Technique for Efficient Evaluation of SRAM Timing Failure

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1109/tvlsi.2012.2212254">http://dx.doi.org/10.1109/tvlsi.2012.2212254</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Electrical and Electronics Engineers (IEEE)</td>
</tr>
<tr>
<td>Version</td>
<td>Author's final manuscript</td>
</tr>
<tr>
<td>Accessed</td>
<td>Mon Dec 31 09:02:37 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/93898">http://hdl.handle.net/1721.1/93898</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-Noncommercial-Share Alike</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a></td>
</tr>
</tbody>
</table>
A Technique for Efficient Evaluation of SRAM Timing Failure

Masood Qazi, Student Member, IEEE, Mehul Tikekar, Student Member, IEEE, Lara Dolecek, Member, IEEE, Devavrat Shah, Member, IEEE, and Anantha P. Chandrakasan, Fellow, IEEE

Abstract—This paper presents a technique to evaluate the timing variation of SRAM. Specifically, a method called loop flattening reduces the evaluation of the timing statistics in the complex, highly structured circuit to that of a single chain of component circuits is justified. To then very quickly evaluate the timing delay of a single chain, a statistical method based on Importance Sampling augmented with targeted, high-dimensional, spherical sampling can be employed. The overall methodology has shown 650X or greater speed-up over the nominal Monte Carlo approach with 10.5% accuracy in probability. Examples based on both the large-signal and small-signal SRAM read path are discussed and a detailed comparison with state of the art accelerated statistical simulation techniques is given.

Index Terms—Cache memories, CMOS memory, random access memory, sense-amplifier, SRAM, process variation

I. INTRODUCTION

Embedded SRAM is a vital component of digital integrated circuits and often constitutes a dominant portion of chip area [1]. Therefore, the specifications of embedded SRAM have significant implications on the overall chip cost, power, performance, and yield. Shown in Fig. 1(a) is a plot of reported cell areas in fully functional SRAM macros versus the technology node for the past few years. The cell area has scaled with the scaling of the critical feature size. Fig. 1(b) plots an unconventional metric—the number of SRAM bits per mm$^2$ of silicon in high performance microprocessor chips—that reveals reduced SRAM cell area does not readily translate into increased SRAM utilization.

This discrepancy in trends is due to a number of limitations of SRAM, all related to local variation: SRAM often needs a separate, elevated power supply; excessive SRAM timing variation degrades performance; and, uncertain aging effects show up first in the highly integrated and highly sensitive SRAM cells. As the overarching goal of this work, we seek to increase the SRAM utilization by propagating the physical trend of shrinking cell area into the overall system-on-chip improvement. This goal can be achieved if designers have a way to quickly assess the impact of circuit solutions on the operating constraints (e.g., minimum VDD, frequency) to ultimately preserve the overall chip yield.

This work focuses on read access yield because it has been observed in measurements that AC fails, manifested as too slow of an access time from one or more addresses, are encountered before DC failures, manifested as the corruption of data at one or more addresses [2]. Therefore, DC stability (write and read margin) is necessary but not sufficient for yielding a memory chip. A significant degree of additional margin must be inserted to meet performance requirements.

In general, the exact distributions of the relevant SRAM performance metrics are not known. As a consequence, any statistical simulation method unavoidably resorts to numerical solvers such as SPICE. Classical approaches like the Monte Carlo method require too many iterations of such SPICE evaluations because of the circuit complexity and extremely low tolerable failure probabilities of individual components ($10^{-8}$ and below). Thus, the primary challenges to any statistical simulation method are: (a) dealing with the structural complexity of the timing delay evaluation problem, and (b) estimating timing delay statistics to a very high accuracy.

Prior Work: A lot of exciting recent work has made important progress towards the eventual goal of designing generically applicable, efficient simulation methodologies for circuit performance evaluation. To begin with, in [3–7], the authors developed efficient sampling based approaches that provide significant speedup over the Monte Carlo method. However these works do not deal with the interconnection complexity, i.e., do not address the challenge (a) stated above.

Other work has addressed the issue of structural complexity. In [8], by modeling the bitline signal and the sense amplifier offset (and the timer circuit) with Gaussian distributions, the authors proposed a linearized model for the read path. As this...
A statistical analysis of a memory critical path requires additional insight into the architecture. For now, consider a single column of 256 memory cells as in Fig. 2(b) with \( R = 256 \). When the wordline goes high at time \( t = 0 \), the memory cell develops a differential signal (voltage difference between BLT and BLC), and when the enable signal goes high at time \( t = T \), the sense amplifier resolves that differential signal to a logic-level output (voltage difference between SAT and SAC). One can represent the bitcell signal of cell \( i \) as 
\[
TX_i = T(\sigma_N X_i + \mu_X) \quad \text{and the sense amplifier offset as} \quad Y = \sigma_Y \bar{Y} \quad (\bar{Y} \text{ and } X_i \text{ are } \mathcal{N}(0,1)).
\]
The failure of this read operation is determined by the interaction of two random variables sampled at different rates. The probability \( P_f \) that this single column memory fails for a given strobe timing \( T \) is the probability that the sense amplifier offset overpowers the bitcell signal for one or more paths in the column:
\[
P_f := Pr \left( \bigcup_{i=1}^{R} \left\{ Y - TX_i > 0 \right\} \right) =: P_u, \quad (1)
\]
where \( P_u \) is the conservative union bound estimate of \( P_f \).

Because of the different rates of repetition, a proper Monte Carlo simulation on a critical path with one cell and one sense amplifier must sample variables in a nested for loop: for each sense amplifier, simulate over 256 cells and check for one or more failing paths, then sample a new sense amplifier and repeat over 256 new cell realizations and so on, as suggested in [8]. If one wishes to apply an accelerated statistical simulation to evaluate the failure of this circuit, the “for loop” sets an unfavorable lower bound on the number of simulations needed just to emulate the architecture.

We observed that the simple union-bound estimate \( P_u \) provides an accurate way to bypass this requirement. Just the path failure probability is simulated and the result is multiplied by \( R \). The estimate is guaranteed to be conservative and in practice agrees very well at low levels of failure probability. In [11] this loop flattening estimate was shown to be accurate for the small signal SRAM read path. As in [8] the bitline signal development and sense amplifier offset were parametrized by Gaussian random variables—
\[
\{ \mu_X = 1mV/ps, \sigma_X = 0.10 \times \mu_X, R = 256, \sigma_Y = 25mV \}
\]
for the expression in Eq. 1. Specifically, the loop flattening estimate was only 1.9% pessimistic in strobe timing for a modestly sized 512 kb memory (2048 memory columns) at a yield of 99%, and increased in accuracy at even lower levels of failure.

The schematic in Fig. 3(a) is the schematic tree of the large signal read path. For the case of cascaded random delays, we can also see the applicability of the loop flattening estimate. This circuit is simulated in a production quality 45 nm CMOS technology, where each shaded transistor (or gate input) exhibits local mismatch modeled as a fluctuation of its threshold voltage. Fig. 4 shows the Monte Carlo SPICE simulation result of this circuit for eight cells per local bitline (\( N_{BL} = 8 \)) and 16 local evaluation networks (\( N_{SEG} = 16 \)). In this picture, there is a delay \( Z_i \) (1 \( \leq i \leq 256 \)) associated with each of the 256 cells on the column and the probability...
of failure associated with a target delay $t$ is

$$P_f := \Pr \left( \bigcup_{i=1}^R \{ Z_i \geq t \} \right)$$

with $R = 256$. The solid curve gives the conventional, nested Monte Carlo simulation result by sampling random variables in proportion to the rate of repetition of their associated transistors. The dashed curve gives the loop flattening estimate in which a simple chain of representative transistors is simulated with all random variables sampled at the same rate. Even for this example, in which the delays are not perfectly normal and delay stages are correlated, the loop flattening technique produces a tight estimate. The single, solid black dot gives a specific example for how an Importance Sampling (IS) simulation with an appropriately chosen mean shift can evaluate the loop flattening (dashed curve) with significant speedup, consuming approximately 1.1 thousand SPICE simulations in this example. The loop flattening approximation suggests that this IS estimate in turn will match the result produced by a proper Monte Carlo simulation with nested sampling, which requires 1.7 million SPICE simulations to observe the 1% failure probability of this single column memory.

For the case of independent, random additive delays, it can be shown analytically that the loop flattening estimate converges to the true failure. Consider the simple tree in Fig. 3(b) where this time the random variables $X_i$ and $Y$ represent delays, and the overall delay of any path is $Z_i = X_i + Y$, associated with the node whose delay is $X_i$. Then, given a time, $t$, the failure probability is defined as the probability that one or more paths in the tree exceeds $t$ as in Eq. (3). The proposed loop flattening estimate treats these paths as independent at low probabilities and approximates the failure probability with the upper bound:

$$P_a := R \cdot \Pr (Z_1 \geq t) .$$

For $X_i$ and $Y$ independent normal random variables, the result in the Appendix shows that:

$$\lim_{\epsilon \to 0} \frac{P_a - P_f}{P_f} = 0 .$$

A similar argument can be developed for the sense amplifier strobe timing case.

For finite $t$, the ratio in Eq. (5) represents the overestimate of the column failure probability as a fraction of the true failure probability ($P_a = P_f(1 + \epsilon)$ with $\epsilon \to 0$). For a memory with $M$ independent columns, the overall memory failure estimate from the loop flattening approximation is:

$$1 - (1 - P_f(1+\epsilon))^M \approx 1 - (1 - MP_f - M\epsilon P_f) = MP_f(1+\epsilon).$$

Therefore, the failure of the overall memory is also overestimated by only $\epsilon \times 100\%$. For a variety of cases including numerical examples and a formal proof, the loop flattening estimate has been justified. This approximation eliminates the constraint of nested sampling which would set an unfavorable lower bound on the required number of Monte Carlo simulations. Accelerated statistical methods such as those described in Section III can be directly applied to quickly determine the value of the loop flattening estimate which produces an accurate, conservative value for the memory failure probability.

III. SPHERICAL IMPORTANCE SAMPLING COST ANALYSIS

In this section, we introduce the Monte Carlo method and then we introduce the Spherical Importance Sampling method with the purpose of significantly reducing the number of SPICE simulations required. Suppose one is given a circuit that fails with probability $p$, and wishes to identify the value of this failure with a Monte Carlo simulation that produces an estimate $\hat{p}_{MC}$. Based on a Chernoff bound, the number of required Monte Carlo simulation runs is given by:

$$N_{MC} > \frac{2 \ln \left( \frac{1}{\delta} \right)}{pe^2}$$

where $\delta = \Pr (\hat{p}_{MC} < (1 - \epsilon)p)$. In plain English, Eq. (6) says that with probability $1 - \delta$, the Monte Carlo estimate $\hat{p}_{MC}$ will not underestimate the true failure $p$ by more than $\epsilon \times 100\%$. For typical values of $\delta$ (0.01 to 0.1) and $\epsilon$ (0.05 to
0.3), this expression indicates $N_{\text{MC}} > 100/p$ to $1000/p$. To accurately estimate low failure probabilities—$10^{-6}$ to $10^{-10}$ for embedded SRAM—the standard Monte Carlo approach would require too many ($10^6$ to $10^{13}$) SPICE simulations as seen from Eq. (6).

Fortunately, Importance Sampling provides a way to speed up such a simulation, with a simple implementation (one needs only to provide an indicator of pass/fail from the SPICE simulation). This work focuses on identifying the most likely region of failure in the parameter space, relying on the well-known notion of a worst-case point [12]. The Spherical Importance Sampling approach defines a dominant point of failure as the point that minimizes the quadratic norm (in a parameter space normalized by the standard deviation along each coordinate). This point is then used as a mean shift in an IS simulation to quickly evaluate the circuit failure probability [5]. The detailed implementation of the method is described in [11] and it is used to evaluate the read timing failure of the small signal SRAM read path (Fig. 2(b)), comprised of a critical path netlist with 12 transistors exhibiting local mismatch variation. Failure is evaluated from a transient SPICE simulation, checking for a polarity of sense amplifier differential output that is consistent with the bitcell data.

In step 1 of the method, the parameter space of threshold voltage fluctuation is sampled on concentric spherical shells, whose radii are adjusted until a sufficient number of failures are observed. For a radius that exhibits sufficient failures (1 to 250 for this work), the failing parameters are averaged to produce an initial IS mean shift that is refined in step 2 of the method. In step 2, threshold voltage parameters are selected in the neighborhood of the current mean shift estimate and are evaluated by a SPICE simulation. The updating of the IS mean shift is designed to gravitate to points of minimum quadratic norm that reside in the failure region. Subsequent trails in step 2 sample within a shrinking neighborhood around the current IS mean shift to increase the resolution of parameter space exploration. Finally, step 3 of the method consists of a conventional Importance Sampling run. As a consistent stopping criterion, empirically validated by accuracy against Monte Carlo and not to be mistaken for the overall variance of the process, the relative sample variance of the IS run is monitored until it falls to 0.1 [5].

The SRAM read path setup and has been discussed in detail in [11], and here we present a full cost break down of the method in Table I. Across a range of probabilities $10^{-6}$ to $10^{-10}$, the exploration cost of Spherical Importance Sampling varied from 1000 to 1,500 simulation runs. The subsequent Importance Sampling stage took 660 to 923 runs. Compared to the previous ISNM work [5], at half the dimensionality (6 instead of 12) and higher probability levels, the previous simulation method remains costlier. This work’s improvement comes from the two-step spherical exploration finding a better shift for the Importance Sampling run. It is also worth highlighting that the local exploration in step 2 is computationally less costly than the directional search in step 1. The two-stage Spherical search effectively handles the dimensionality of 12, considering that over 4,000 simulations are needed just to check all the corners of a 12D cube. With Spherical Importance Sampling much fewer directions are examined while still identifying a suitable mean shift.

A general comparison across a variety of simulation methods [3, 5–7, 13] is presented in Fig. 5. The y-axis gives failure probability levels and the horizontal axis gives the number of total circuit simulations (e.g., SPICE runs) to evaluate the failure probability. Also indicated is the dimensionality of the problem (number of random variables). All methods require less than 10,000 circuit evaluations to identify failure probabilities from $10^{-3}$ to $10^{-9}$, and the relation between failure probability and number of simulation trials is much steeper than the Monte Carlo behavior of $\approx 100/p$. The Spherical Importance Sampling method compares favorably with other works. Indeed, Monte Carlo simulation is much less sensitive to dimensionality than the accelerated statistical evaluation techniques in Fig. 5 which all rely on some type of classification of the parameter space. Developing an accelerated simulation method for a higher dimensional parameter space (e.g., 50) will broaden the applicability of quick statistical simulation techniques.

### IV. Conclusion

This paper discussed two techniques—Loop Flattening and Spherical Importance Sampling—and a method to synthesize them to reduce the statistical analysis of an SRAM block to an Importance Sampling simulation of a chain of component circuits. The key challenge of searching for the most likely failure mechanism in a high dimensionality (12 in this work)

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>COMPARISON OF SIMULATION COST BETWEEN THIS WORK AND [5]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>This Work 12 Dimensions 2-step Spherical Samples</td>
</tr>
<tr>
<td>$P$</td>
<td>$9.08 \times 10^6$</td>
</tr>
<tr>
<td>Step 1</td>
<td>500</td>
</tr>
<tr>
<td>Step 2</td>
<td>500</td>
</tr>
<tr>
<td>Total Exploration</td>
<td>1000</td>
</tr>
<tr>
<td>IS run</td>
<td>660</td>
</tr>
<tr>
<td>Total</td>
<td>1660</td>
</tr>
</tbody>
</table>

Fig. 5. Simulation cost comparison. Not shown is a point for evaluating a failure probability of $1.8 \cdot 10^{-7}$ in 300,000 simulations in a 24 dimensional space by [6].

1Global variation from one chip to another is modeled as a common shift in the mean of threshold voltage distributions. This type of variation is also a concern (but does not dominate simulation cost). It must be separately treated after the impact of local variation is evaluated.
parameter space is addressed by a two-stage process in which a
course direction is obtained first, followed by a local sampling
of increasing resolution. As future work, this method can be
extended to the complete, global row and column path of
large embedded SRAM, as well as to other highly structured
circuits such as adders, FIR filters, and FFT accelerators. Such
highly symmetric, multiplexing structures will become more
prevalemt in the ascent of multi-core chip design.

APPENDIX A

PROOF OF THE LOOP FLATTENING APPROXIMATION

Here, Eq. (5) is shown for the case where \( X \) and \( Y \) are
\( N(0, 1) \) and independent additive delays \( (Z_i = X_i + Y_i)^2 \)
Recall that failure is given by Eq. (3) and the conservative loop
flattening estimate \( P_u \) is defined in Eq. (4).

By the union bound,

\[
P_f := \Pr \left( \max_{1 \leq i \leq R} Z_i \geq t \right) = \Pr \left( \bigcup_{i=1}^R Z_i \geq t \right) \\
\leq \sum_{i=1}^R \Pr (Z_i \geq t) = R \cdot \Pr (Z_1 \geq t) =: P_u .
\]

Then, by incorporating the pair-wise intersection probabilities,
we can introduce an optimistic (lower bound) estimate \( P_x \):

\[
P_f = \Pr \left( \bigcup_{i=1}^R Z_i \geq t \right) \\
\geq \sum_{i=1}^R \Pr (Z_i \geq t) - \frac{1}{2} \sum_{(i,j), i \neq j} \Pr (Z_i \geq t \cap Z_j \geq t) \\
\geq P_u - \frac{R(R-1)}{2} \Pr \left( Z_1 \geq t \cap Z_2 \geq t \right) =: P_x .
\]

This implies

\[
\frac{P_u - P_f}{P_f} \leq \frac{P_u - P_x}{P_x} = \frac{1}{R^2 - 1} \frac{\Pr (Z_1 \geq t)}{\Pr (Z_1 \geq t \cap Z_2 \geq t)} - 1 . \quad (7)
\]

We use the following well-known bound on the tail probability
of the standard Gaussian (let \( W \sim N(0, 1) \), \( w > 0 \)):

\[
\frac{1}{\sqrt{2\pi w}} \left( 1 - \frac{1}{w^2} \right) e^{-\frac{w^2}{2}} < \Pr (W \geq w) \leq \frac{1}{2} e^{-\frac{w^2}{2}} . \quad (8)
\]

Since \( \var(Z) = 2 \),

\[
\frac{1}{\sqrt{\pi t}} \left( 1 - \frac{2}{t^2} \right) e^{-\frac{t^2}{2}} < \Pr (Z_1 \geq t) . \quad (9)
\]

We now examine the intersection probability \( \Pr (Z_1 \geq t \cap Z_2 \geq t) \). Conditioning on \( Y = y \), the two
events \( Z_1 \geq t \) and \( Z_2 \geq t \) are independent. Therefore the
joint probability can be written as:

\[
\Pr \left( Z_1 \geq t \cap Z_2 \geq t \right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_Y(y) \left[ \Pr (X_1 + y \geq t) \right]^2 dy \\
\leq \Pr \left( Y \leq 0 \right) \left[ \Pr (X_1 \geq t) \right]^2 + \\
\int_0^t \int_0^t f_Y(y) \left[ \Pr (X_1 + y \geq t) \right]^2 dy + \Pr (Y \geq t) . \quad (10)
\]

This proof assumes zero mean standard normal variables. The same
arguments can be applied for the general case of non-zero mean normal
random variables of different variance and will hold for generally well-
behaved distributions that decay sufficiently fast.

ACKNOWLEDGMENT

The authors acknowledge the support of the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity.

REFERENCES