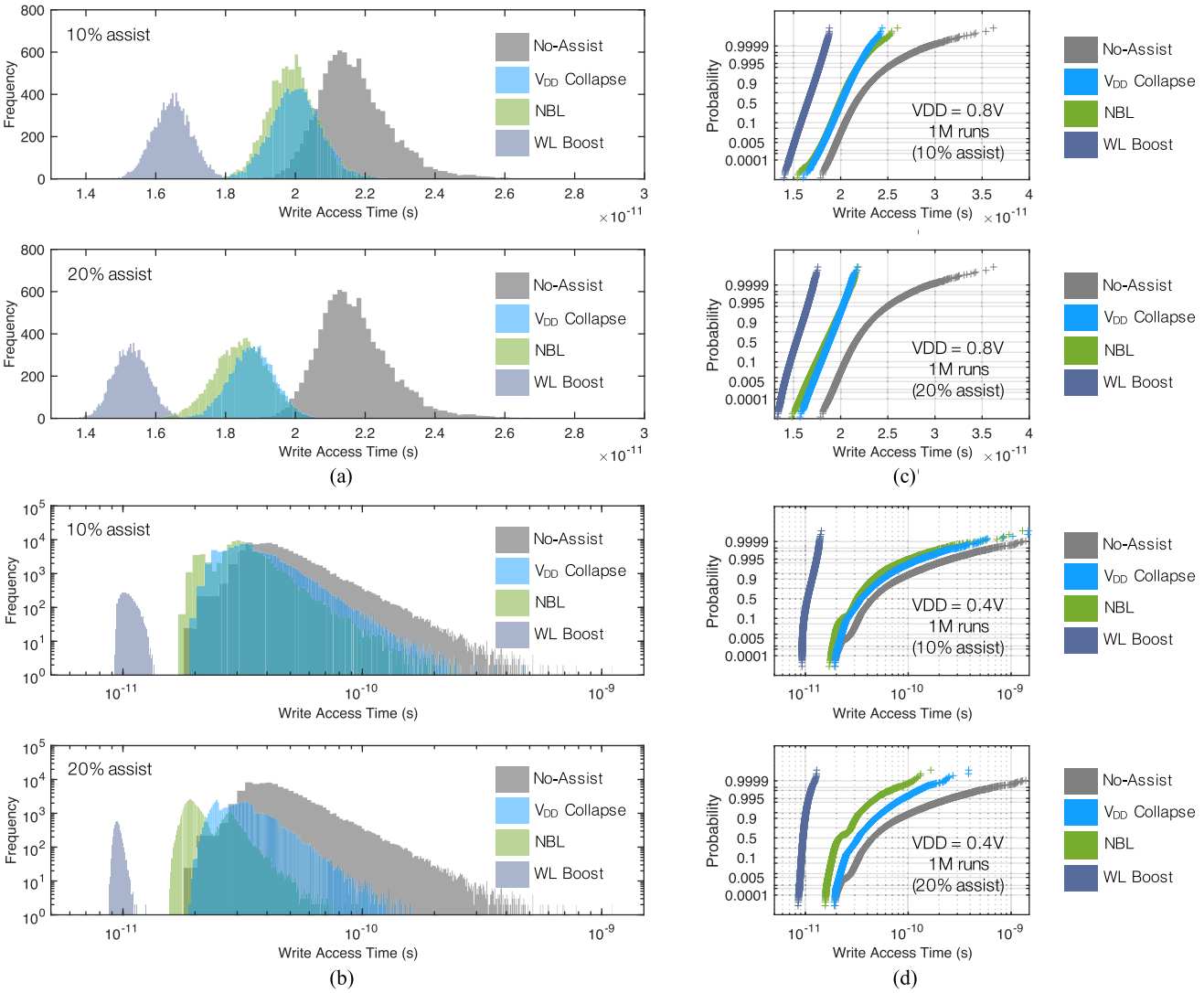

Fig. 6. Performance comparison of write-access operation histograms with various write assist techniques at (a) $V_{DD} = 0.8V$ (b) $V_{DD} = 0.4V$ and write-access operation probability plots at (c) $V_{DD} = 0.8V$ (d) $V_{DD} = 0.4V$ in 12nm FinFET process.

IV. Conclusion

In this work, we showed how the write-access operation in an SRAM has a skewed and long tailed distribution which sets the failure threshold for the dynamic write operation. Analytical approaches which use transformation and sensitivity analysis are viable methodologies to evaluate the tail. While previous analytical approaches are successfully able to trade off speed and accuracy in comparison to MC simulations, they are only able to do so in super-threshold region of operation and introduce large errors in subthreshold region. We presented an analytical methodology to estimate the tail of the write-access operation using a log transformation model which works well in all regions of operation.

We also provided an analytical solution to the tail of the write-access operation distribution which is known to

$$p(x) = e^{-\lambda/2+(\lambda n_1 x)/[2(n_2+n_1 x)]} n_1^{n_1/2} n_2^{n_2/2} x^{n_1/2-1} (n_2+n_1 x)^{-(n_1+n_2)/2} \times \frac{\Gamma\left(\frac{1}{2}n_1\right)\Gamma\left(1+\frac{1}{2}n_2\right) L_{n_2/2}^{n_1/2-1}\left(-\frac{\lambda n_1 x}{2(n_2+n_1 x)}\right)}{B\left(\frac{1}{2}n_1, \frac{1}{2}n_2\right)\Gamma\left(\frac{1}{2}(n_1+n_2)\right)}$$

$$= \frac{1}{B\left(\frac{1}{2}n_1, \frac{1}{2}n_2\right)} e^{-\lambda/2} n_1^{n_1/2} n_2^{n_2/2} x^{n_1/2-1} \tag{A1}$$

$$(n_2+n_1 x)^{-(n_1+n_2)/2} {}_1F_1\left(\frac{1}{2}(n_2+n_1); \frac{1}{2}n_1; \frac{\lambda n_1 x}{2(n_2+n_1 x)}\right) \tag{A2}$$

$$\mu_p\big|_1 = \mu = \frac{n_2\left(\lambda + n_1\right)}{n_1\left(n_2 - 2\right)} \tag{A3}$$

$$\mu_p\big|_2 = \sigma^2 = \frac{2n_2^2\left[\lambda^2 + 2\left(n_1 + n_2 - 2\right)\lambda + n_1\left(n_1 + n_2 - 2\right)\right]}{n_1^2\left(n_2 - 2\right)^2\left(n_2 - 4\right)} \tag{A4}$$

$$\mu_p\big|_3 = \frac{8n_2^3\left[\begin{array}{l} 2\lambda^3 + 6\left(n_1 + n_2 - 2\right)\lambda^2 \\ +3\left(n_1 + n_2 - 2\right)\left(2n_1 + n_2 - 2\right)\lambda \\ +n_1\left(n_1 + n_2 - 2\right)\left(2n_1 + n_2 - 2\right) \end{array}\right]}{n_1^3\left(n_2 - 2\right)^3\left(n_2 - 4\right)\left(n_2 - 6\right)} \tag{A5}$$

$$\mu_p\big|_4 = \frac{12n_2^4\left[\begin{array}{l} \left(n_2 + 10\right)\lambda^4 + 4\left(n_2 + 10\right)\left(n_1 + n_2 - 2\right)\lambda^3 \\ +2\left(3n_1 + 2n_2 - 4\right)\left(n_2 + 10\right)\left(n_1 + n_2 - 2\right)\lambda^2 \\ +8\left(n_1 + n_2 - 2\right)\left(3n_1 + n_2 - 2\right)\left(2n_1 + n_2 - 2\right)\lambda \\ +4\left(n_1 + n_2 - 2\right)^2\left(n_2 - 2\right)\left(n_1 + 2\right)\lambda \\ +2n_1\left(n_1 + n_2 - 2\right)\left(3n_1 + n_2 - 2\right)\left(2n_1 + n_2 - 2\right) \\ +n_1\left(n_1 + n_2 - 2\right)^2\left(n_2 - 2\right)\left(n_1 + 2\right) \end{array}\right]}{n_1^4\left(n_2 - 2\right)^4\left(n_2 - 4\right)\left(n_2 - 6\right)\left(n_2 - 8\right)} \tag{A6}$$

closely resemble the noncentral F distribution. Furthermore, we presented a sensitivity analysis-based evaluation method to determine the write-access operation distribution in advanced FinFET technologies.

The dynamic write-access failures can also include failures which are independent of pulse-width and are caused by insufficiency of static write margin. Therefore, we presented a methodology to extend all these methods to include the evaluation of contention-limited write-access failures as well.

Since in advanced technologies, assist circuits play a crucial role in enabling performance and stability, we evaluated and compared various write assist techniques for different bit-cell fin ratios and regions of operation.

APPENDIX
MOMENTS OF THE NON-CENTRAL F DISTRIBUTION

Let $(X_1, X_2, \ldots, X_i, \ldots, X_k)$ be k independent, normally distributed with means μ_i and unit variances. Then the random variable $\sum_{i=1}^{k} X_i^2$ is distributed according to the noncentral chi-squared (χ^2) distribution. It has two parameters: k specifies the number of degrees of freedom and λ (also called the non-centrality parameter) is related to the mean of the random variables X_i by $\lambda = \sum_{i=1}^{k} \mu_i^2$. Then, if $X : \chi_{n_1}^2(\lambda_1)$ and $Y : \chi_{n_2}^2(\lambda_2)$ are two independently distributed noncentral chi-squared variables with n_1 and n_2 degrees of freedom respectively, then the doubly noncentral F distribution can be defined as $F = \frac{\chi_{n_1}^2/n_1}{\chi_{n_2}^2/n_2}$ [22]. With $\lambda_1 \neq 0$, $\lambda_2 = 0$, the distribution is called the singly noncentral F distribution and its probability density function is defined as (A1) and (A2), as shown at the bottom of the previous page where $\Gamma(z)$ is the Gamma function, $B(\alpha, \beta)$ is the Beta function, $L_m^n(z)$ is a generalized Laguerre polynomial, and ${}_pF_q$ is a Hypergeometric Function. The first few central moments are then evaluated as (A3)–(A6), as shown at the top of the page.

REFERENCES

[1] B. H. Calhoun and A. P. 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.

[2] A. Sheikholeslami, "Process variation and Pelgrom's law," *IEEE Solid-State Circuits Mag.*, vol. 7, no. 1, pp. 8–9, Feb. 2015.

[3] S. Gupta and B. H. Calhoun, "Dynamic read VMIN and yield estimation for nanoscale SRAMs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 3, pp. 1171–1182, Mar. 2021.

[4] S. Gupta, K. Gupta, and N. Pandey, "Pentavariate V_{min} analysis of a subthreshold 10T SRAM bit cell with variation tolerant write and divided bit-line read," in *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 10, pp. 3326–3337, Oct. 2018.

[5] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 12, pp. 1859–1880, Dec. 2005.

[6] J. Boley, V. Chandra, R. Aitken, and B. Calhoun, "Leveraging sensitivity analysis for fast, accurate estimation of SRAM dynamic write VMIN," in *Proc. DATE*, Mar. 2013, pp. 1819–1824.

[7] K. Agarwal and S. Nassif, "Statistical analysis of SRAM cell stability," in *Proc. 43rd Annu. Conf. Design Autom.*, 2006, pp. 57–62.

[8] M. H. Abu-Rahma and M. Anis, "A statistical design-oriented delay variation model accounting for within-die variations," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 27, no. 11, pp. 1983–1995, Nov. 2008.

[9] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1440, Oct. 1989.

[10] V. Wang, K. Agarwal, S. R. Nassif, K. J. Nowka, and D. Markovic, "A simplified design model for random process variability," *IEEE Trans. Semicond. Manuf.*, vol. 22, no. 1, pp. 12–21, Feb. 2009.

[11] C. Couso *et al.*, "Dependence of MOSFETs threshold voltage variability on channel dimensions," in *Proc. Joint Int. EUROSOI Workshop Int. Conf. Ultimate Integr. Silicon (EUROSOI-ULIS)*, Apr. 2017, pp. 87–90.

[12] J. Wang, A. Singhee, R. A. Rutenbar, and B. H. Calhoun, "Two fast methods for estimating the minimum standby supply voltage for large SRAMs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 29, no. 12, pp. 1908–1920, Dec. 2010.

[13] D. Burnett, S. Parihar, H. Ramamurthy, and S. Balasubramanian, "FinFET SRAM design challenges," in *Proc. IEEE Int. Conf. Design Technol.*, Austin, TX, USA, May 2014, pp. 1–4.

[14] A. Asenov, "Simulation of statistical variability in nano MOSFETs," in *Proc. IEEE Symp. VLSI*, Jun. 2007, pp. 86–87.

[15] H. Nam and C. Shin, "Study of high-k/metal-gate work function variation in FinFET: The modified RGG concept," *IEEE Electron Device Lett.*, vol. 34, no. 12, pp. 1560–1562, Dec. 2013.

[16] T. Matsukawa *et al.*, "Comprehensive analysis of variability sources of FinFET characteristics," in *Proc. Symp. VLSI Technol.*, Honolulu, HI, USA, 2009, pp. 118–119.
[17] H. F. Dadgour, K. Endo, V. K. De, and K. Banerjee, "Grain-orientation induced work function variation in nanoscale metal-gate transistors—Part I: Modeling, analysis, and experimental validation," *IEEE Trans. Electron Devices*, vol. 57, no. 10, pp. 2504–2514, Oct. 2010.
[18] B. Zimmer *et al.*, "SRAM assist techniques for operation in a wide voltage range in 28-nm CMOS," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 853–857, Dec. 2012.
[19] H. Makino *et al.*, "Reexamination of SRAM cell write margin definitions in view of predicting the distribution," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 58, no. 4, pp. 230–234, Apr. 2011.
[20] S. Gupta, K. Gupta, and N. Pandey, "A 32-nm subthreshold 7T SRAM bit cell with read assist," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 12, pp. 3473–3483, Dec. 2017.
[21] S. Gupta, K. Gupta, B. H. Calhoun, and N. Pandey, "Low-power near-threshold 10T SRAM bit cells with enhanced data-independent read port leakage for array augmentation in 32-nm CMOS," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 3, pp. 978–988, Mar. 2019.
[22] R. Chattamvelli, "On the doubly noncentral f distribution," *Comput. Statist. Data Anal.*, vol. 20, no. 5, pp. 481–489, Nov. 1995.

Shourya Gupta (Graduate Student Member, IEEE) was born in New Delhi, India, in 1994. He received the B.Tech. degree in electronics and communication engineering from Guru Gobind Singh Indraprastha University, New Delhi, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Virginia, Charlottesville, VA, USA.

His current research interests include the design of low-power logic and memory circuits, and circuit design automation.

Benton H. Calhoun (Fellow, IEEE) received the B.S. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2000, and the M.S. degree and the Ph.D. degree in electrical engineering from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2002 and 2006, respectively.

In January 2006, he joined the Department of Electrical and Computer Engineering, University of Virginia, where he is currently a Professor. His research has emphasized energy-efficient and sub-threshold circuit design for self-powered and batteryless wireless sensing systems. Starting from fundamental advances in sub-threshold circuits, he has expanded his work to include complete self-powered nodes for the Internet-of-Things (IoT) and body-worn applications. He is the Campus Director and the Technical Thrust Leader of the NSF Nanosystems Engineering Research Center (ERC) for Advanced Self-Powered Systems of Integrated Sensors and Technologies (ASSIST). He co-founded and is the Co-CTO of Everactive, Inc., Charlottesville, which is selling self-powered and energy-harvesting wireless sensing solutions in the industrial IoT market. He is the coauthor of *Sub-threshold Design for Ultra Low-Power Systems* (Springer, 2006) and the author of *Design Principles for Digital CMOS Integrated Circuits* (NTS Press, 2012). He has over 200 peer-reviewed publications and 22 issued U.S. patents that contribute to the field of energy-efficient circuits and systems for self-powered and energy-constrained applications. His research interests include self-powered wireless sensors for the IoT, batteryless systems, body area sensor networks, low-power digital circuit design, system-on-chip architecture and circuits for energy-constrained applications, system-driven embedded hardware/software design, wakeup receivers, energy-harvesting-power management units, sub-threshold digital circuits, sub-threshold SRAM, energy-efficient communication, power harvesting and delivery circuits, low-power mixed-signal design, and medical applications for low-energy electronics.