# Evaluating the Accuracy of SRAM Margin Simulation Through Large Scale Monte-Carlo Simulations with Accurate Compact Models

Plamen Asenov, Scott Roy, Asen Asenov
University of Glasgow,
Device Modelling Group
plamen.asenov.1@research.gla.ac.uk

Dave Reid, Campbell Millar
Gold Standard Simulations Ltd.
Scotland, UK

David New
ARM Ltd.
Cambridge, UK

***Abstract*— Statistical variability due to the discreteness and granularity of charge and matter has a large impact on SRAM performance due to its stochastic nature. In this paper we have performed 5 million SRAM dynamic write simulations with accurate compact models which capture all aspects of statistical variability and use them to benchmark the accuracy of Gaussian threshold voltage modeling strategies and a common SRAM margining technique, MPV. The results show that while MPV and Gaussian VT are proven approaches, deep into the tails of the distribution NPM simulation may present significant opportunities for improved design.**

## I. INTRODUCTION

Statistical variability, associated with the granularity of matter and discreteness of charge, has a significant impact on the performance, power and yield of SRAM. At the 20/22nm technology generation, the major sources of statistical variability include Random Discrete Dopants (RDD) [1], Line Edge Roughness (LER) [2] and Metal Gate Granularity (MGG) [3].

Because of this, it is important to include accurate statistical variability information in the SRAM design and evaluation process. Unlike digital logic circuits where timing delays within a chain typically average out, missed SRAM cell timing may render a whole SRAM block non-functional. In order to increase SRAM density, designers attempt to optimise cells until minimally T-sized transistors can be found which provide a functioning cell that gives the required yield; as the magnitude of variability is inversely proportional to device area this leaves the transistors in SRAM cells acutely sensitive to statistical variability. The huge numbers of SRAM cells in modern memory arrays provide a strong motivation to simulate SRAM performance accurately up to and beyond 5 σ. Due to its stochastic nature, statistical variability is resistant to standard yield improvement methodologies, like Supply Voltage Scaling (SVS) [4] or Adaptive Body Biasing (ABB) [5] and cannot be modelled using 'process corner' modelling approaches.

The impact of statistical variability on SRAM operation has been a hot topic since the 100nm technology generation, with multiple approaches proposed to evaluate the impact of statistical variability on the SRAM standard cell. Most of these, however, have been limited to considering only the threshold voltage of devices, modelled as a Gaussian distribution (Gaussian $V_T$); an approach known *as idealisation of statistical chaos in a single variable* [6]. This approach has been favoured due to its ease of implementation and relative ease of technology characterisation. In addition, the Gaussian $V_T$ approach also simplifies statistical analysis and enables margining approaches like Most Probable Vector (MPV) [6], which can drastically reduce simulation time. It is well known [7,8] that in order to do this, Gaussian $V_T$ captures only the 1st order effects of variability, and over estimates the correlation between threshold voltage and other device figures of merit. It is of considerable industrial interest to examine quantitatively the effects of this approximation.

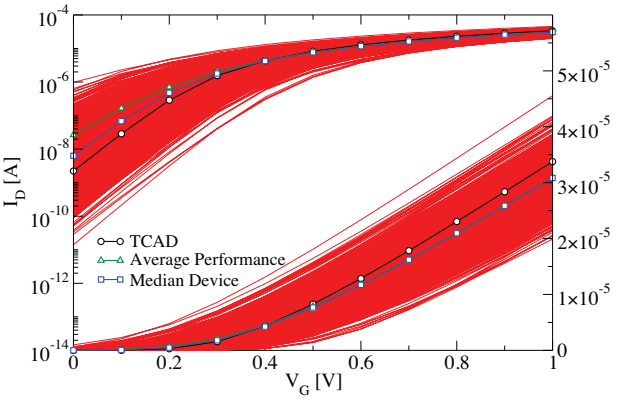


Figure 1. 10,000 simulated device at high drain bias

In this paper we perform 10,000 3D TCAD device simulations using the GSS variability simulator GARAND [9] using template Bulk-MOSFET devices designed by GSS. These devices have physical gate length of 25nm and are representative of 20/22nm bulk technology generation. The TCAD simulations include the dominant statistical variability sources at this technology node (RDD, LER and MGG) and provide accurate statistical variability information on device performance. The transfer characteristics obtained from simulation of 10,000 n-channel devices can be seen in Figure 1. Statistical compact models are then extracted and the advanced compact model generation strategy — Non-linear


978-1-4673-5736-4/13/$31.00 ©2013 IEEE

Power Method (NPM) [10] — is used for the purpose of statistical SRAM simulation, as it produces an effectively unlimited number of unique devices, which very accurately reproduce the statistical behaviour of underlying technology.

In order to evaluate the effect of statistical variability on SRAM in an industrially relevant application, this project has been carried out in collaboration with ARM Ltd. The purpose of the project was to evaluate a representative industry 'margin' simulation, (in this case the MPV method) against comprehensive Monte-Carlo simulation with compact models generated using NPM. SRAM dynamic write margin was chosen due to its paramount importance in determining word line pulse width.

## II. SCM Extraction and Generation

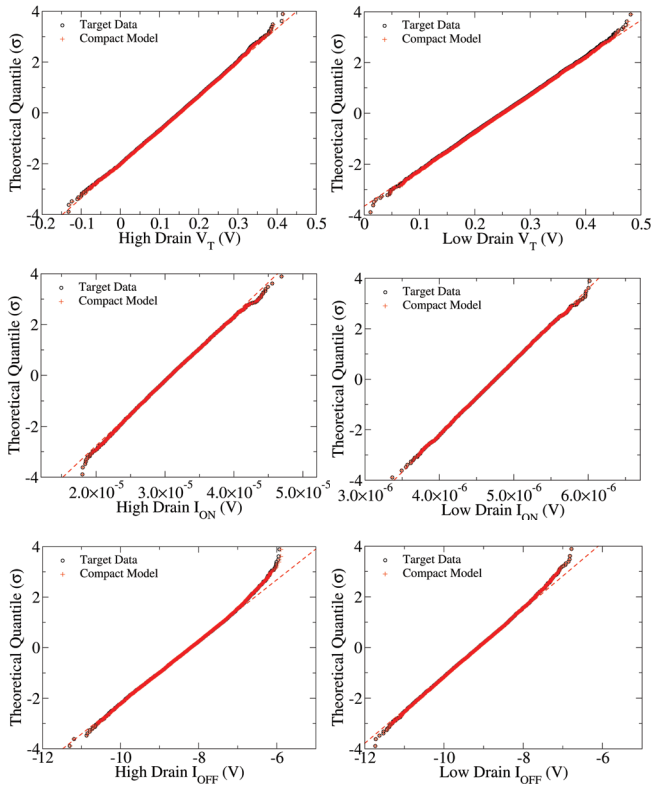


Figure 2. Comparison between the distribution of key figures of merit of the nMOSFET extracted from the TCAD simulation and from the compact model.

The statistical extraction strategy used has been previously described and is proven to accurately capture the impact of statistical variability on device performance [7], it is a two-stage process, and employs the GSS statistical compact model extractor Mystic [9]. First a nominal compact model, based on 'uniform' TCAD simulation or average device measurements containing no effects of variability, is extracted. This model is calibrated to capture gate length, body bias and temperature dependence for the underlying technology. A subset of the standard BSIM4 parameters is then selected to capture the variations in device performance due to atomistic effects. The analysis combines an in-depth knowledge of device physics, the effects of statistical variability on device performance and an intimate knowledge of the BSIM4 compact model equations and parameters. Each physical effect of stochastic variability at high and low drain voltage (namely threshold voltage, on-current, off current, subthreshold slope, drain induced barrier lowering (DIBL), mobility, vertical field dependence) is modelled by a specifically selected parameter, which is directly extracted.

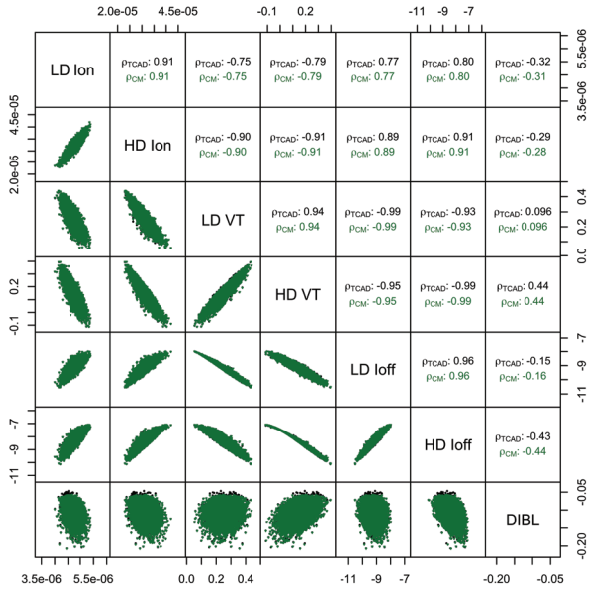


Figure 3. Comparison of correlations between extracted figures of merit for the nMOSFET from TCAD simulations and from the compact models

In the case of the 10,000 25nm n-channel ensemble, it was found that 9 carefully chosen and controlled parameters can accurately and completely capture the performance of the device ensemble. Figure 2 provides a comparison between the device figures of merit of device simulated using the extracted compact models and those extracted from atomistic simulations while Figure 3 illustrates shows that the extraction strategy developed accurately captures their correlations.

In order to generate continuous distribution of devices, which accurately reproduce the statistical behavior of the underlying transistor distributions, we use the Non-linear Power Method (NPM) calculated using the compact model parameter distributions obtained from statistical extraction.

## III. SRAM Dynamic Write Margin Simulation

A test memory system design was supplied by ARM for the purposes of this project, including addressing, word-line pulse generation, pre-charge, sense-amp and clock generation circuitry. The SRAM bit cell is representative of a 'low power' cell. Dynamic write simulations model a realistic operating condition for the whole SRAM system, and as such can be directly related to actual SRAM system operation and yield. The simulation of dynamic write margin (WM) was chosen for this comparison due to its paramount importance in defining the word line pulse width, which limits the cycletime for write operations. If this is longer than the read cycletime it will be the limiting factor for the memory cycletime.

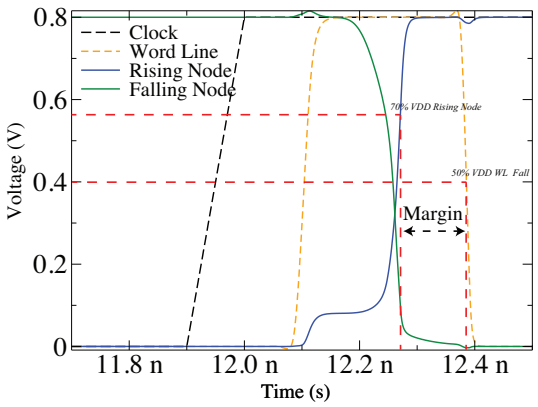


Figure 4. Dynamic Write Measurement

The dynamic write margin is defined as the time between the rising internal node reaching 70% of $V_{DD}$ before the word line falls to 50% of $V_{DD}$. The measurement is depicted in Figure 4. For the purpose of these simulations we only consider the effect of statistical variability on the SRAM cell, assuming that the word line pulse is constant. The reason behind this is that the surrounding digital circuitry is more significantly affected by process variability as it is standard digital logic and a global drift has a larger impact on digital circuit timing performance as statistical variability effects can average out. This assumption holds true for dynamic write margin simulation, as we do not use the sense amplifier, which may be highly sensitive to statistical variability, during the write operation.

WM is largely dependent on the falling node which transitions from '1' or high to '0' or low. Initially the bit lines are pre-charged to '1', then the bitline on the rising node side is actively driven. The bit-line on the side of the internal node falling from '1' to '0' is left floating and slowly discharges on to the bit-line. Due to this, the rate at which the internal node is pulled down from '1' to '0' is dominated by the performance of the pass transistor, enabled by the word line, and the pull-up transistor, which is 'on' as the node is currently storing a '1'. The node voltage slowly discharges and starts to feedback to the pull up and pull down transistors of the opposite node. This creates a feedback loop, which slowly decreases the voltage on the falling node until the cell reaches its metastable point and the cell almost instantly changes state. At this point the falling node and equivalent bitline are quickly discharged through the pulldown transistor and the write operation is complete. The dynamic write margin can be broken down into two components: the falling node discharge time, which dominates due to its relatively slow nature and is defined by the falling node pull-up to pass transistors, and the cell inverter pair metastable point, which is defined by the cross-coupled inverter pair. In order simulate the circuit at its most likely failing point we perform all simulation and analysis a the Slow N-Fast P process corner and at T=-40°C .

### *A. MPV Analysis*

The MPV method relies on the assumption that Gaussian $V_T$ accurately captures the effect of statistical variability, and involves calculating the standard deviation of the threshold voltage of each transistor in the circuit. Offsets to the threshold voltage of each transistor are calculated based on Equation 1,

$$u_i = \frac{G_i \sigma_i^2}{\sigma M} \delta x_i \quad (1)$$

where $M$ is defined as the performance metric, $u_i = \delta x_i$ is the unit perturbation which degrades $M$ by $1\sigma$, and $G_i$ is the gradient of degradation of $M$. MPV signifies the shortest path of degradation of the metric. These offsets are applied in the direction which most degrades cell performance and allows the calculation of the most probable fail point in the circuit. We already know that two assumptions inherent to this method are potentially problematic, typically requiring industry to introduce more robust margins into the design methodology to guarantee customer yield. First, threshold voltage is Gaussian distributed to high sigma [1,11], and second, Gaussian $V_T$ accurately represents statistical variability on a device and circuit level [7,8]. These assumptions will be evaluated through large-scale Monte Carlo simulation using NPM model simulations as a reference.

## IV. Simulation Results

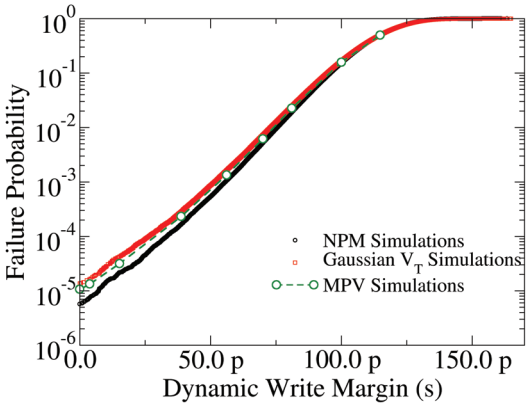


Figure 5. CDF plot of dynamic write margin obtained through Gaussian $V_T$, NPM and MPV simulation.

The results of dynamic write simulations for Gaussian $V_T$, NPM and MPV approaches are shown in Figure 5 in the form of logarithmic Empirical Cumulative Distribution Function (ECDF) plots which provide an estimate of the CDF of the input distribution. In each result set there are 5 million NPM and Gaussian $V_T$ simulations, which characterize the results to ~4.5 σ and allow analysis deep into the tails of the dynamic write distribution. These results show that, while MPV reproduces the results of Gaussian $V_T$ simulation, both of these results are pessimistic in comparison with the NPM model simulations. Full compact model simulations generated with NPM show a WM failure point for this cell design and word line pulse width which is close to 4.4 σ, compared to 4.1-4.2 σ predicted by Gaussian $V_T$ and MPV. Converting this to a parts per million failure rate, Gaussian $V_T$ / MPV simulations predict approximately 20 fails per million, while full model simulations show that the actual fail rate is closer to 4 fails per million. This indicates that while MPV and Gaussian VT are proven approaches, deep into the tails of the

distribution NPM simulation may present significant opportunities for improved design.

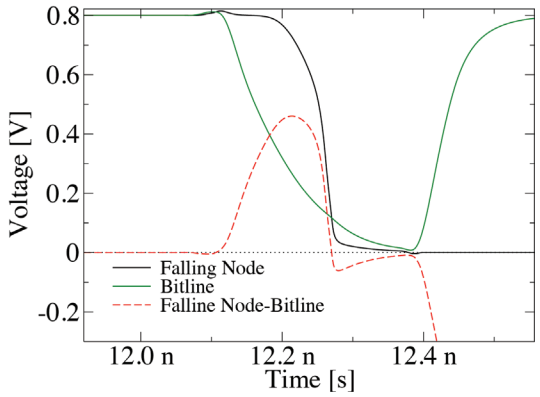


Figure 6. The bit line voltage is represented by Bitline, the falling node voltage is represented by Falling Node and the difference is represented by Falling Node-Bitline

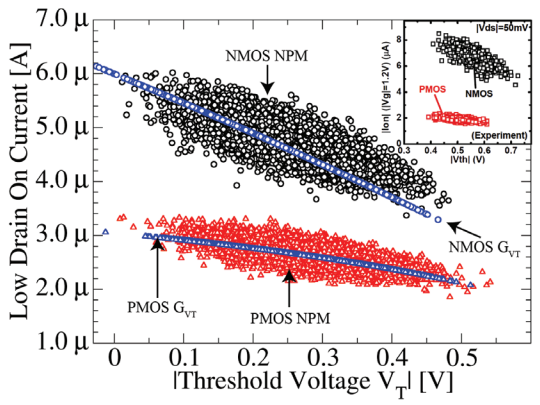


Figure 7. Generated device threshold voltage plotted against the corresponding device low drain on current, for NPM and GVT Devices, inset are device measurements at 65nm from Hiramoto et al. [8]

In order to understand the differences between Gaussian $V_T$ and NPM based full model simulations we analyse the most dominant transistor in the circuit as indicated by MPV analysis, which is the pass gate transistor on the falling SRAM cell node. Analysis shows that the worst Dynamic Write margin is obtained when this device is 'weak' i.e. when it has a high threshold voltage and low on-current, write margin is reduced. Considering the bias conditions of the pass transistor during the write operation we see that the gate bias is at $V_{DD}$ due to the word line pulse and that the source and drain are close to $V_{DD}$ due to the pre-charged bit line and internal node state of the cell. The bit line voltage then falls as it's capacitance discharges, creating a potential difference across the source and drain of the pass transistor and current begins to flow. The node voltage of the source and drain of the pass transistor in a typical device simulation with no variability is shown in Figure 6. From this we conclude that due to the fact that the peak drain bias of the critical pass gate is 0.4V, the limiting factor of its operation is therefore the low drain bias on-current of the pass device.

In order to understand the deviation of the predictions from Gaussian $V_T$ and NPM simulations we compare 10,000 n-channel pass gate devices generated using the NPM generation approach and Gaussian $V_T$. Figure 7 shows the pass gate threshold voltage compared to low drain on current for both NPM and Gaussian $V_T$ generated devices. It is clear that the Gaussian $V_T$ approach greatly overestimates the correlation between threshold voltage and low drain on-current. The use of Gaussian $V_T$ yields a correlation coefficient of 1, while NPM data shows a correlation coefficient of ~0.7. The data also shows a systematic offset between the Gaussian $V_T$ generated devices and the NPM devices, which is greatest for devices with higher threshold voltages, where Gaussian $V_T$ simulation underestimates the on current of the transistors. This result explains the pessimistic predictions of the Gaussian $V_T$ method in Dynamic Write Margin simulations, as they predict pass gates with high threshold voltages have significantly lower on-current than the actual devices.

## V. CONCLUSIONS

This paper presents the results of circuit simulations based on advanced statistical compact models extracted using a physical extraction strategy which captures the effects of statistical variability on circuit performance. NPM was then used to create compact model generators which reproduced the statistical behavior of the underlying technology. NPM based simulations were compared to Gaussian $V_T$ based simulations and the MPV margining technique on sample sizes of 5 million simulations. The results show that while MPV and Gaussian VT are proven approaches, deep into the tails of the distribution NPM simulation may present significant opportunities for improved design.

## REFERENCES


[1] D Reid and C Millar and G Roy and S Roy and A Asenov, "Analysis of threshold voltage distribution due to random dopants: A 100,000 sample 3-D simulation study", *IEEE Transactions on Electron Devices* 56 (2009), pp. 2255-2263.

[2] A. Asenov, S. Kaya and A. R. Brown, "Intrinsic Parameter Fluctuations in Decananometer MOSFETs Introduced by Gate Line Edge Roughness", *IEEE Transactions on Electron Devices* 50, 5 (2003), pp. 1254-1260.

[3] A R Brown, N Idris, J R Watling and A Asenov, "Impact of Metal Gate Granularity on Threshold Voltage Variability: A Full-Scale Three-Dimensional Statistical Simulation Study", *IEEE Electron Device Letters* 31, 11 (2010), pp. 1199-1201.

[4] E I Vatajelu and J Figueras, "Supply Voltage Reduction in SRAMs: Impact on Static Noise Margin", *Proceedings of Automation, Quality and Testing* (2008), pp. 73-78.

[5] T. Chen and S. Naffziger, "Comparison of ABB and ASV for Improving Delay and Leakage Under the Presence of Process Variation", *IEEE Transactions of VLSI Systems* 11, 5 (2003), pp. 888-899.

[6] Amith Singhee and Rob Rutenbar, *Extreme Statistics in Nanoscale Memory Design* (Springer, 2010).

[7] P. Asenov, D. Reid, S. Roy, C. Millar, A. Asenov. "An Advanced Statistical Compact Model Strategy for SRAM Simulation at Reduced VDD", *in Proc. ESSDERC 2012, Bordeaux,* pp. 205-209

[8] T. Hiramoto, *et al.*, "Direct measurement of correlation between SRAM noise margin and individual cell transistor variability by using device matrix array", *IEEE Transactions on Electron Devices* 58, 8 (2011).

[9] www.goldstandardsimulations.com

[10] C. Sohrmann, R Jancke, J Hasse, B Cheng, U Kovac and A Asenov. "A general approach for multivariate statistical MOSFET compact modeling preserving correlations", *in Proc. ESSDERC 2011* pp. 163-166

[11] T. Mizutani, A Kumar and T Hiramoto, "Measuring threshold voltage variability of 10G transistors", *in Proc. IEDM 2011,* pp. 25.2.1-25.2.4