Statistical Methodology for Modeling Non-IID Memory Fails Events

*[1]Sabine Francis, [1]Rouwaida Kanj, [2]Rajiv Joshi, [1]Ayman Kayssi, [1]Ali Chehab
[1]American University of Beirut, Beirut, Lebanon
[2]IBM Yorktown Heights, New York, USA

ABSTRACT

We propose a comprehensive and computationally efficient methodology for the estimation of correlated memory fail probabilities. The methodology allows, for the first time, to accurately predict the number of failing memory parts in the presence of correlation between the memory fails due to shared peripheral logic. It relies on importance sampling to model the fail region and emulate different memory architectures. Unlike traditional yield analysis methodologies that assume independent and identically distributed fail events, we record a high overdispersion rate, variance to mean ratio, due to the correlations. This in turn leads to an increase in the number of memory bit fails compared to IID events and is itself a strong function of the memory/peripheral logic grouping. Under extreme operating conditions, our experiments demonstrate overdispersion ratios larger than 10, and more than 23% increase in the number of fails at the 90th percentile level of the array samples/population. However, from a redundancy perspective, the correlations result in reduced column redundancy quota requirements with demonstrated cases with more than 20% reduction in the required quota compared to the IID assumption.

1. Introduction

With the aggressive scaling of CMOS technology, intra-die process variation effects are increasing dramatically [1, 2]. Driven by density requirements, memory designs use the smallest devices and most aggressive design rules and tend to suffer from variability most. For SRAM designs this translates to unavoidable manufacturing variations between neighboring transistors of the same cell. Threshold voltage mismatch between neighboring devices can lead to large number of fails in memory designs and can degrade SRAM performance and yield [3-5]. When combined with other effects such as line edge roughness, narrow width effects, end-of-life effects, and other reliability effects, the scaling of SRAMs becomes increasingly difficult due to reduced margins.

Statistical methodologies addressing the memory yield problem so far typically focus on efficient techniques for estimating rare fail probabilities. This involves integrating the fail region probability via variance reduction techniques or fast sampling methods [6-9], as opposed to traditional Monte Carlo methods. To improve the accuracy of the analysis, memory cross-section abstractions typically account for critical peripheral logic elements and loading effects. However, process variations impact the peripheral logic as well [3, 10]. Statistical analysis for custom logic and memory-interfacing logic is critical; this is important not only from the performance perspective but also from the functional behavior perspective. While statistical timing techniques have been used to achieve full-chip and full-process coverage based on high-level models, the models involved do sacrifice accuracy and do not capture the memory cell variability. Most importantly, it is necessary to analyze the yield in the presence of combined memory/peripheral logic variability effects. This raises the need for simultaneous statistical analysis of the memory cell/logic unit, i.e., variability in the full cross-section. This is a crucial step for design yield optimization and key to lowering design operating voltages (Vmin). This step, as is, still lacks modeling the correlation between the different memory cells that share the same peripheral logic unit.

The authors in [3, 9] extend the rare fail event estimation methods [6-9] to analyze such cross-sections. However, all of these methods focus on the memory cell fails being independent and identically distributed (IID). The fact that peripheral logic is shared among multiple cells, correlations arise between memory cell fails, and existing methodologies fail to capture these correlations. The authors in [10] rely on Gumble distribution to model the worst-case cell per column distribution in order to optimize the sense amplifier set time. Their goal is to identify the set time in terms of worst-case cell distribution. They model the impact of the sense amplifier set time in terms of the signal differential drop developed prior to activation of the sense amplifier during Read; the differential itself is a function of the cell strength and the sense amplifier threshold voltage mismatch. They then build a table for a compensation factor to compensate for interaction between the cell and sense amplifier past the activation of the latter. They only take into consideration variability in the pass gate (one device) in the SRAM cell, and the mismatch representing the variability between two transistors of the sense amplifier. However, they do not consider correlation effects in a true statistical manner. Their method aims at a pessimistic bound and relies on several simplifying assumptions.

In this work, we propose an efficient statistical simulation based analysis methodology for hierarchical memory designs. The methodology guides designers to ensure proper functionality and yield in the presence of variability not only in the cell devices but also in peripheral logic where more than one cell share the same sense amplifier. The methodology avoids pessimism of traditional approaches and captures true yield of hierarchical memory sense amplifier designs under non-IID fail assumptions. The methodology relies on fast sampling methods. This makes it very suitable for rare probability fail estimation of complex designs where millions of cells and thousands of sense amplifiers are subject to random mismatch fails. The paper is organized as follows. Section 2 provides an overview of memory architecture and fail mechanisms. Section 3 provides an overview of non-IID events modeling and estimation. Section 4 presents the proposed methodology. Section 5 presents the results and analysis and section 6 presents the conclusions.

2. Memory Architecture and Fail Mechanisms

Fig. 1 presents the general hierarchy of a memory block. Several columns share a sense amplifier, and multiple cells share a bitline. To access a cell, the address signals are translated into column enable and wordline activation signals. During a Read operation, the cell draws charge from the precharged bitline on the side of '0' storage node. To reduce cell access time during Read

*S. Francis is currently with the University of Texas Austin

operation, the sense amplifier is designed to detect small changes in the bitline voltage levels. Mismatch in the cell and/or sense amplifier transistors can lead to misread and in extreme conditions to node upset or flipping of the contents of the memory cell. In addition to process variations, many factors play a role in triggering such events. Key design concerns include the sense amplifier set time [10], the operating logic and memory supplies, the number of cells per bitline and other architectural considerations. A successful yield analysis methodology aims at capturing and abstracting these effects to allow for proper modeling of the feasibility space. Fig. 2 illustrates a sketch of the signals involved during the Read operation along with the schematic of an SRAM cell. The Read operation can be described as follows assuming Read '0' at node L. The cell is selected by turning ON the column enable signal (C.E.) and the wordline signal (WL). When the WL turns ON, the '0' on node L pulls down the precharged bitline BLT; BLC remains high. Thus a voltage difference builds between the two bitlines. At time t1, the sense amplifier is activated by turning ON the sense enable signal (S.E). The sense amplifier then senses this difference and helps amplify it. Eventually, the sense amplifier node true is pulled down to '0'. For our analysis we focus on the following two key design metrics.

1. Readability: A proper Read '0' on node L discharges the precharged node true, which in turn charges a pre-discharged global read signal (Fig.1). In general the Read operation is successful if the voltage difference between the two bitlines of the selected column is larger than the threshold voltage mismatch, in the sense amplifier devices M1 and M2 [10], that arises due to variability. The S.E. timing and C.E. signal timing also affect proper functionality of the Read operation. Particularly, there is an increased sensitivity to the sense amplifier mismatch in early sense enable situations. A wide column enable pulse may help the sense amplifier as more difference builds in the bitline, but this leads to more power consumption.

2. Stability: Upon Read, the cell '0' storage node is subject to noise injected from the precharged bitlines. If the cell is weak due to process variations, this may lead to flipping the contents of the cell and hence cause a fail. Early sense enable with a strong sense amplifier may increase the risk of flipping the cell contents.

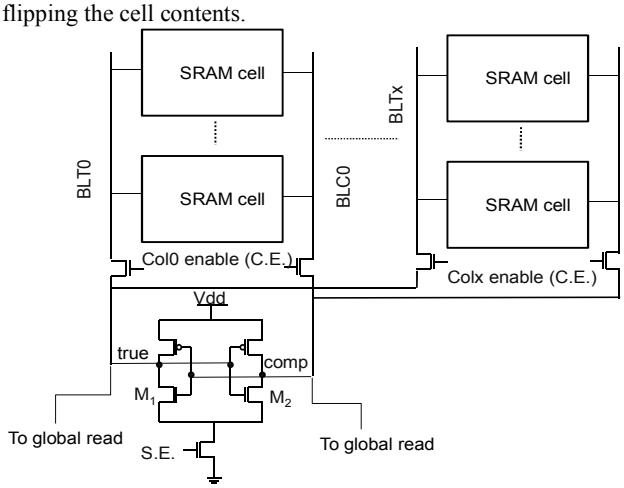


Figure 1. Memory cross-section. A sense amplifier may be shared among several columns.

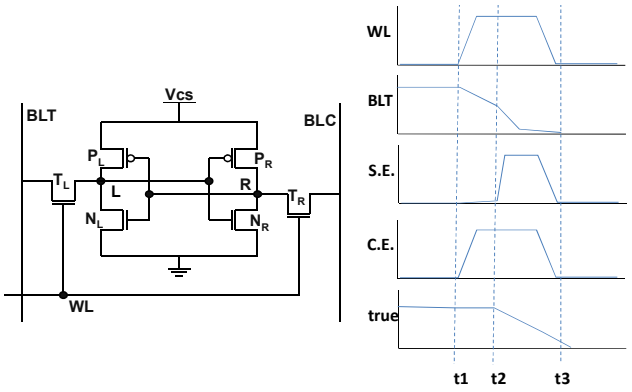


Figure 2. Enable signals during Read '0' on node 'L'.

3. IID and Non-IID Events

3.1 Statistical Sampling and Binomial Events

Under extreme conditions of variability, mismatch in the cell and/or sense amplifier devices can lead to functional fails. From a statistical simulation perspective, the process variation parameters, x_i, represent independent random variables. One defines a pass/fail condition in relation to a set of predefined metrics; and each statistical sample point is accompanied with a statistical simulation to evaluate the pass/fail conditions. Each simulation is treated as a Bernoulli experiment [6] with an indicator function I(x).

$$I(x) = \begin{cases} 0 & success \\ 1 & fail \end{cases} \quad (1)$$

In Monte Carlo, the "sample points are selected according to the design parameter distributions" [6]. For the case of the SRAM cell, these would represent the threshold voltage variations of the devices due to random dopant fluctuations and other parameters. The probability of failure and the corresponding variance are defined as follows, where N_M is the number of Monte Carlo samples.

$$P_f = \frac{1}{N_M}\sum_{j=1}^{N_M} I(x^j) \quad \text{and} \quad \sigma^2 = \frac{P_f \times (1-P_f)}{N_M} \quad (2)$$

For low fail probabilities, the convergence of the standard deviation of the estimate is very slow; this leads to significant simulation time and raises the need for variance reduction and importance sampling methods [6-9]. Importance sampling enables fast convergence by distorting the natural sampling function *p(x)* and generating more points around the tail. Theoretically, it is backed up by equation (3), where *g(x)* is the distorted sampling function with a shift μ_s obtained from an initial sampling phase.

$$E_{p(x)}[\theta] = E_{g(x)}(\theta \cdot \frac{p(x)}{g(x)}) \quad \text{where} \quad g(x) = p(x - \mu_s) \quad (3)$$

The corresponding probability of failure is presented below.

$$P_f = \frac{\sum_{i=1}^{N_{IS}} y(x^i)}{\sum_{i=1}^{N_{IS}} w(x^i)} = \frac{\overline{y(x)}}{\overline{w(x)}} \quad (4)$$

where $w(x) = \frac{p(x)}{g(x)}$ and $y(x) = I(x)w(x)$

N_{IS} is the number of samples drawn from *g(x)*. The method achieves remarkable speedups compared to the standard Monte Carlo method should the proper importance sampling function be identified [6, 7].

Given a memory array of 1 million cells, each cell can be modeled as a trial with probability of fail P_f. Assuming that the cell fails are independent, the probability distribution of the

number of cell fails is modeled after a binomial distribution, *Binom*(). For example if we assume that a good array has at most 10 cells failing, then the yield can be estimated as follows,

$$Binom(10, N = 1e6, P_f) = P(num_{Fails} \le 10) = \sum_{n=0}^{10} P(\text{n } fails) = \sum_{n=0}^{10} {}^{N}C_{n}\ P_f^n * (1 - P_f)^{N-n} \quad (5)$$

where N is the number of independent trials (memory size). In the limit, the law of rare events indicates that the occurrence of events (fails) follows approximately the Poisson distribution whenever the number of trials is high and the probability of occurrence (P_f in this case) is low. Finally, for IID cells, the expected number of fails and its variance for the memory array can be derived according to (6). When P_f is small, the variance of the expected number of fails is equal to the mean, $\mu \cong \sigma^2$.This is referred to as *equi-dispersion*.

$$\mu = N * P_f \quad \text{and} \quad \sigma^2 = N * P_f(1 - P_f) \quad (6)$$

3.2 Non-IID Events

Often in count data, and in real life problems, the assumption that the events are IID is too strong [11-15]. It is violated in real data; overdispersion or underdispersion can occur due to correlation in the data sets. Most of the existing literature propose that a non-IID system is to be fit to a negative binomial distribution, correlated binomial, generalized Poisson distribution, beta binomial distribution or other statistical models that assume some form of correlation between the events. Under these models, the equi-dispersion assumption is relaxed because the data almost always rejects the assumption that the variance equals the mean. The model reduces to the binomial distribution when the correlation factor drops to zero. For a general negative binomial (NB) model described in equation (7), the variance is related to the mean according to (8).

$$P_f(y|\mu, \alpha) = \frac{\Gamma(y+v)}{\Gamma(v)\times\Gamma(y+1)} \times \left(\frac{v}{v+\mu}\right)^v \times \left(\frac{\mu}{v+\mu}\right)^y \quad (7)$$

$$\sigma^2 = \mu + \alpha\mu^p \qquad \text{with } \alpha = v^{-1} \quad (8)$$

$\alpha(\geq 0)$ and v are referred to as the overdispersion coefficients and are estimated using either the maximum likelihood method or the method of moments [14]. If $\alpha = 0$, (hence v very large) the model reduces to the Poisson distribution as $\sigma^2 = \mu$. NB1(2) are the most popular negative binomial models with power factor p=1(2). NB2 is the standard formulation for the negative binomial.

To illustrate how the overdispersion affects the yield, we setup the following experiment: $N=10^6$, $P_f=10^{-6}$, and the maximum allowable number of fails is equal to 3. Fig. 3 illustrates a plot of the (constant) yield assuming a binomial model under IID events, and the corresponding negative binomial yield estimate as a function of the overdispersion coefficient v.

Revisiting the memory array example, each trial represents a cell being accessed, and the number of trials typically corresponds to the number of cells per chip. However, when variability kicks in the peripheral logic and particularly in the sense amplifier, the unit understudy should include both the cell and sense amplifier. Because many cells share the same sense amplifier, the assumption that the fail events (of two cells that share the same sense amplifier) are independent no longer holds because of the many-to-one organization of cell to sense amplifier in memory designs. In fact, the memory cell fails are correlated for the cells that share the same sense amplifier. So if there are N cells and M cells per sense amplifier, the N Read accesses will occur through N/M sense amplifiers.

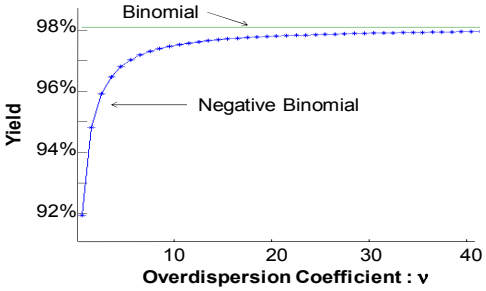
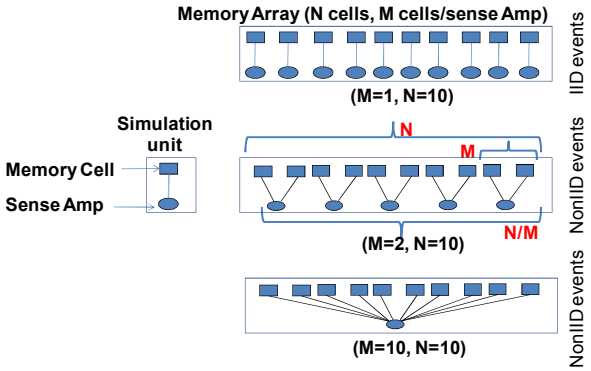


Figure 3. Negative binomial yield estimate as function of the overdispersion coefficient $v = \alpha^{-1}$ for the case $N=10^6$, $Pf=10^{-6}$ and maximum number of fails equal to 3.

Fig. 4 illustrates a hypothetical overdispersion example where the cells are all assumed to be perfect, and the fail occurs only due to a bad sense amplifier. For our rough analysis, we assume that on average every tenth sense amplifier will be bad.

We consider the three possible grouping cases in Fig. 4. We also have 10 replications for each case. M is the number of cells/sense amplifier and N is the total number of cells.

1. (M=1, N=10): each replication has 10 sense amplifiers, and hence will have one bad sense amplifier and one Read fail on average.
2. (M=2, N=10): each replication has 5 sense amplifiers, and every other replication will have a bad sense amplifier on average. But that amplifier controls 2 cells. Hence every other replication may have 2 fails, and hence on average we get one fail per replication.
3. (M=10, N=10): each replication has 1 sense amplifier, and every 10^{th} replications, we get one bad sense amplifier. This will lead to 10 Read fails in the 10^{th} replication and none in the others. Hence on average the number of fails is 1 fail per replication, but the variance is very large. Hence, we have overdispersion.



P_f(Sense Amp| Ideal cells)	(M, N)	Expected # Fails	Variance (# Fails)
0.1	(2, 10)	1	1.16
0.1	(10, 10)	1	9.98

Figure 4. An illustration of the many-to-one mapping.

4. Proposed Methodology

Generating fail distributions by relying on true simulation of the group-based architecture can be very costly especially when dealing with large memory designs. Traditional statistical analysis generates sample points to estimate P_f for the unit (single access

of a cell and sense amplifier) and not the grouping effect. Only when the IID assumption holds can the distribution of the expected number of fails be derived from P_f. However, one can exploit the unit analysis to build models which in turn can be used for non-IID event simulation as we propose next.

4.1 Methodology Overview

Fig. 5 presents an overview of the proposed methodology. It relies on three main steps.

- *Step I:* We perform traditional statistical analysis of the memory unit via importance sampling. This step finds the center of gravity (COG) of the critical fail region in the process variation space of the memory unit [6, 7]. It also finds the unit probability of fail P_f which is otherwise difficult to derive via Monte Carlo.
- *Step II:* P_f and COG are used to construct an approximate unit fail boundary. By relying on the closest point method, the parameters of the bounding hyperplane are derived according to Table I. The hyperplane is constructed along the direction of the COG at a distance from the origin that is derived to match the fail probability P_f. A sketch of the fail boundary is presented in Fig. 6.
- *Step III:* We rely on numerical simulation to estimate the number of memory fails for the memory design. We generate sample points representing process variation parameters of the (N cells x N/M sense amplifier) for an instance of the memory architecture and grouping understudy. Each cell is then paired with its amplifier and the resultant unit parameters are tested against the fail boundary model. By emulating replicas of the memory design understudy, we obtain distributions of the expected number of fails and derive the corresponding maximum likelihood estimates of the negative binomial distributions [11].

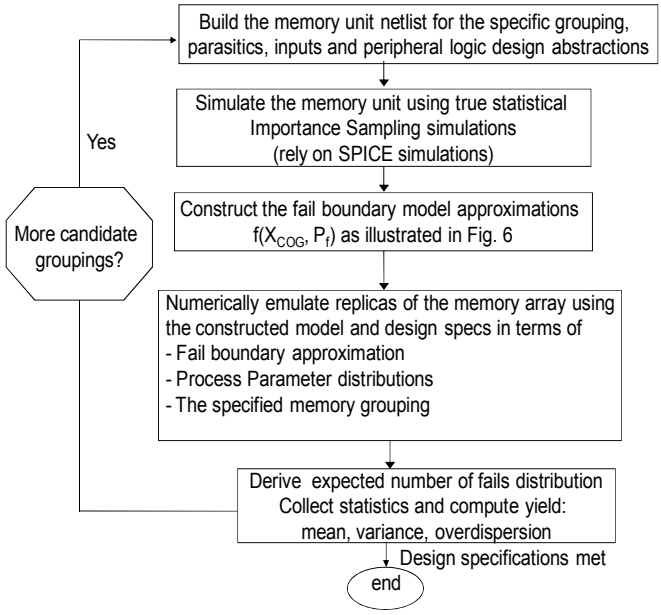


Figure 5. Methodology flow diagram.

5. Analysis and Results

In this section, we apply the proposed methodology to 45nm SRAM cell designs. First we study the proposed methodology when applied to theoretical problems.

Table I. Method for determining parameters of the bounding hyperplane.

Consider a "d" dimensional system representing the process variation space (x_1, x_2, …, x_d) all distributed according to a normal Gaussian distribution $N_G(0, 1)$.

Let plane P: $y_0 = \sum a_i x_i$ to be the desired bounding hyperplane. We choose y_0 and a_i to satisfy the following requirements.

(i) We require that the COG of fails vector x_{COG} be the direction normal to the plane, and hence we use it to derive the coefficients a_i of the hyperplane based on the definition of a normal to a hyperplane.

$$a_i = {}^{x_{COGi}}/_{|\boldsymbol{x}_{COG}|}$$

(ii) We require that the fail region probability equals to P_f derived from the importance sampling stage. Let the distance from the origin to the closest point $\boldsymbol{x_{CP}}$ on the bounding hyperplane be d_p, then according to the closest failing point approximation [6], the probability of fail, P_f,

$$P_f = P(z > d_p) \Leftrightarrow d_p = \varphi^{-1}(1 - P_f)$$

where φ is cumulative distribution function of a Gaussian function. The closest point lies on the normal direction at a distance d_p, thus, the coordinates of $\boldsymbol{x_{CP}}$ are

$$\boldsymbol{x_{CP}} = d_p * {}^{x_{COG}}/_{|x_{COG}|}$$

Finally, $\boldsymbol{x_{CP}}$ belongs to the hyperplane. Hence

$$y_0 = \sum a_i \boldsymbol{x_{CPi}}$$

$$y_0 = d_p * \sum a_i \frac{x_{COGi}}{|x_{COG}|}$$

$$\therefore y_0 = d_p = \varphi^{-1}(1 - P_f)$$

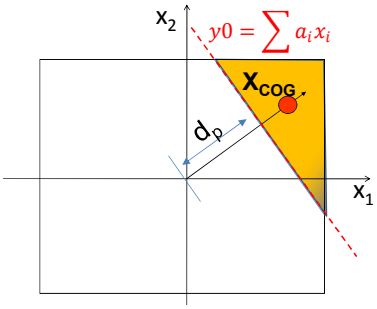


Figure 6. Fail boundary modeling. x_1 and x_2 represent two process variation parameters without loss of generality. COG stands for center of gravity of fails; the fail region is shaded.

5.1 Theoretical Analysis

We extend the proposed methodology to study theoretical examples of Fig. 4. Without loss of generality and for purposes of illustration, we model the memory unit by two variables: x_a for the cell and x_b for the sense amplifier. The bounding hyperplane is of the form (9). We also assume that x_a and x_b are normally distributed N(0, 1), and that a and b satisfy (10).

$$a * x_a + b * x_b = y_0 \quad (9)$$

$$\sqrt{a^2 + b^2} = 1 \quad (10)$$

For this assumption, $y_0 = d_p$ represents an d_p-sigma fail probability in the normalized space according to table I. The higher the coefficient a, the higher is the significance of the cell for fails. We study the impact of grouping on overdispersion as a function of the grouping ratio, the number of cells/sense amplifier, $M \in [10, 100, 1000, 10000]$ and $a \in [0,1]$ when the number of cells N=1e6.

Fig. 7 plots the logarithm of the ratio of variance to the mean of the number of fails for 100 replications. Overdispersion is found to increase with both the group size M and the sense amplifier coefficient 'b' (hence decreases with the cell coefficient 'a'). This trend is still maintained as the fails become rarer. However, for rare fail events the number of replications is no longer sufficient to capture the extreme cases. An example would be (a=0, y_0=5 sigma, N=1e6, M=1e4); this circuit has 100 sense amplifiers and the 100 replications only involve 100x100 sense amplifiers whereas the fails only happen in the sense amplifier with rate of 1 to 2.7 million. Fig. 8 illustrates a closer look at the cumulative distribution function of the number of fails for the case y_0=3 sigma. We clearly see the wider spread of the number of fails as opposed to the narrow spread for the IID assumption.

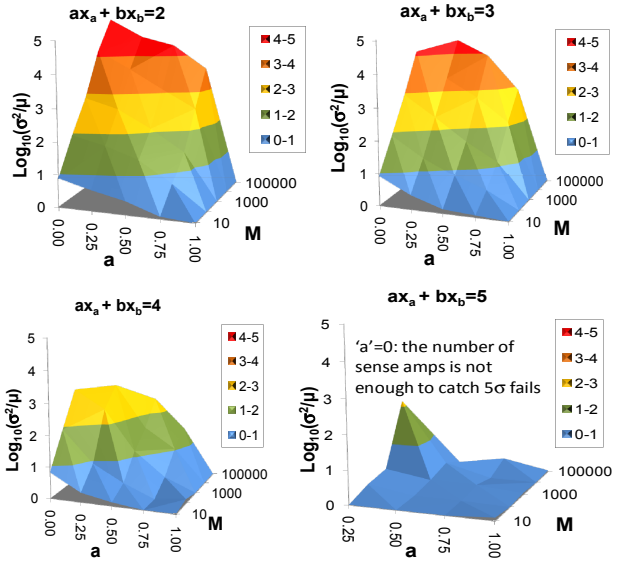


Figure 7. Overdispersion trends for 2, 3, 4 and 5 sigma fails.

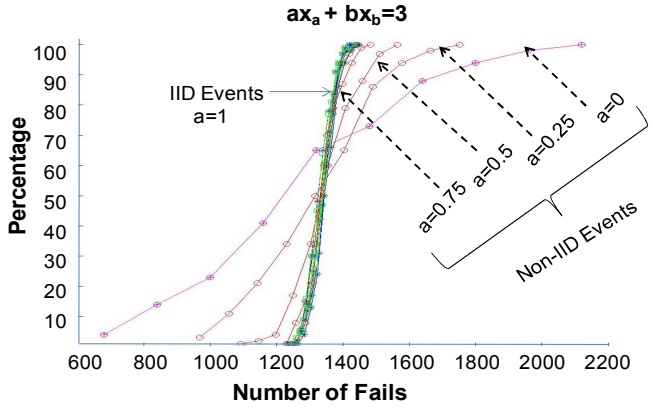


Figure 8. Cumulative distribution function for y_0=3 sigma.

5.2 SRAM Analysis

We study a 45nm minimum sized SRAM cell design with a cell beta ratio of 2. For the memory unit analysis, we apply random threshold voltage variations to the 6 transistors of the SRAM cell as well as to the 4 transistors of the sense amplifier (2 PMOS devices and M1 and M2 in Fig. 1). σ_{Vt} is set to 12% Vt for the minimum sized SRAM cell devices, and 8% Vt for the sense amplifier devices. We perform SPICE simulations using predictive technology models [16].We apply variability in both cell and sense amplifier and we analyze the design for the corners presented in table II. Low voltage operation is found to incur fails. Logic supply, Vdd, set to higher values than the cell supply Vcs is meant to study a stronger sense amplifier at negative Vdiff=Vcs-Vdd. Early sensing also aggravates misread. If the column enable turns off early, misread may happen without affecting the cell contents. A regular or late column enable can alleviate the misread as the cell continues to help the sense amplifier at the expense of extra power; yet this leads to higher risk for a weak cell to flip its contents. For the early sense enable, we let the sense enable signal arrive 50ps earlier than the typical time in a 2GHz design. For early column enable, we turn OFF the column enable at the onset of sense enable signal.

Table II. Predefined simulation corners.

Corner	Vdd	Vcs	Sense Enable
1	0.7	0.7	Regular
2	0.8	0.7	Regular
3	0.7	0.7	Early
4	0.8	0.7	Early
5	0.8	0.7	Early w/ Early C.E.

The proposed methodology is applied and 100 replications are used to emulate the variance and mean of the expected number of fails. Table III presents the results for corners 1 and 2. The number of fails is small and represents the base case for a 1e6 array design. Table IV presents a summary of the Readability study for two different array sizes N=1e5 and N=1e6 for the critical corners 3, 4 and 5. Along with a high overdispersion ratio, we notice 16%- 30% increase in the number of fails at 3 sigma estimates compared to IID assumption. Fig. 9 plots log(σ^2/μ) versus the grouping number M. As expected, overdispersion increases with M and the sensitivity to the sense amplifier variation is maximum in corner 5 when the column enable is turned OFF early.

Table III. Statistics for corners 1 and 2, Readability.

Corner	M	Expected # Fails	Variance # Fails
1	**64**	1.23	1.27
	128	1.35	2.03
2	**64**	0.88	0.73
	128	1.08	1.09

Table IV. Statistics for corners 3, 4 and 5, Readability.

Corner	Array size	M	Expected # Fails	Variance # Fails	σ^2/μ	% increase #fails at 3 sigma
3	**1e5**	**64**	23	46	2.0	16%
	1e5	**128**	24	83	3.5	33%
4	**1e5**	**64**	68	232	3.4	23%
	1e5	**128**	69	302	4.4	29%
5	**1e5**	**64**	346	2010	5.8	20%
	1e5	**128**	349	3820	10.9	32%
3	**1e6**	**64**	232	555	2.4	9%
	1e6	**128**	235	907	3.9	16%
4	**1e6**	**64**	690	1886	2.7	7%
	1e6	**128**	696	3500	5.0	13%
5	**1e6**	**64**	3451	18196	5.3	6%
	1e6	**128**	3521	42796	12.2	12%

We remove the Read assist circuitry (not illustrated here), and study the stability fails of the SRAM cell (cell node flipping) for corners 3 and 4. Tables V and VI summarize the results for different array size (N=1e5, 1e6). Again we note an increase in

the overdispersion with the grouping number M and the corner conditions in terms of sensitivity to sense amplifier. Next we study the implications of overdispersion on the near 100% yield of the memory design. We rely on the statistics to fit the negative binomial model and we compare the yields as illustrated in the examples of Figs 10-11. We note for the same yield requirement, an increase of the order of 23% in the allowed number of fails compared to the IID assumption for critical combinations.

Table V. Statistics for corners 3 and 4, Stability, N=1e5.

Corner	M	Expected # Fails	Variance # Fails	σ^2/μ	% increase #fails at 3 sigma
3	**64**	15	19	1.3	6%
	128	15	20	1.3	6%
	256	16	25	1.5	10%
	512	16	42	2.6	26%
4	**64**	59	107	1.8	10%
	128	57	116	2.0	12%
	256	57	231	4.1	29%
	512	56	244	4.4	31%

Table VI. Statistics for corners 3 and 4, Stability, N=1e6.

Corner	M	Expected # Fails	Variance # Fails	σ^2/μ	% increase #fails at 3 sigma
3	**64**	159	214	1.3	3%
	128	159	257	1.6	5%
	256	159	345	2.2	9%
	512	156	445	2.9	13%
4	**64**	583	845	1.5	2%
	128	583	1627	2.8	7%
	256	584	2044	3.5	10%
	512	583	3684	6.3	17%

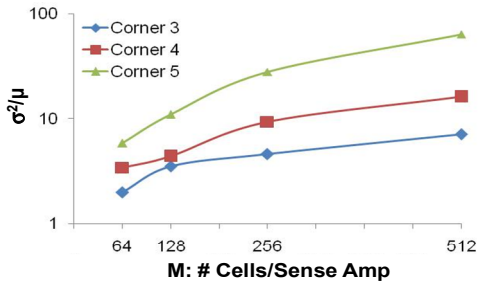


Figure 9. Overdispersion ratio trends for Readability, N=1e5.

Next we study the number of impacted sense amplifier groups from a column redundancy perspective. Our analysis shows that while the variance of the number of fails increases, the fails tend to be clustered and hence the number of failing/bad sense amplifier groups under the nonIID assumption is smaller as opposed to that predicted in an IID model. To emulate this, we rely on the methodology to model two sets of data for a given corner/grouping analysis. Set 1 represents the proposed nonIID model. Set 2 represents the IID model. The IID cell fail assumption assumes that fails can happen anywhere in the array with the same probability of fail. For each set we emulate 300 replications of the design understudy, and collect the following:

a) The distribution of the number of fails
b) The distribution of the number of failed groups (aka number of sense amplifier groups with failing cells).

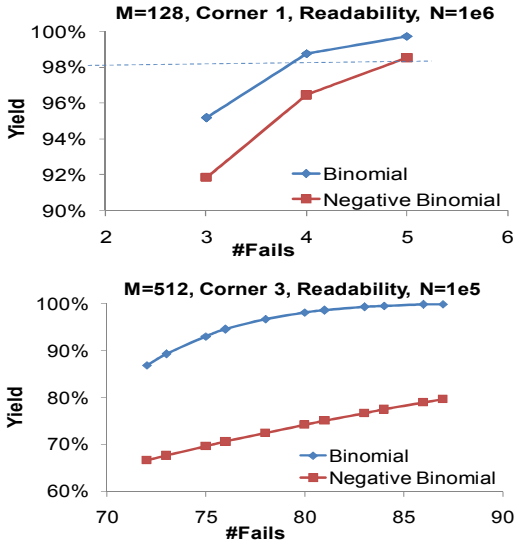


Figure 10. Yield versus # allowed Fails, Readability. Negative Binomial corresponds to nonIID model.

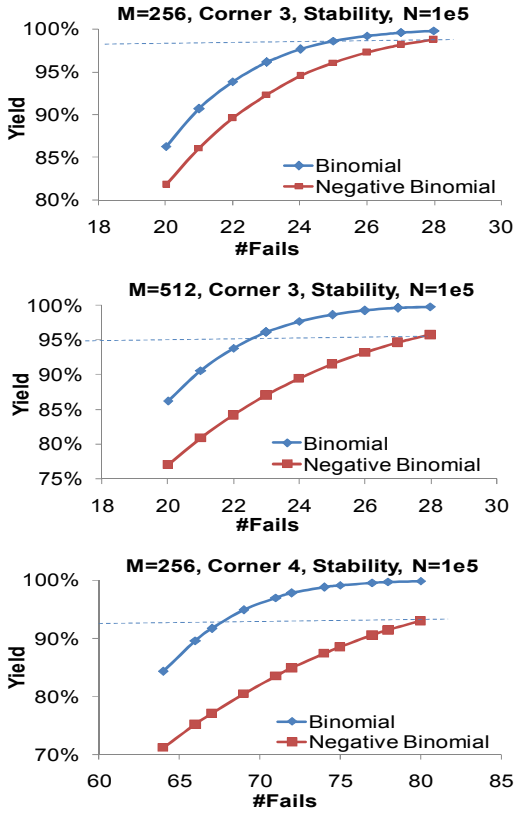


Figure 11. Yield versus # allowed Fails, Stability. Negative Binomial corresponds to nonIID model.

Figures 12-14 present case studies to analyze redundancy requirements. While the number of fails (bit error rate) increases in the presence of overdispersion, the number of impacted groups is found smaller in the nonIID assumption mainly because of the dependency of fail in part on the sense amplifiers. For the example in Fig. 13, the nonIID model results in a 25% reduction in the required column redundancy quota. For the extreme case in

Fig.14, we notice significant reduction in the failed groups compared to the IID assumption. Hence from a bit error rate the nonIID model leads to significant increase in the number of fails, however the correlated fails require reduced redundancy quota when dealing with sense amplifier group repairs.

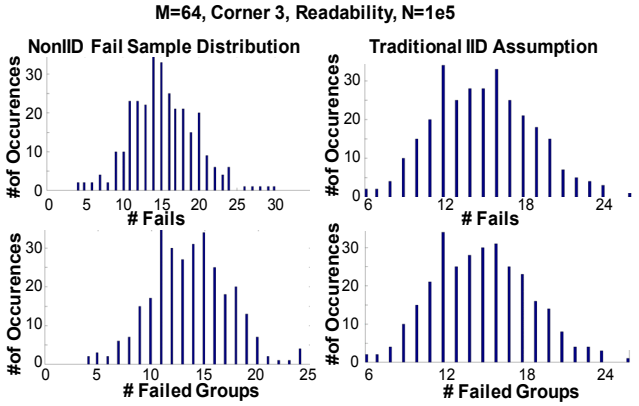


Figure 12. NonIID expected number of failed groups spread (left) is comparable to IID assumption (right); number of fails is small.

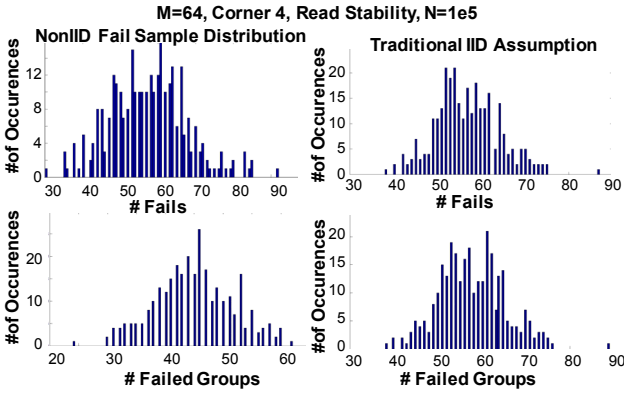


Figure 13. NonIID number of failed groups spread (left) is smaller than the IID assumption (right).

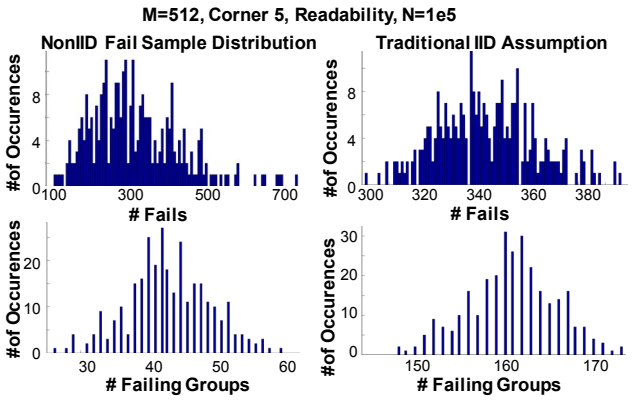


Figure 14. Extreme case where the expected number of nonIID failed groups (left) is much smaller than that of the IID assumption (right) due to high overdispersion.

6. CONCLUSIONS

We propose an efficient statistical methodology for modeling correlated memory fails. The methodology is comprehensive and captures hierarchical grouping effects in the presence of variability in the cell and peripheral logic. The methodology is universal, can be extended to different peripheral logic styles and is solely based on statistical simulation and modeling of the memory unit. We demonstrate an increase of the bit error rate due to overdispersion in practical SRAM designs. We also demonstrate a reduction in the column, sense amplifier group, redundancy requirement compared to the traditional IID assumptions which can be quite pessimistic in extreme conditions.

Acknowledgement. The authors would like to acknowledge the University Research Board (URB) at the American University of Beirut for funding this research study.

7. REFERENCES

[1] T. Karnik, V. De, S. Borkar, "Statistical design for variation tolerance: key to continued Moore's Low," IEEE ICICDT, 2004, pp. 175-176.

[2] S. R. Nassif, "Design for variability in DSM technologies [deep submicron]," IEEE ISQED, 2000, pp. 451 – 454.

[3] R. V. Joshi et al., "The dawn of Predictive chip yield design: along and beyond the memory lane," *IEEE Design & Test of Computers*, Dec 2010.

[4] R. V. Joshi et al., "A low power and high performance SOI SRAM circuit design with improved cell stability," SOI Conf., 2006, pp. 211 – 214.

[5] A. Pelella et al., "Evaluation and alleviation of SOI impacts on SRAM functionality and yield," SOI Conf., 2008, pp. 41 – 42.

[6] A. Signhee, and R. Rutenbar, Editors. *Extreme Statistics in Nanoscale Memory Design*. Springer. 2010.

[7] R. Kanj et al., "Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events," IEEE DAC, 2006, pp. 69-72.

[8] A. Singhee et al., "Statistical blockade: a novel method for very fast Monte Carlo simulation of rare circuit events, and its application," *IEEE DATE,* 2007, pp. 1-6.

[9] C. Dong, and X. Li, "Efficient SRAM failure rate prediction via Gibbs sampling," IEEE DAC, 2011, pp. 200-205.

[10] R.M. Houle, "Simple statistical analysis techniques to determine optimum sense amp set times", *IEEE Journal of Solid-State Circuits*, Aug 2008.

[11] A. Cameron and P. Trivedi. *Regression Analysis of Count Data*. Econometric Society Monographs No.30, Cambridge University Press, 2007.

[12] N. Ismail, A. Jemain, "Handling overdispersion with negative binomial and generalized Poisson regression models," *Casualty Actuarial Society forum*, winter 2007.

[13] V. Nicola et al., "Modeling of correlated failure and community error recovery in multiversion software", IEE Trans on Software Engineering, 1990

[14] Z. Feng et al., "Correlated binomial variates: properties of estimator intraclass", Statistics in Medicine II, 1992

[15] M. Bakaloglu et al., "Modeling correlated failures in survivable storage systems", 2002

[16] Predictive Technology models http://ptm.asu.edu