Copula Models of Correlation: A DRAM Case Study

C. Glenn Shirley and W. Robert Daasch, *Member*, *IEEE*

Abstract—Variable bit retention time observed in a 65-nm dynamic random access memory (DRAM) case study will cause miscorrelation between retention times occurring in Test and Use. Conventional multivariate normal statistics cannot adequately model this miscorrelation. A more general copula-based modeling approach, widely used in financial and actuarial modeling, solves this problem. The DRAM case study shows by example how to use copula models in test applications. The method includes acquiring data using a test vehicle, fitting the data to a copula-based statistical model, and then using the model to compute producer- and customer-oriented figures of merit of a product, different from the test vehicle. Different array sizes, fault tolerance schemes, test coverage, end-use (datasheet), and test condition specifications of the product are modeled.

Index Terms—Integrated circuits, dynamic random access memory (DRAM), testing, fault tolerance, reliability, yield models

1 Introduction

Each bit of a dynamic random access memory (DRAM) retains its information as stored charge on a capacitor. After the bit has been written to, the charge leaks away so that valid data has a characteristic retention time. To retain the information, the bit must be read and refreshed with a specified time interval between refreshes. DRAM memory cells in every technology node can have a defect which causes a some bits to have a variable retention time (VRT), while most bits have stable retention times (SRT) [1]–[3]. The VRT behavior is an example of random telegraph noise (RTN) in gate-induced leakage (GIDL) current caused by a trap in gate oxide [4] or a defect in silicon [5] at the near-surface drain-gate boundary of the transistor in the DRAM cell. At any time the defect can transition reversibly between two states. One of the states is associated with a higher leakage current and shorter retention time. $t_{\max}$ and $t_{\min}$ are scale parameters of exponential distributions of duration of the maximum and minimum dwell times of a bit in the maximum and minimum retention time states. The states are maintained for many minutes [5], so retention times are manifested at the test process step in manufacturing ("Test") differently from how they are manifested in end-use ("Use"). Test, being brief, may "see" a VRT bit in either the high or low leakage state. The probability that Test will find a bit in the maximum retention time state is $s = t_{\max}/(t_{\max} + t_{\min})$. On the other hand, since Use has an indefinite duration, a VRT bit's high leakage state (worst-case) will certainly occur in Use. If a VRT bit passes a Test screen in a low-leakage state, and the high leakage state causes the retention time in Use to be shorter than the specified refresh time, then the VRT bit will fail in Use.

The proportion of VRT bits can be controlled in the silicon fabrication process by reducing the density of, or passivating, RTN-inducing defects, or by reducing the mechanical stress which activates the defect [6]. But Test screens and fault-tolerant array design are still needed to meet yield and quality targets for a product array of bits. Data with VRT bits which fail in Use are not suited to correction by ECC schemes used for soft errors (for example, due to cosmic rays) because of the performance impact of repeated error correction of a bit stuck in a failing state for many minutes. Run-time-repair schemes suited to "hard" bit failures [7] can be used.

The performance and quality requirements of a memory product may be met in different ways with possibly different costs. For example, the fraction of arrays with VRT bits escaping to Use, and failing, can be reduced by setting the Test retention time much longer than the refresh time in Use. This has a high cost of rejecting many good arrays (overkill) or repairing many "innocent" bits at Test. On the other hand, if a run-time-repair scheme is employed, the Test condition may be set closer to the Use condition and overkill may be reduced at the cost of design complexity. Tradeoffs like this occur at all stages of the product lifecycle, from product definition to manufacturing. Decision-making requires a statistical model of the memory product which adds considerations of array size, array repair capacity, Test conditions and datasheet (Use) specifications to the bit-level instability characteristics measured in recent studies [8]. This paper describes a new approach to the statistical modeling.

The paper breaks new ground in two aspects of statistical modeling: 1) Model-fitting involves selection of the mathematical forms of distributions to be used, and determination of goodness-of-fit of data to the models. 2) Inference involves "what-if" transformation of the fitted mathematical models to conditions different from the data (different array sizes, fault tolerance, different Test and Use conditions), and definition of the rules of decision-making. Decision-making rules use

• *The authors are with the Integrated Circuits Design and Test Laboratory, Portland State University, Portland, OR 97207. E-mail: cshirley@pdx.edu.*

Manuscript received 01 Oct. 2012; revised 07 Apr. 2013; accepted 05 June 2013. Date of publication 13 June 2013; date of current version 12 Sep. 2014. Recommended for acceptance by S. Shukla.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TC.2013.129

0018-9340 © 2013 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

```
Group 1: 000110111111
Group 2: 000001111111
Group 3: 000111101111
Group 4: 000001111111
Group 5: 000111111111
                ↓First zero from right.
    AND: 000000101111 i_max = 8
     OR: 000111111111 i_min = 3
            ↑Last zero from left.
```

Fig. 1. Example of the pass/fail pattern of a VRT bit, and extraction of i_{max} and i_{min}. Pass/Fail is indicated by 0/1.

TABLE 1
Bit Categories with $r < 604$ au Failing in At Least One of the 18 Environmental Conditions

SRT-only	SRT, but not VRT	A	1610
VRT {	VRT and SRT	B	288
	VRT, but not SRT	C	64
	Total bits	N	48750000
	SRT PPM	A/N	33
	VRT PPM	(B+C)/N	7

carefully-defined figures of merit closely related to cost models, such as yield loss (*YL*), overkill loss (*OL*), and customer-perceived defect level (*DL*).

Model-fitting for semiconductor products usually involves fitting a multivariate normal (Gaussian) distribution to the data. But when a bivariate normal model was fitted to DRAM bit retention time data acquired at PSU's ICDT lab (in Section 2.2 below) it was found that the model did not properly characterize the deep tail of the data. Unlike the data, the correlation in the Gaussian model fades away as one moves deeper into the bivariate tail. Actuarial and financial applications have also encountered this problem [9]. Inadequacy of multivariate normal models for financial and actuarial applications has motivated rapid development over the past decade of copula-based modeling methods because copulas provide a completely general approach to modeling multivariate dependency. Nelsen [10] and Trivedi and Zimmer [11] give good introductions to copulas. This study finds that the Clayton copula, which differs fundamentally from the usual bivariate normal Gaussian model, is needed to describe the underlying dependency (correlation) structure of DRAM VRT behavior, and the way it is manifested in Test and in Use.

Semiconductor product applications require unique methods to handle scaling to various array sizes, for handling fault tolerance, for modeling Test and Use conditions, and for computing and using figures of merit closely related to product cost and quality models. This paper develops the necessary statistical machinery to do all of this for the DRAM application. The methods are, however, quite general and may be applied to any semiconductor product for which miscorrelation between Test and Use or among Test operations needs to be modeled.

The plan of the paper is as follows: In Section 3 copula models are extracted from the DRAM data described in Section 2. The central problem of model extraction is choice of the copula used to fit the data. Section 3 shows the shortcomings of the Gaussian copula, which mirrors the problem of multivariate normal models, and uses the Clayton copula which is well-suited to the DRAM data. Section 4 covers the inferential aspects of the application of copulas to test. These include:

- Modeling how Test and Use are manifested.
- Scaling from bit-level to array level.
- Modeling tolerance to single-bit faults.
- Modeling active repair at test.
- Definition of figures of merit (FOMs).

Section 5 describes where statistical copula-based modeling method fits in a wider context, and Section 6 indicates future directions.

2 DRAM Case Study

2.1 Experimental Design

The experiment follows a design similar in principle to that of Kim et al. [8] except that only the retention time minimum and maximum for each bit was determined. The experiment did not determine the time constants $t_{\max}$ and $t_{\min}$ of the maximum and minimum retention time states.

Test chips with four identical DRAM arrays on each chip were fabricated in a 65 nm process. Each of the four arrays on a test chip has 1,218,750 bits. Test chips packaged in ball grid array packages and 10 test chips, prescreened for gross failures, were selected for the experiment. So the number of bits tested is $10 \times 4 \times 1,218,750 = 48,750,000 (49\ \mathrm{Mb})$.

The arrays were tested in PSU's ICDT Lab on a Credence Quartet tester with temperature controlled by a Silicon Thermal Powercool LB300-i controller. Temperature was measured by a calibrated sensor on the silicon die. Pass/fail at 12 retention times for each bit in the array was determined at 18 environmental conditions, and the physical x, y location and retention time of each failing bit in the array was recorded. The environmental conditions were:

- Three temperatures: $105°\mathrm{C}$, $115°\mathrm{C}$, $125°\mathrm{C}$.
- Three values of supply voltage, V_d: 0.8, 1.0, 1.2 volts.
- Two values of substrate bias, V_p: 0.4, 0.45 volts.

For each bit, 60 retention times in five groups of 12 were measured as follows:

- 12 retention times, r, were tested, increasing from 60 au to 604 au in steps of 49.5 au: $r = 10 + i \times 49.5$ au, $i = 1$ to 12, with pass/fail determined at each test stop, i. Retention times are given in arbitrary units (au), related to the true retention times by a numerical ratio.
- Each group of 12 retention times was repeated five times. Groups were separated by variable durations, typically many hours.

Fig. 1 shows how the maximum observed retention time index $i_{\max}$ and the minimum observed retention time index $i_{\min}$ were extracted from the pass/fail pattern of each failing bit. If $i_{\max} - i_{\min} > 2$ the bit is classified as VRT, otherwise it is SRT. A difference of 2 eliminates tester quantization effects which might misclassify SRT bits as VRT bits, at the risk of classifying some less variable VRT bits as SRT bits. If the leftmost retention time index in any group is 1 (fail) the bit is "dead" and is excluded from the study.

2.2 Experimental Results

Table 1 summarizes failing bit counts sampled from 49 Mb across all environmental conditions. Six bits were found to be dead and were excluded from analysis. At each environmental condition "live" bits passing at least the first test stop

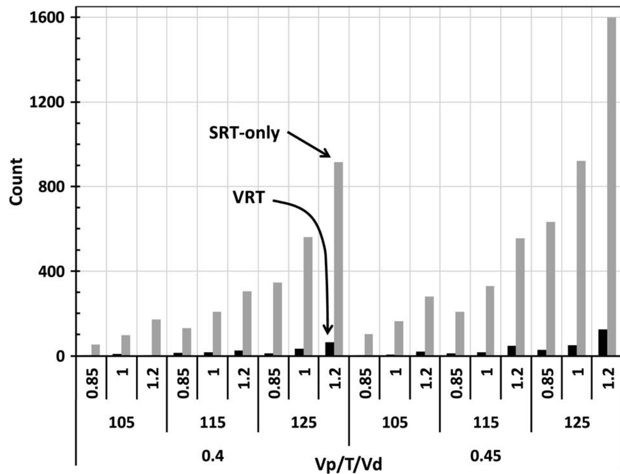


Fig. 2. VRT and SRT-only counts vs. environmental condition sampled from $10 \times 4 \times 1,218,750 = 48,750,000$ bits.

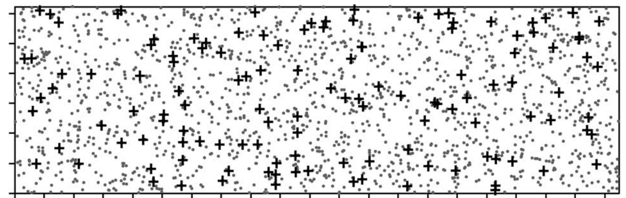


Fig. 3. Map of spatial xy locations of all bits with $r < 604\text{au}$ from 4 arrays on 10 chips, sampling 49 Mb. VRT bits are shown as $+$.

and with $r < 604$ au are classified as SRT or VRT. A given bit may be classified differently in different environmental conditions. For example, 1610 failing bits showed only SRT behavior, and 288 bits showed SRT behavior in at least one environmental condition and VRT behavior in at least one environmental condition. The total number of bits with $r < 604$ au observed to fail in the sample of 48,750,000 bits (minus 6 dead bits) was $\mathrm{A} + \mathrm{B} + \mathrm{C} = 1962$.

Fig. 2 shows bit categories by environmental condition, and Fig. 3 gives a spatial map of the xy location of failing bits. Important observations are:

- At the highest stress, retention times less than 604 au were observed for only a small fraction (40 PPM) of the population of bits. At less stressful environmental conditions, the fraction is smaller. These bits are representative of the "tail distribution" of retention times observed by White et al. [12].
- Statistical analysis of the spatial distribution of failing bits in Fig. 3 shows no evidence of clustering. The distribution of bit failures from die to die and array-to-array within dies is also indistinguishable from random. So the experiment may be regarded as sampling 48,750,000 independent bits.
- Yield loss for a 1 Mb array with 1.2 DPPM of bits defective, corresponding to the lowest environmental condition in the experiment, is 72%. Since array sizes of 1 Mb and larger are generally used in applications, this shows that fault tolerance is required for any product array made from the bits studied here.
- 18% of the bits with $r < 604$ au, $(\mathrm{B} + \mathrm{C})/(\mathrm{A} + \mathrm{B} + \mathrm{C})$, show VRT behavior. This shows that VRT behavior must be included in any statistical model of DRAM retention time.

Each failing bit has a minimum and a maximum retention time. For stable bits, these retention times are equal. The fraction of time that an unstable bit is in the maximum versus minimum retention time state could not be empirically determined because the DRAM arrays in the test chips were indirectly accessed through a BIST controller which gives only pass/fail for a given refresh time. So, to construct a model of Test/Use correlation from the data it is necessary to additionally specify how retention time is manifested in Test and in Use. The manifestation will be different in Test and Use because in Use a given bit will be accessed an indefinite number of times and the minimum retention time will certainly occur, whereas Test is a single brief measurement for which the maximum or minimum retention time occurs with probability depending on the fraction of time-in-state.

Model-fitting is simplified by displaying the retention time data in a way that is different from any plausible Test/Use model with the understanding that, for decision-making, the data or fitted model will be transformed later into a plausible Test/Use model. The data display in Table 2 is constructed by assigning $(r_{\max}, r_{\min})$ to (r_1, r_2) or (r_2, r_1) with equal probability. Similar tables were generated for the 17 other environmental conditions. Several observations can be made: 1) The display of r_1 and r_2 in Table 2 does not represent a plausible Test/Use scenario of sequentially observed retention times because it does not include the possibilities of $(r_1, r_2) = (r_{\min}, r_{\min})$ and $(r_1, r_2) = (r_{\max}, r_{\max})$. 2) Fitting a model to the data in Table 2 is greatly simplified because only exchangeable copulas (symmetrical about $r_1 = r_2$) and a single marginal distribution (the same for r_1 and r_2) need to be fitted due to symmetry of the data. 3) The data in Table 2 and any model fitted to it will be transformed later into a plausible Test/Use model via Eqs. (14) or (15).

The marginal empirical cumulative distributions (F as a function of r_1 or r_2) given in Table 2 were fitted to a single Weibull distribution following Lieneweg et al. [13] and White et al. [12], as shown in Fig. 4. The slope and intercept of the fitted lines give the shape, β, and scale, α, parameters of the Weibull distribution of retention time:

$$F(r) = 1 - \exp\left[-\left(\frac{r}{\alpha}\right)^{\beta}\right]. \quad (1)$$

Fits like Fig. 4 were done for every environmental condition. The shape parameter β was always nearly 2, so the model was simplified by forcing β to 2. $\beta = 2$ results in a small underestimate of the retention time at short retention times, which is conservative. Arrhenius temperature and exponential voltage dependence gave an excellent fit (Fig. 5) to the scale parameters, α, extracted from all environmental conditions:

$$\ln \alpha = \ln \alpha_0 + a(V_p - V_{p0}) + b(V_d - V_{d0}) + \frac{Q}{k_B}\left(\frac{1}{T} - \frac{1}{T_0}\right). \quad (2)$$

The good fit shown in Fig. 5 means that the single parameter, ln α, computed from Eq. (2) is a measure of the combined effect of V_d, V_p, and T, simplifying the display of the environmental dependencies observed in this study. The leftmost point in Fig. 5 (smallest α) corresponds to Fig. 4 and Table 2.

TABLE 2
Maximum and Minimum Retention Times at the Highest Environmental Condition (Rightmost Bars in Fig. 2) Binned into Cells Using a "Symmetrical" Method of Displaying the Data (See Text)

T = 125 C
Vp = 0.45 V
Vd = 1.2 V
SS = 48750000

r_2: F = Cum N/SS (PPM)	Cum N	N	r (au)														
			604	0	0	0	0	1	0	1	1	1	1	18	69	N/A	
33.6	1639	273	555	0	0	0	0	0	0	0	0	1	14	104	95	59	
28.0	1366	343	505	0	0	1	0	0	0	1	2	10	97	92	120	20	
21.0	1023	237	456	0	0	0	0	0	0	1	5	46	83	86	14	2	
16.1	786	211	406	0	0	0	0	0	1	4	59	80	56	6	2	3	
11.8	575	156	357	0	0	0	0	0	1	29	68	53	2	1	2	0	
8.6	419	134	307	0	0	0	0	0	36	56	38	2	1	1	0	0	
5.8	285	109	258	0	0	0	0	15	71	20	2	1	0	0	0	0	
3.6	176	69	208	0	0	0	12	38	18	0	0	0	1	0	0	0	
2.2	107	61	159	0	0	8	43	9	0	0	0	0	0	0	0	1	
0.9	46	27	109	0	1	19	6	0	0	0	0	1	0	0	0	0	
0.4	19	13	60	0	11	2	0	0	0	0	0	0	0	0	0	0	
0.1	6	6	0	5	1	0	0	0	0	0	0	0	0	0	0	0	
				0	60	109	159	208	258	307	357	406	456	505	555	604	r_1: r (au)
				5	13	30	61	63	127	112	175	195	255	308	302		N
				5	18	48	109	172	299	411	586	781	1036	1344	1646		Cum N
				0.1	0.4	1.0	2.2	3.5	6.1	8.4	12.0	16.0	21.3	27.6	33.8		F=Cum N/SS (PPM)

A given value of ln α defines a surface (nearly a plane) of "statistically equivalent" test set points in (V_p, V_d, T) space, providing useful flexibility when integrating different kinds of test into a test program.

The distribution of data across r_1 and r_2 cells in Table 2, and at 17 other environmental conditions was characterized by Kendall's tau. Suppose the retention times, r_1 and r_2, are known exactly for each of n measured bits so that every bit may be ranked by r_1 and by r_2 without ties. The number of bit-pairs, $n(n-1)/2$ comprises k "concordant" pairs and d "discordant" pairs. For a concordant pair, the relative ranks of r_1 for a bit pair is the same as the relative ranks of r_2 of the bit pair. For a discordant pair the relative ranks are opposite. Kendall's tau for the sample (indicated by the prime) is $\tau' = (k-d)/(k+d)$.

Test data is typically binned into cells as in Table 2 so that Kendall's tau must be calculated for data with many ties. A method for computing τ' from data with ties is given in [14] as:

$$\tau' = \frac{k-d}{\sqrt{\frac{1}{2}n(n-1)-U}\sqrt{\frac{1}{2}n(n-1)-V}},$$
$$U = \frac{1}{2}\sum u(u-1), \quad V = \frac{1}{2}\sum v(v-1), \tag{3}$$

where any bit pairs that are tied in r_1 or r_2 are not counted in k or d and where, in U, u is the number of tied r_1 values in each set. V is defined in the same way, but for r_2 values. Code to compute τ' from data with ties is available in many statistical software packages. τ' for the data in Table 2, and 17 other environmental conditions was computed by Eq. (3) and plotted in Fig. 6. Also plotted in Fig. 6 is the fraction of the population sampled at each environmental condition,

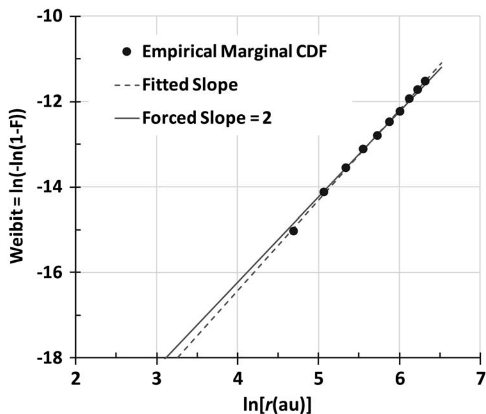


Fig. 4. Weibull fit of marginal distributions at the highest environmental condition, from Table 2.

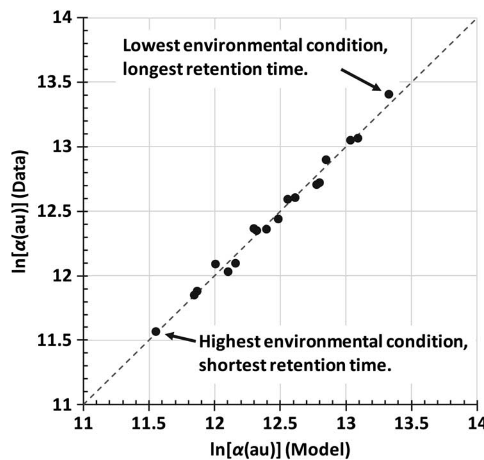


Fig. 5. Extracted vs. model-fitted scale parameter α, Eq. (2), for all 18 environmental conditions.

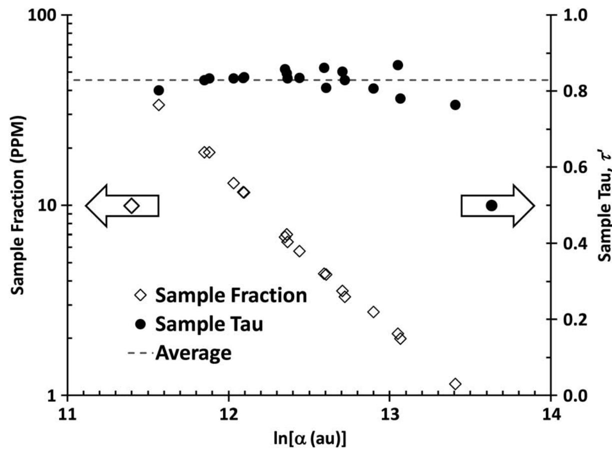


Fig. 6. Sample tau, τ', is independent of sample fraction and environmental conditions (ln α).

ranging from 35 PPM at the highest stress (smallest α) to 1.2 PPM at the lowest stress. A remarkably constant value of $\tau' = 0.828$, independent of environmental condition, is observed. If only the diagonal cells in Table 2 were populated, the value of τ' would be unity.

Extracted parameters of the marginal scaling model, Eq. (2), are given in Table 3 along with parameters describing the correlation (aka "Dependence") including copula parameters described in the following Section.

TABLE 3
Parameters of Extracted Marginal and Dependence Models

Margin	β	2.0
	ln[a_0 (au)]	11.57
	a (V^{-1})	-5.79
	b (V^{-1})	-1.55
	Q (eV)	0.605
	V_{p0} (V)	0.45
	V_{d0} (V)	1.2
	T_0 (°C)	125.0
Dependence	Sample Tau, τ'	0.828
	Clayton Copula, θ	9.74
	Gaussian Copula (1-ρ)x10^3	0.695

3 Modeling Dependence Using Copulas

3.1 Copula Background

If the cell-counts in Table 2 are divided by the sample size to give the probability mass in the cell, then the table is an empirical 2-dimensional probability density function (pdf) sampling a population pdf $h(r_1, r_2)$. The corresponding bivariate cumulative distribution function (cdf) is

$$H(r_1, r_2) = \int_0^{r_1} dx \int_0^{r_2} dy h(x, y). \tag{4}$$

For marginal distributions $F(r_1) = H(r_1, \infty)$, and $G(r_2) = H(\infty, r_2)$ (in the present application, $F = G$), the definition of a copula C is given by H written as a function of the marginal distributions

$$H(r_1, r_2) = C[F(r_1), G(r_2)], \tag{5}$$

or

$$C(u, v) = H\left[F^{-1}(u), G^{-1}(v)\right]. \tag{6}$$

A two-dimensional copula is a function on the unit square domain with range [0,1], which:

- Is grounded. $C(u, 0) = 0 = C(0, v)$.
- Is normalized. $C(1, 1) = 1$.
- Has uniform marginal distributions. $C(u, 1) = u$, $C(1, v) = v$.
- Is 2-increasing, so that for every u_1, u_2, v_1, v_2 in [0,1] such that $u_1 \leq u_2$ and $v_1 \leq v_2$ the probability mass in the rectangular area defined by (u_1, v_1) and (u_2, v_2) is positive definite

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

The definitions have generalizations to more than two dimensions.

Sklar showed that for a given H, the copula C is unique. And Schweizer and Wolff showed that C is invariant under monotonic transformations of F and G. This history, and more, is covered by Nelsen [10]. These results are profound because they imply that C contains *all* of the rank-dependency information in *any* multivariate cdf, and that the study and modeling of this dependency can be completely decoupled from details of the marginal distributions.

Two copulas are especially important:

$$C(u, v) = \begin{cases} \min[u, v], & \text{Perfect correlation,} \\ uv, & \text{Independence.} \end{cases} \tag{7}$$

Also important is the definition of tail dependency in the low tail

$$LT = \lim_{u \to 0+} \frac{C(u, u)}{u}. \tag{8}$$

Notice that LT for perfect correlation is unity, whereas for independence LT vanishes.

The population Kendall's tau may be calculated analytically from a copula by

$$\tau = 4 \int_0^1 \int_0^1 C(u, v) \frac{\partial^2 C(u, v)}{\partial u \partial v} du dv - 1. \tag{9}$$

For perfect correlation $\tau = 1$, for perfect anti-correlation $\tau = -1$, and for independence $\tau = 0$. Equation (9) can be generalized to compute τ for truncated regions of the copula [15].

Copulas come in families spanned by adjustable parameters, so Eqs. (8) and (9) provide a way to fit the model parameters to data. For example, if the empirical (sample) value of τ' is known from Eq. (3), then the value of the parameter of a single-parameter copula may be determined by comparison with τ computed from Eq. (9) integrated over a truncated region of the model copula corresponding to the data sample.

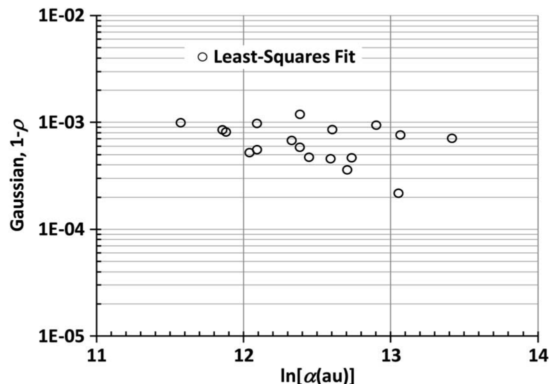


Fig. 7. Values of $1-\rho$ for the Gaussian copula fitted by least squares to data like Table 2, as a function of environmental condition.

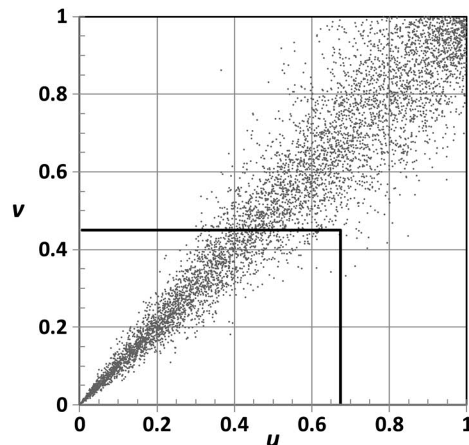


Fig. 8. Probability density map for Clayton copula with $\theta = 9.74$, and example rectangular truncation.

Copulas give a complete generalization of the usual multivariate normal approach for modeling statistical dependency. But this leads to the main problem of copula modeling; choosing the appropriate copula. The universe of possible functions even in two dimensions is vast, so some application-specific guidance is needed. In the course of this work several kinds of copula were tried for the DRAM; the Gaussian copula, geometrical copulas, various kinds of Archimedean copulas, the Marshall-Olkin copula, and convex combinations of various copulas [10]. Just two of the fitting attempts are described in the following; the Gaussian copula, and the Clayton copula (an Archimedean copula). The Gaussian copula is described because it is the conventional multivariate normal modeling approach in copula guise and therefore shows the problem with conventional multivariate normal modeling. The Clayton copula is shown because it is the best model found, and was used in subsequent application of the model.

3.2 The Gaussian Copula

In two dimensions, the Gaussian copula is

$$C(u, v; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \times \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right] dy dx, \tag{10}$$

where Φ is the standard normal distribution. Numerical methods to compute bivariate and trivariate integrals like Eq. (10) are available [16]. The Gaussian copula was fitted to the data of Table 2 by finding ρ which minimizes the sum of squares:

$$SSQ(\rho) = \sum_i \sum_j \left(N \times \delta C_{ij}(\rho) - n_{ij}\right)^2 \Big/ \sum_i \sum_j n_{ij}^2, \tag{11}$$

where i and j are cell indexes in Table 2 and

$$\delta C_{ij} = C_{ij} - C_{i-1,j} - C_{i,j-1} + C_{i-1,j-1}. \tag{12}$$

δC_{ij} is the probability mass in cell ij with C_{ij} computed via (10) at each cell ij, n_{ij} is the count in a cell, and $N = 48,750,000$. This was repeated for each of the 18 environmental conditions and the fitted values of $1-\rho$ were plotted vs environmental condition in Fig. 7.

The value of ρ must be forced to within a few parts in 10,000 of unity to fit the observed data which are the deep tail (1.2 to 35 PPM) of the bit population. The tiny value of $1-\rho$ shows a key problem with multivariate normal modeling and with the Gaussian copula which has been recognized in other fields [9]. As one moves from the bulk of the population into the tails the correlation in the Gaussian copula fades away unless ρ is exactly unity. That is, for the Gaussian copula, $LT = 0$ except when $\rho \equiv 1$. The Gaussian copula's tail dependency may be valid for intrinsic properties of devices, but for defect-related mechanisms such as the retention time mechanisms of the DRAM one would expect that dependency would be maintained no matter how far into the tail the sample is taken. That is, one would prefer a copula for which $LT \neq 0$. The significant scatter in $1-\rho$ as a function of environmental conditions in Fig. 7 also shows that the Gaussian copula is not a "natural" fit to the data. The average of $1-\rho$ across environmental conditions is given in Table 3 above.

3.3 The Clayton Copula

The Clayton copula (see [10], p116) for the range of θ of interest is

$$C(u, v; \theta) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta} \quad (0 < \theta < \infty), \tag{13}$$

where $\theta \to 0$ corresponds to independence, and $\theta \to \infty$ corresponds to perfect correlation. The probability density map corresponding to Eq. (13) for the parameter value $\theta = 9.74$ which fits the DRAM data is shown by the density of (u, v) points in Fig. 8 synthesized by standard methods described in the appendix of [11] and in [17]. Properties of the Clayton copula include the low-tail dependence, $LT = 2^{-1/\theta}$. and Kendall's tau for the entire probability space, $\tau = \theta/(\theta + 2)$ derived using Eqs. (8) and (9), respectively, with Eq. (13). Recently Oakes [15] showed that the Clayton copula is the only absolutely continuous copula with a remarkable "truncation invariance" property: If the

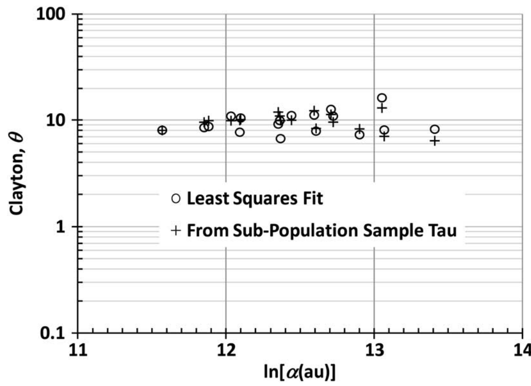


Fig. 9. Values of θ vs environmental condition (ln α) for the Clayton copula fitted by least squares to data like Table 2, and by $\theta = 2\tau/(1-\tau)$ from τ' in Fig. 6.

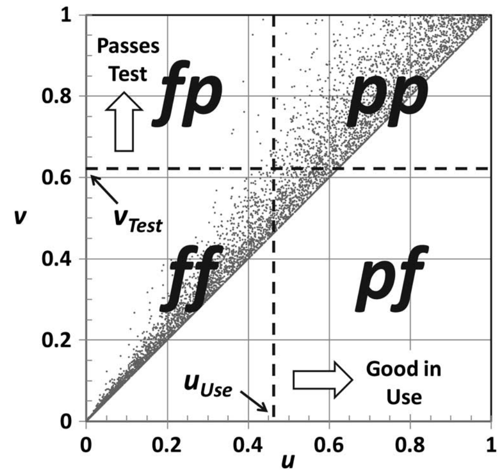


Fig. 10. Test and Use conditions divide the bit pseudo-copula D (with $\theta = 9.74$, $s = 0.8$) into four categories labeled by fp, pf, ff, pp, where the first character refers to Use and the second to Test.

probability map of any rectangular truncation of the copula with one corner pinned at (0,0) as shown in Fig. 8 is remapped to the entire copula domain, $[0,1]^2$, then the same Clayton copula (same θ) is recovered. A consequence of this is that τ computed from any rectangular truncation of the probability density map with one corner pinned at (0,0) is always $\theta/(\theta+2)$.

Truncation invariance of the DRAM data was seen in Fig. 6 because Kendall's tau for the failing bits remains constant as larger and larger samples of the 49 Mb population were exposed by increasing the test environmental condition (reducing ln α). The samples correspond to square truncations of the empirical copula with one corner fixed at (0,0), and the opposite corner at (x, x), where x varies with the environmental condition. Tau for the Clayton copula is also invariant as data is truncated by rectangles (and, *a fortiori*, squares) like the one shown in Fig. 8. So the Clayton copula is a plausible model for the observed dependency behavior. Occam's razor was used to choose the Clayton copula over others with small but non-vanishing truncation variation of τ. Fig. 9 shows θ for the Clayton copula fitted at each of the 18 environmental conditions by the same least squares method used for the Gaussian copula. Fig. 9 also shows θ determined from Fig. 6 using the inverse of the relation between τ and θ for the Clayton copula: $\theta = 2\tau/(1-\tau)$.

Monte-Carlo synthesis of random vectors (u_i, v_i) from a copula is often needed to "play back" a model to validate it, or to do numerical calculations when analytical calculations are intractable. The Clayton copula has a very useful property for Monte-Carlo simulation stemming from the truncation invariance. It is possible to synthesize points in a truncated region of the copula, such as shown in Fig. 8, *without rejection*. This feature of the Clayton copula and of certain other copulas is very important for efficient simulation because only the extreme tail of the distribution (40 PPM in the current example–a tiny area near the origin in Fig. 8) is of practical interest and needs to be synthesized. The Gaussian copula does not have this feature and requires extensive rejection to generate tail samples. This is another significant disadvantage of the conventional multivariate normal approach.

4 Application

4.1 Model of Test and Use

The symmetrically displayed bit data in Table 2 were fitted to an exchangeable copula (symmetrical in its arguments, $C(u,v) = C(v,u)$). Although the fitted exchangeable copula is not a plausible model of Test and Use, the exchangeable copula may be transformed into a pseudo-copula (having properties of a copula except for non-uniform margins) which is a plausible model of the way Test and Use are manifested. An advantage of this approach is that different Test/Use scenarios may be explored by varying the transformation of the fitted copula.

A plausible model of Test and Use manifestation is one in which the maximum retention time of a given bit is exhibited at Test with probability $s = t_{\max}/(t_{\max}+t_{\min})$ (and minimum retention time is exhibited at Test with probability $1-s$) while the minimum retention time for the bit is always exhibited in Use. If r_1 and r_2 are retention times sampled from the symmetrically displayed data of Table 2 or sampled by Monte-Carlo from the fitted symmetrical model (the Clayton copula) then the plausible Test/Use model is

$$\begin{aligned} r_{Use} &= \min[r_1, r_2] \quad = r_{\min} \qquad \text{All the time,} \\ r_{Test} &= \begin{cases} \max[r_1, r_2] = r_{\max}, & s \text{ of the time,} \\ \min[r_1, r_2] = r_{\min}, & 1-s \text{ of the time.} \end{cases} \end{aligned} \tag{14}$$

The probability density map obtained by using Eq. (14) with $s = 0.8$ to transform the symmetrical model Clayton copula density map in Fig. 8 is shown in Fig. 10.

Association of minimum retention time of a bit with Use is realistic because a bit will be accessed an indefinite number of times making it certain the minimum retention time will occur eventually. Test, however is a single brief event so association of the maximum or minimum retention time with Test depends on the probability, s, that the bit happens to be in the maximum retention time state when tested. The assumption that the bit is always in the maximum retention time state at Test ($s = 1$) is conservative from the customer perspective because a model based on this will over-estimate DPPM in

Use, and thereby lead to customer-conservative Test and Use specifications.

Using methods of Nelsen (problem 2.16 on p26 of [10]), and Navarro and Spizzichino [18], it can be shown that[1] since the marginal distributions for r_1 and r_2 are the same in the symmetrical model by construction, the transformation of (14) may be written as a transformation of the fitted exchangeable copula C into the pseudo-copula D:

$$D(u,v) = s[C(u,v) + C(v,z) - C(u,z)] + (1-s)[2z - C(z,z)], \quad z = \min[u,v]. \quad (15)$$

Fig. 10 shows the density map of D with $s = 0.8$ when C is the Clayton copula with $\theta = 9.74$ (Fig. 8).

4.2 Test Set Points and Datasheet Specifications

Test set points and datasheet specifications (Use conditions) are expressed in terms of environmental conditions (V_p, V_d, T) and retention time, r. These four parameters at Test and at Use are usually set so that the Test set point is more "stressful" (causes more failures) than the Use condition. The environmental conditions are mapped into α_{Test} in Test and α_{Use} in Use by (2). So a single parameter, u, depending on both the environmental condition via α_{Use} and retention time limit r_{Use} in the datasheet defines the datasheet specification (Use condition), and a single parameter, v, defines the Test set point:

$$u = 1 - \exp\left[-\left(\frac{r_{\text{Use}}}{\alpha_{\text{Use}}}\right)^{\beta}\right], \quad v = 1 - \exp\left[-\left(\frac{r_{\text{Test}}}{\alpha_{\text{Test}}}\right)^{\beta}\right]. \quad (16)$$

When the datasheet specification and Test set point u and v are superimposed on the bit pseudo-copula, D, as in Fig. 10, the probability masses associated with each of the four regions are given by

$$p_{fp} = D(u,1) - D(u,v) \quad p_{pf} = D(1,v) - D(u,v)$$
$$p_{ff} = D(u,v) \quad p_{pp} = 1 - p_{fp} - p_{pf} - p_{ff}, \quad (17)$$

where, for example, p_{fp} is the fraction of bits failing in Use and passing in Test.

4.3 Array Statistics and Fault Tolerance[1]

The random spatial distribution of bit failures (Fig. 3), the large sample size of bits (49 Mb), and the small probability of failure (1 to 40 PPM, depending on environmental condition) easily justifies use of the Poisson approximation to model the statistics of arrays of bits. Consider an array of n bits. The probability that the array has exactly n_{fp}, n_{pf}, and n_{ff} bits in the mutually exclusive categories defined in Fig. 10 is

$$P(N_{fp} = n_{fp}, N_{pf} = n_{pf}, N_{ff} = n_{ff}) = \frac{\lambda_{fp}^{n_{fp}} \exp(-\lambda_{fp})}{n_{fp}!} \frac{\lambda_{pf}^{n_{pf}} \exp(-\lambda_{pf})}{n_{pf}!} \frac{\lambda_{ff}^{n_{ff}} \exp(-\lambda_{ff})}{n_{ff}!}, \quad (18)$$

where

$$\lambda_{fp} = np_{fp}, \quad \lambda_{pf} = np_{pf}, \quad \lambda_{ff} = np_{ff}. \quad (19)$$

1. Derivation of Eq. (15) and details of manipulations in Section 4.3 are given in supplementary material available online at http://doi.ieeecomputersociety.org/10.1109/TC.2013.129.

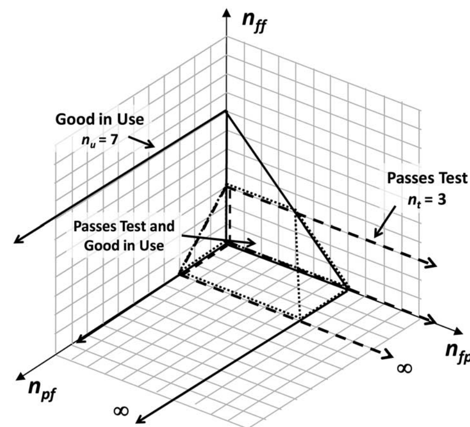


Fig. 11. Zones of bit category space corresponding to three array category probabilities for bad bits tolerated but not repaired at Test, and tolerated in Use. For $n_u = 7$ and $n_t = 3$.

Fault tolerance is modeled by expanding the definition of a "good" array to include arrays with some "bad" bits. Bad bits in arrays that are considered good are taken to be covered by a fault tolerance scheme. The maximum number of bad bits which can be tolerated is a measure of the capacity of the fault tolerance scheme. Suppose an array can tolerate up to n_t bits bad in Test and up to n_u bits bad in Use. Also suppose that the bits tolerated in Test are not repaired, but are included in the bad bits tolerated in Use. Then the probability that the array "Passes Test" is the sum of Eq. (18) over sets of integers n_{fp}, n_{pf} and n_{ff} allowed by the constraint $n_{ff} + n_{pf} \le n_t$ (n_{fp} is unconstrained). And the probability the array is "Good in Use" is a sum constrained by $n_{ff} + n_{fp} \le n_u$ (n_{pf} is unconstrained). The probability that an array "Passes Test and is Good in Use" is a sum over values of n_{fp}, n_{pf} and n_{ff} which satisfy both $n_{ff} + n_{pf} \le n_t$ and $n_{ff} + n_{fp} \le n_u$. A geometrical interpretation of the regions of bit category index space corresponding to three array categories is shown in Fig. 11.

Analytical expressions for the sums over terms like Eq. (18) corresponding to the zones in Fig. 11 are expressible in terms of the bivariate correlated Poisson distribution introduced by Campbell [19], derived as follows: If an array has exactly n_u bits which are bad in Use and exactly n_t bits which are bad at Test, then n_{fp}, n_{pf} and n_{ff} may vary within the following constraints:

$$n_u = n_{ff} + n_{fp} \quad n_t = n_{ff} + n_{pf} \quad 0 \le n_{ff} \le \min[n_u, n_t], \quad (20)$$

where the last inequality is a way of expressing the constraints $n_{pf} \ge 0$ and $n_{fp} \ge 0$. So if N_u and N_t are random variables giving the number of failing bits in Use and Test respectively, the probability that an array has exactly n_u bits failing in Use and exactly n_t bits failing in Test is the sum of Eq. (18) over values of n_{ff} allowed by Eq. (20):

$$P(N_u = n_u, N_t = n_t) = e^{-(\lambda_{fp}+\lambda_{pf}+\lambda_{ff})} \sum_{n_{ff}=0}^{\min[n_u,n_t]} \frac{\lambda_{fp}^{n_u - n_{ff}} \lambda_{pf}^{n_t - n_{ff}} \lambda_{ff}^{n_{ff}}}{(n_u - n_{ff})!(n_t - n_{ff})! n_{ff}!} \equiv \text{pois}(n_u, n_t; \lambda_{fp}, \lambda_{pf}, \lambda_{ff}), \quad (21)$$

and the cumulative form of this distribution is

$$\begin{aligned}P(N_u \le n_u, N_t \le n_t) &= \sum_{m=0}^{n_u}\sum_{n=0}^{n_t} \text{pois}\big(m, n; \lambda_{fp}, \lambda_{pf}, \lambda_{ff}\big) \\ &= \sum_{i=0}^{\min[n_u,n_t]} \frac{\lambda_{ff}^i e^{-\lambda_{ff}}}{i!} R\big(\lambda_{fp}, n_u - i\big) R\big(\lambda_{pf}, n_t - i\big) \\ &\equiv \text{Pois}\big(n_u, n_t; \lambda_{fp}, \lambda_{pf}, \lambda_{ff}\big),\end{aligned} \quad (22)$$

where R is the univariate cumulative Poisson distribution available in many software packages:

$$R(x, n) \equiv e^{-x} \sum_{0 \le i \le n} \frac{x^i}{i!}. \quad (23)$$

which vanishes when $n < 0$. n_u and n_t are usually small integers so calculation of the cumulative bivariate Poisson distribution using the second equality in Eq. (22) is easy. Equations (21) and (22) are Campbell's [19] bivariate correlated Poisson distribution. Johnson et al. [20] point out that (21) is the distribution of

$$N_u = N_{fp} + N_{ff} \quad N_t = N_{pf} + N_{ff}, \quad (24)$$

where N_{fp}, N_{pf}, and N_{ff} are mutually independent Poisson random variables with means λ_{fp}, λ_{pf}, and λ_{ff}. On the margins (that is, with $n_t = \infty$ or $n_u = \infty$), N_u and N_t have Poisson distributions with means $\lambda_{ff} + \lambda_{fp}$ and $\lambda_{ff} + \lambda_{pf}$, respectively.

The mapping of proportions of three Test/Use categories of *bits* into proportions of three Test/Use categories of *arrays* with specified fault tolerance is given by Campbell's correlated Poisson distribution, Eq. (22). If an array can tolerate up to n_u bad bits in Use and up to n_t bad bits in Test then the probability that the array "Passes Test and is Good in Use" corresponds to sums over bit category indexes in the intersection of the two infinite prisms along the n_{pf} and n_{fp} axes in Fig. 11, and is

$$\begin{aligned}&P(\text{Passes Test and Good in Use}) \\ &\quad = \text{Pois}\big(n_u, n_t; \lambda_{fp}, \lambda_{pf}, \lambda_{ff}\big).\end{aligned} \quad (25)$$

The probability that the array tolerates n_u bits in Use, irrespective of the number of bad bits in Test the sum of (18) over bit category indexes in the prism running down the n_{pf} axis in Fig. 11 and is

$$\begin{aligned}P(\text{Good in Use}) &= \text{Pois}\big(n_u, n_t = \infty; \lambda_{fp}, \lambda_{pf}, \lambda_{ff}\big) \\ &= R\big(\lambda_{ff} + \lambda_{fp}, n_u\big).\end{aligned} \quad (26)$$

The probability that the array tolerates n_t bits in Test, irrespective of the number of bad bits in Use corresponds to a sum of (18) over bit category indexes in the prism running down the n_{fp} axis in Fig. 11 and is

$$\begin{aligned}P(\text{Passes Test}) &= \text{Pois}\big(n_u = \infty, n_t; \lambda_{fp}, \lambda_{pf}, \lambda_{ff}\big) \\ &= R\big(\lambda_{ff} + \lambda_{pf}, n_t\big).\end{aligned} \quad (27)$$

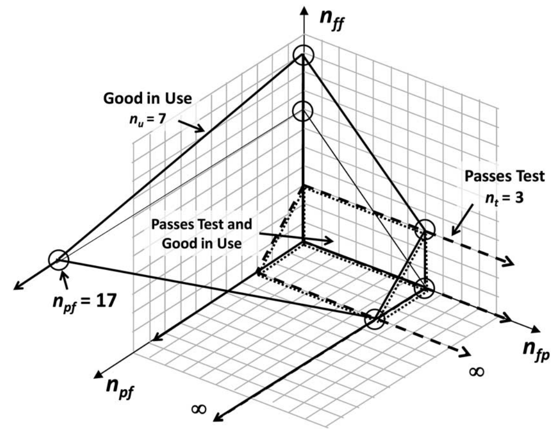


Fig. 12. Zones of bit category space corresponding to three array category probabilities for bad bits tolerated and repaired at Test, and tolerated in Use. For $n_u = 7$ and $n_t = 3$. Corners of the extra volume of "Good in Use" due to active repair at Test are shown by circles.

In practice Eqs. (25), (26), and (27) would be applied when the fault tolerance mechanism on the chip is enabled in both Test and Use ($n_t = n_u > 0$).

When the tester *actively repairs* the bits that it tolerates, the constraints on the integers n_{fp}, n_{pf} and n_{ff} allowed in the sums over terms like Eq. (18) are changed from the constraints shown in Fig. 11 to the constraints shown in Fig. 12. Repair of bits tolerated at Test causes the tolerance mechanisms in Use to have fewer ff category bits to tolerate. So, more tolerance capacity is available in Use for test escape bits (fp category bits) and ff bits exceeding the repair capacity of Test. The effect is seen in Fig. 12 as an extra volume of bit category index space on top of the "Good in Use" volume shown for the no-repair-at-Test case in Fig. 11. (Details are given in supplementary materials.[1])

The "Passes Test and Good in Use" volume is the intersection of the "Good in Use" and the "Passes Test" volumes, so its probability sum changes too. But the "Passes Test" volume is the same as for the no-repair-in-test case.

The probability expressions corresponding to the volumes in Fig. 12 are

$$\begin{aligned}P(\text{Good in Use}) &= L\big(\lambda_{ff}, \lambda_{fp}, \lambda_{pf}, n_u, n_t\big) \\ &\quad + R\big(\lambda_{ff} + \lambda_{fp}, n_u\big),\end{aligned} \quad (28)$$

where L is the sum of terms like Eq. (18) over the "extra" volume of "Good-in-Use" bit category space in Fig. 12. L does not have a tidy analytical expression, but it is easily evaluated because the number of terms in the "extra" volume is finite and small. The "Passes Test and Good in Use" volume in Fig. 12 is a simpler truncated prism than the corresponding shape in Fig. 11, giving

$$\begin{aligned}&P(\text{Passes Test and Good in Use}) \\ &\quad = R\big(\lambda_{fp}, n_u\big) R\big(\lambda_{ff} + \lambda_{pf}, n_t\big).\end{aligned} \quad (29)$$

And finally the "Passes Test" probability for active repair is the same as Eq. (27) for no-repair-at-test, as it must be because this proportion will be agnostic to the repair status of Test-tolerated bits.

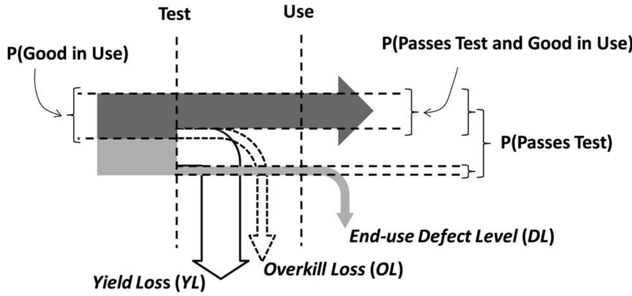


Fig. 13. Relationship of figures of merit (italic) to population category probabilities.

4.4 Figures of Merit and Decision-Making

Manufacturing and quality figures of merit (FOMs) can be expressed in terms of the three array probabilities, *P*(Passes Test), *P*(Good in Use), and *P*(Passes Test and Good in Use) derived in Section 4.3. The FOMs are required to meet target values to determine fault tolerance, test specifications, and datasheet specifications of the product. The FOMs are yield loss (*YL*), overkill loss (*OL*), and end-use defect level (*DL*). *YL* and *OL* are producer-oriented cost-related FOMs, and *DL* is a customer-oriented quality FOM. Fig. 13 shows how these FOMs are related to the probabilities derived in Section 4.3.

Yield Loss given by

$$YL = P(\text{Fails Test}) = 1 - P(\text{Passes Test}) \tag{30}$$

is the fraction of manufactured arrays rejected by Test. YL is a primary manufacturing indicator since it directly affects producer costs.

Overkill Loss given by

$$OL = P(\text{Good in Use}) - P(\text{Passes Test and Good in Use}) \tag{31}$$

is the fraction of manufactured arrays (a subset of YL) invalidly rejected by Test. OL affects the manufacturing cost charged to Test.

End Use Defect Level given by

$$\begin{aligned} DL &= P(\text{Fails in Use}|\text{Passes Test}) \\ &= 1 - P(\text{Good in Use}|\text{Passes Test}) \\ &= 1 - \frac{P(\text{Passes Test and Good in Use})}{P(\text{Passes Test})} \end{aligned} \tag{32}$$

is the customer-perceived proportion of defective arrays. It is the fraction of units classified as failing in Use, given that they have passed Test (a conditional probability). DL is a quality indicator since it affects the customer.

For business decision-making FOMs corresponding to hypothetical Test, Use (datasheet), and fault tolerance specifications are compared with targets. Only specifications meeting all three targets are acceptable. The FOMs defined here are designed to lie in the range [0,1] such that a larger value is less desirable. Therefore "target" values are regarded as the maximum acceptable values of the FOMs. Arbitrarily chosen typical targets for the product example shown below are $YL \leq 20\%, OL \leq 2\%$, and $DL \leq 200$ DPPM.

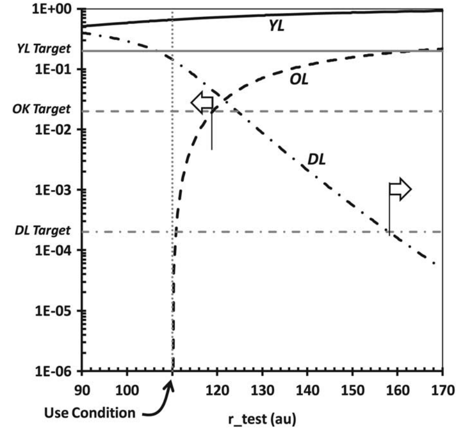


Fig. 14. Figures of merit vs. Test retention time set point for a 1 Mb array with no fault tolerance $m = 0$, assuming $s = 1$. There is no Test retention time setting for which all three FOM targets are met. Arrows show set point ranges which meet targets.

Equations (2), (13), (15), (16), (17), (19), (25) or (29), (26) or (28), (27), (30), (31), and (32) provide a fully deterministic analytical model readily implemented in, say, Excel to do "what-if" calculations of FOMs as a function of array size, fault tolerance, Test, and datasheet (Use) specifications. The sensitivity to models of Test/Use manifestation may be explored by adjusting parameters of the transformation, Eq. (15), of the fitted copula C into the pseudo-copula D embodying these models. Since the experiment did not give the fraction of time a VRT bit is in the long retention time state, the customer-conservative assumption used in examples described next is that at Test VRT bits are in the long retention time state all of the time ($s = 1$).

As an example, suppose the n-bit array has an internal mechanism, which can tolerate up to m bad bits, enabled in both Test and Use ($m = n_u = n_t$). Figs. 14 and 15 show FOMs computed using the model for a 1 Mb (2^{20} bits) array at the maximum environmental condition of the experiment ($V_p = 0.45$ V, $V_d = 1.2$ V, $T = 125°$C) for both Test and Use, with a datasheet (Use) refresh time specification of 110 au. The FOMs are plotted as a function of the Test retention time setting which is swept past the datasheet refresh time specification ("Use Condition").

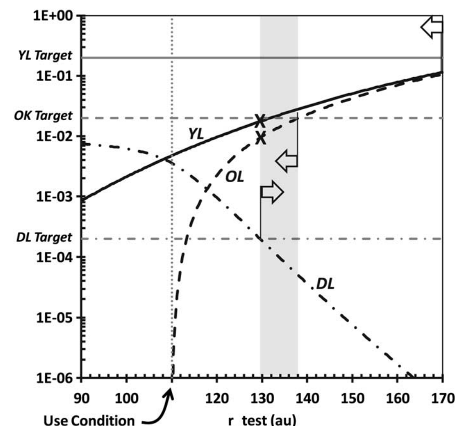


Fig. 15. Figures of merit vs Test retention time set point for a 1 Mb array which can tolerate $m = 4$ bad bits, assuming $s = 1$. All FOM targets can be met for Test retention time settings between 130 and 138 au. Arrows show set point ranges which meet targets.

The design of the FOMs allows all of the FOMs to be plotted together and to be compared with their target values. Fig. 14 shows that there is no setting for which a 1 Mb array with no fault tolerance ($m = 0$) can meet all targets. The minimum fault tolerance capacity for which targets can all be met is $m = 4$ bits and Fig. 15 shows this case. Notice the greatly improved yield loss in Fig. 15 compared to Fig. 14.

5 Discussion

When the rank statistics of the DRAM VRT effect is separated from complicated details of environmental dependence and shapes of marginal (Weibull) distributions an unexpected, yet simple, picture emerges. Unexpected because the usual method of fitting a bivariate normal distribution (or equivalently, Gaussian copula) *cannot* represent the invariance of tau under sample truncation as shown in Fig. 6. Simple, because a single parameter Clayton copula with a single value for the parameter *can* represent invariance of tau across *all* environmental conditions. More broadly, the DRAM case study shows the necessity of using copula methods to generalize the usual multivariate normal methods of statistical modeling of miscorrelation in semiconductor applications. The main challenge of copula methods is the need to choose a particular copula from the vast number of possibilities. For the DRAM the number of possibilities is greatly reduced because not many copulas have truncation invariance to a degree sufficient to match the data, and only the Clayton copula has complete truncation invariance.

Copula methods offer considerable practical convenience. The fitting of marginal models and copula models is decoupled, and can be done in any order. Moreover, many copulas are well-suited to synthesizing data in limited parts of the population without rejection of Monte-Carlo-generated samples. The Clayton copula is an example of such a copula. But the Gaussian copula (and therefore the multivariate normal distribution) cannot be synthesized in a limited part of the population without rejection. Rejectionless simulation makes Monte-Carlo simulation of the small but critical tail regions of interest in semiconductor applications highly efficient.

Key decisions at various stages of the product lifecycle require statistical models connecting device-level (bit-level for DRAM) to product-level cost and quality models. This paper provides all the machinery needed to do copula-based "what-if" analyses of effects of scaling of array size, fault tolerance (including active repair at test), datasheet (Use) specifications, and Test specifications. The statistical model may be "played forward" from bit-level to product-level to make product decisions based on computed figures of merit. But it may also be "played backwards" to discover data requirements at the device level (bit-level for DRAM) or silicon process level actually needed for product decision-making. This is important because device-level and process level characterizations can be expensive or, more significantly, time-consuming. For example, one may wish to understand the benefit of extraction of models of the RTN "duty cycle" time constants t_{max} and t_{min}. In the example of Fig. 15 $s = 1$ was used as a customer-conservative assumption to determine FOMs. Fig. 16 shows the effect of relaxing the assumption that Test always finds the bit in the maximum retention time state, that is, of allowing s to be less than unity. Yield loss and overkill loss figures of merit are shown for $s < 1$ at test settings for which all FOMs satisfy targets. The test settings which satisfy targets all have $DL = 200$ DPPM at the left-hand edge of the zone shown in Fig. 15 as the limiting constraint.

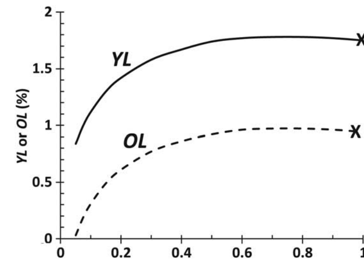


Fig. 16. Yield-loss and overkill at $DL = 200$ DPPM for the conditions of Fig. 15 as a function of hypothetical VRT duty cycle, s. Values shown by X at $s = 1$ correspond to X's in Fig. 15.

Fig. 16 shows that precise knowledge of s has no beneficial effect (reduction) on the overkill component of yield loss except for $s < 0.3$. But Kim et al. [21] found values of s varying from bit-to-bit within a memory, ranging from ~ 0.1 to nearly unity. And Kim et al. [8] found $s \cong 0.55$, weakly dependent on voltage and temperature (t_{max} and t_{min} individually depend more strongly on temperature). Unless $s < 0.3$ for all bits and all environmental conditions covered in Test and Use, there is no downside to setting $s = 1$ for all model calculations. One would conclude, at least for the DRAM case study here, that detailed knowledge of VRT duty cycle is not needed for product cost and quality decision-making.

An essential part of the method, not discussed in detail, is estimation of risks due to, a), sampling error and, b), model selection. Sampling error may be estimated using standard bootstrap methods. The semi-analytical form of the model facilitates the use of bootstrap methods [22]. When bootstrap methods are "played forward" confidence levels at which FOMs meet targets may be computed. And when "played backwards", aspects of the design-of-experiments for bit-level data acquisition (such as sample size) required to meet targets at specific confidence levels may be determined.

Model selection risk estimation requires evaluation of FOMs using copulas of various kinds fitted to the data. The decoupling of copula models from the marginal models makes it easy to "plug in" different copulas, recompute FOMs, and thereby quantify the risk of copula selection. For the DRAM example, truncation invariance was used to select the Clayton copula. But although the Clayton copula has absolute truncation invariance, another "geometrical" copula (not shown) which had truncation invariance to a degree sufficient to fit the data nearly as well was constructed. The shape of the FOM characteristics such as Figs. 14 and 15 is sensitive to whether the Clayton or the geometrical copula is chosen. The Clayton copula gave the more customer-conservative (larger value of r_{Test}) Test setting.

An often-overlooked requirement to balance producer and customer risk in integrated circuit test manufacturing is careful design of a *complete* set of FOMs and targets. Traditional fault tolerance modeling [23] assumes perfect correlation between Test and Use and focuses on only yield loss (YL) and sometimes the customer-perceived defectivity (DL). But

miscorrelation introduces another degree of freedom so that the three FOMs (*YL*, *DL*, and *OL*) discussed in this paper are needed for decision-making. It's also useful to design FOMs to cover [0,1] and map to "$\text{good} = 0$ and $\text{bad} = 1$" for the stakeholder most interested in the FOM. The problem is more complicated but same approach works for multiple test steps.

The DRAM VRT phenomenon does not fall neatly into classical notions of "hard" and "soft" reliability mechanisms. Since VRT bits can be stuck in a state for many minutes, or even hours, VRT bit errors are "soft" as far as Test is concerned, but "hard" as far as fault tolerance in Use is concerned. VRT is soft in Test because Test, being brief, cannot detect some bits which may fail in Use. But VRT is "hard" in Use because of the unacceptable performance effect of soft data correction of bits stuck in a failing state for extended durations. The VRT mechanism is very different from the classical [24] picture of foreign material particles causing clusters of bad bits which are either hard failures or latent reliability defects causing infant mortality that can be made to fail (permanently) and be screened by burn-in. The VRT phenomenon is also different from classical cosmic-ray soft-error mechanisms which cause only a momentary upset in Use. But the VRT mechanism is similar to RTN instabilities observed in other devices, such as SRAMs [25], [26]. Key characteristics of RTN-based mechanisms are random spatial distributions (Fig. 3 and [26]) and lack of memory in normal operation. However, stress can alter the properties of defects, changing the marginal distributions [12] (reducing α in our model) and increasing the miscorrelation [27] (reducing θ in our model).

6 Conclusions

There are several ways to extend the method described in this paper without introducing new concepts. First, more than one Test step increases the dimensionality of the copula and the multivariate mathematical manipulations. The increased dimensionality exacerbates the "copula choice" problem. Second, the method may be extended to multiple kinds of sub-elements (instead of bits) with differing critical areas, and multiple kinds of defects. Another extension is when the marginal variables are different, including different environmental dependence, such as Isb and Fmax. Yet another extension is to replace the Poisson model in Eq. (18) by a negative binomial model to describe "large-area" wafer-to-wafer, or lot-to-lot probability density variation [28].

An extension of the method which does require new concepts is exploration of principles governing the form of copulas in the semiconductor context in order to guide copula model selection. This is important because error associated with copula model selection is hard to gauge. A hint of the conceptual framework needed is seen in the way the copula model depends on how Test and Use are manifested, Eq. (15). And, one may ask, what is behind the remarkable truncation invariance seen in the DRAM data? The history of the development of the Clayton copula may provide a clue (see [15]).

Acknowledgments

Thanks are due to Satoshi Suzuki for acquiring the data, and to Roger Nelsen for pointing out the truncation-invariance of the Clayton copula described by Oakes. Support of this project by the Semiconductor Research Corporation under task numbers 1608.1 and 2095.1 is gratefully acknowledged.

References

[1] D. S. Yaney, C. Y. Lu, R. A. Kohler, M. J. Kelly, and J. T. Nelson, "A Meta-Stable Leakage Phenomenon in DRAM Charge Storage—Variable Hold Time," *Proc. Int'l Electron Devices Meeting (IEDM)*, vol. 33, pp. 336–339, http://dx.doi.org/10.1109/IEDM.1987.191425, 1987.

[2] P. J. Restle, J. W. Park, and B. F. Lloyd, "DRAM Variable Retention Time," *Proc. Int'l Electron Devices Meeting (IEDM)*, pp. 807–810, http://dx.doi.org/10.1109/IEDM.1992.307481, 1992.

[3] Y. Mori, K. Ohyu, K. Okonogi, and R.-I. Yamada, "The Origin of Variable Retention Time in DRAM," *Proc. Int'l Electron Devices Meeting (IEDM)*, pp. 1034–1037, http://dx.doi.org/10.1109/IEDM.2005.1609541, 2005.

[4] Y. Mori, H. Yoshimoto, K. Takeda, and R.-I. Yamada, "Mechanism of Random Telegraph Noise in Junction Leakage Current of Metal-Oxide-Semiconductor Field-Effect Transistor," *J. Applied Physics*, vol. 111, p. 104513, http://dx.doi.org/10.1063/1.4721658, 2012.

[5] T. Umeda et al., "Single Silicon Vacancy-Oxygen Complex Defect and Variable Retention Time Phenomenon in Dynamic Random Access Memories," *Applied Physics Letters*, vol. 88, p. 253504, http://dx.doi.org/10.1063/1.2213966, 2006.

[6] K. Ohyu et al., "Quantitative Identification for the Physical Origin of Variable Retention Time: A Vacancy-Oxygen Complex Defect Model," *Proc. Int'l Electron Devices Meeting (IEDM'06)*, pp. 1–4, http://dx.doi.org/10.1109/IEDM.2006.346792, 2006.

[7] B. D. Boatright, B. J. Eapen, C. G. Shirley, and C. Scafidi, "Methods and Apparatuses for Reducing Infant Mortality in Semiconductor Devices Utilizing Static Random Access Memory (SRAM)," U.S. Patent 7197670 B2, 27 Mar. 2007.

[8] H. Kim et al., "Characterization of the Variable Retention Time in Dynamic Random Access Memory," *IEEE Trans. Electron Devices*, vol. 58, no. 9, pp. 2952–2958, http://dx.doi.org/10.1109/TED.2011.2160066, Sept. 2011.

[9] P. Embrechts, A. McNeil, and D. Straumann, "Correlation: Pitfalls and Alternatives," *RISK Magazine*, pp. 69–71, http://www.math.ethz.ch/~embrecht/ftp/risk_pitt_alters_1999.pdf, May 1999.

[10] R. B. Nelsen, *An Introduction to Copulas*, 2nd ed. Springer, 2010.

[11] P. K. Trivedi and D. M. Zimmer, "Copula Modeling: An Introduction for Practitioners," *Foundations and Trends in Econometrics*, vol. 1, no. 1, pp. 1–111, http://dx.doi.org/10.1561/0800000005, 2005.

[12] M. White, J. Qin, and J. B. Bernstein, "A Study of Scaling Effects on DRAM Reliability," *Proc. Ann. Reliability and Maintainability Symp. (RAMS)*, pp. 1–6, http://dx.doi.org/10.1109/RAMS.2011.5754522, 2011.

[13] U. Lieneweg, D. N. Nguyen, and B. R. Blaes, "Assessment of DRAM Reliability from Retention Time Measurements," *Jet Propulsion Lab., California Inst. of Technol., Pasadena, Flight Readiness Technol. Assessment, NASA EEE Parts Prog.*, http://parts.jpl.nasa.gov/cots/external/dram_ret.pdf, 1998.

[14] M. G. Kendall and J. D. Gibbons, *Rank Correlation Methods*, 5th ed. USA: Oxford University Press, 1990.

[15] D. Oakes, "On the Preservation of Copula Structure under Truncation," *Canadian J. Statistics*, vol. 33, no. 3, pp. 465–468, http://www.jstor.org/stable/25046191?origin=JSTOR-pdf, Sept. 2005.

[16] A. Genz, "Numerical Computation of Rectangular Bivariate and Trivariate Normal and t Probabilities," *Statistics Computing*, vol. 14, pp. 251–260, http://dx.doi.org/10.1023/B:STCO.0000035304.20635.31, 2004.

[17] M. Armstrong, "Copula Catalogue. Part 1: Bivariate Archimedean Copulas," *CERNA—Centre d'économie industrielle*, http://www.cerna.ensmp.fr/Documents/MA-CopulaCatalogue.pdf, 2003.

[18] J. Navarro and F. Spizzichino, "On the Relationships Between Copulas of Order Statistics and Marginal Distributions,"*Statistics & Probability Letters*, vol. 80, no. 5–6, pp. 473–479, http://dx.doi.org/10.1016/j.spl.2009.11.025, Mar. 2010.

[19] J. T. Campbell, "The Poisson Correlation Function," *Proc. Edinburgh Math. Soc.*, vol. 4, no. 1, pp. 18–26, http://dx.doi.org/10.1017/S0013091500024135, Mar. 1934.

[20] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions*. John Wiley & Sons, 1997.

[21] H. Kim et al., "RTS-like Fluctuation in Gate Induced Drain Leakage current of Saddle-Fin Type DRAM cell transistor,"*Proc. IEEE Int'l Electron Devices Meeting (IEDM)*, pp. 1–4, http://dx.doi.org/10.1109/IEDM.2009.5424370, 2009.

[22] B. Efron and G. Gong, "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *Am. Statistician*, vol. 37, no. 1, pp. 36–48, http://www.jstor.org/stable/2685844, Feb. 1983.

[23] C. H. Stapper, A. N. McLaren, and M. Dreckman, "Yield Model for Productivity Optimization of VLSI Memory Chips with Redundancy and Partially Good Product," *IBM J. Research and Development*, vol. 24, no. 3, pp. 398–409, May 1980.

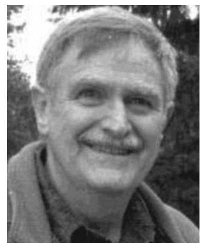
[24] C. H. Stapper, "Correlation Analysis of Particle Clusters on Integrated Circuit Wafers," *IBM J. Research and Development*, vol. 31, no. 6, pp. 641–650, Nov. 1987.

[25] M. Agostinelli et al., "Erratic Fluctuations of SRAM Cache Vmin at the 90 nm Process Technology Node," *Proc. IEEE Int'l Electron Devices Meeting, IEDM Technical Digest*, pp. 655–658, http://dx.doi.org/10.1109/IEDM.2005.1609436, 2005.

[26] K. Takeuchi et al., "Direct Observation of RTN-Induced SRAM Failure by Accelerated Testing and Its Application to Product Reliability Assessment," *Proc. Symp. VLSI Technology (VLSIT)*, pp. 189–190, http://dx.doi.org/10.1109/VLSIT.2010.5556222, 2010.

[27] Y. I. Kim, K. H. Yang, and W. S. Lee, "Thermal Degradation of DRAM Retention Time: Characterization and Improving Techniques," *Proc. IEEE Int'l Reliability Physics Symp.*, pp. 667–668, http://dx.doi.org/10.1109/RELPHY.2004.1315442, 2004.

[28] C. H. Stapper, "Large Area Fault-Clusters and Fault Tolerance in VLSI Circuits: A Review," *IBM J. Research and Development*, vol. 33, no. 2, pp. 162–172, Mar. 1989.



C. Glenn Shirley received the PhD degree in physics from Arizona State University, Tempe, AZ, and the MSc degree in physics from the University of Melbourne, Australia. He is a research professor with the Integrated Circuit Design and Test Laboratory in the ECE Department, Portland State University, Oregon. He retired in 2007 after 23 years at Intel, and a prior 10 years at Motorola,
U.S. Steel, and Carnegie-Mellon University (postdoctoral researcher). At Intel, he worked and published on package reliability, moisture reliability of silicon, accelerated moisture test hardware (HAST), and industry standards. He also led the development of Intel's burn-in methodology, founded a manufacturing test technology development Q&R group, and started a Q&R statistical modeling group. Subsequently, he co-directed as Intel's Q&R systems architect, a department responsible for Intel's quality systems. His current interests include yield, quality, and reliability statistical modeling of manufacturing test.

W. Robert Daasch is a professor of electrical and computer engineering at Portland State University, Oregon. He is the founder and co-director of the Portland State University Integrated Circuits Design and Test Laboratory. His current research interests are applying statistical methods to digital and analog integrated circuit design and test. He and his research team received the Semiconductor Research Corporation's Technical Excellence Research Award in 2008 for their research on statistical methods for burn-in reduction. In 2009, he received the Oregon Sigma Xi chapter's award for Engineering Researcher of the Year. He presented an invited plenary address at the International Test Conference in 2006. He was coordinator and co-author of the ITC Outstanding Lecture Series for 2005, first author on the International Test Conference Best Paper for 2001, co-author on the VLSI Test Symposium Best Paper for 2003, and first author on the ITC Honorable Mention Paper in 2004. He has three patents. He is a life member of Sigma XI.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.