Relating Yield Models to Burn-In Fall-Out in Time

Thomas S. Barnett
IBM Microelectronics
Logic Test Division
Essex Junction, Vt. 05452
email: tbarnet@us.ibm.com

Adit D. Singh
Electrical and Computer Engineering
Auburn University
Auburn, AL. 36849
email: adsingh@eng.auburn.edu

Abstract

An early-life reliability model is presented that allows wafer test information to be used to predict not only the total number of burn-in failures that occur for a given product, but also the time at which they occur during burn-in testing. The model is a novel extension of an experimentally verified yield-reliability model based on the fact that defects that cause early-life reliability (burn-in) failures are "smaller", more subtle versions of the defects that cause failures at wafer test. Consequently, knowledge of defect densities following wafer test (inferred from wafer probe failures) provides knowledge of the relative magnitude of early-life reliability defect densities. It is shown that this fact can be exploited to produce die with varying burn-in duration requirements. This is accomplished by sorting die into "bins" based on known reliability indicators. Presently, two such indicators are known: the local region yield of the die in question, and the number of repairs performed on the die in question. The early-life reliability model presented in this work will demonstrate that chips sorted based on these criterion have different fall-out or failure rate curves in burn-in. This information can be used to select optimal burn-in durations while maintaining outgoing product reliability.

1 Introduction

The significant cost associated with burn-in testing has forced researches in industry as well as in academia to seek reliability solutions that minimize the number of electronic components that must be subjected to burn-in testing. Significant cost reductions generally require limiting the number of burn-in tools a semiconductor manufacturer must purchase. Indeed, a set of burn-in boards (BIB), which hold the die in place during burn-in, can run into the $200,000 range, and high-end burn-in ovens, which regulate temperature and voltage during burn-in, can approach $1,000,000. Eliminating the need for even a single BIB set or oven can therefore result in significant financial savings. Of course, such reductions must address the implications with regard to product reliability. Intelligent burn-in strategies therefore seek to minimize the number of burn-in tools required while simultaneously maintaining outgoing reliability objectives.

The key to optimizing burn-in lies in identifying those die most likely to fail burn-in before burn-in is actually performed. Once identified, die of higher reliability risk may be subjected to more stringent stress testing, (e.g. longer burn-in durations), while those die deemed more reliable may see a reduced stress, or no stress at all. The chief difficulty in this procedure is, of course, in assigning a reliability or quality level to a die before burn-in. Fortunately, failures that occur at wafer test also give information about reliability. Indeed, it has been experimentally verified that defects that cause early-life reliability failures (i.e. burn-in failures) are fundamentally the same in nature as those defects that cause wafer probe failures, with defect size and placement distinguishing between the two [1, 2, 3, 4, 5]. Thus, indicators of high defect densities at wafer test also indicate relatively high reliability defect densities. Consequently, die that pass wafer probe testing, yet come from regions of relatively high defect density, are more likely to contain subtle early-life reliability defects than die from regions with lower defect densities.

Presently, two key reliability indicators have been identified: local region yield and, for chips containing redundancy, the number of defects that have been repaired. Local region yield is simply the yield of those chips in the vicinity of the die under question. One way to define local region yield is by examining the yield of a die's adjacent neighbors. Die can then be binned based on the number of faulty neighbors present. Die with 0 faulty neighbors go in bin 0, die with 1 faulty neighbor

go in bin 1, and so on, up to bin 8, where all neighbors are faulty. Extensions of this binning approach to include the 24 surrounding neighbors is, of course, possible as well. Because die in the higher numbered bins come from regions with a relatively large number of killer defects (i.e. defects that cause wafer probe failures), they should also be more likely to contain latent defects (i.e. defects causing early-life reliability failures). Indeed, this supposition has been experimentally verified through large independent studies conducted at Intel [2, 3] and IBM [4]. In both of these works it has been shown that die with many faulty neighbors can pose a significantly greater early-life reliability risk than chips with few faulty neighbors. Local region yield is therefore a strong indicator of die reliability.

In addition to local region yield, the number of repaired defects has also been shown to be an indicator of die reliability. This follows from the fact that the number of repairs performed is essentially a direct count of defects appearing at wafer test. Since latent defects are generally found near killer defects, chips that have been repaired are more likely to contain additional latent defects than chips with no repairs. Indeed, the more repairs performed on a chip, the more likely the chip is to contain a latent defect. This fact has been experimentally verified for SRAM as well as DRAM products manufactured by IBM Microelectronics in Burlington, Vermont [5].

The integrated yield-reliability model presented in [4, 5] allows one to extend current defect-based yield models to predict die reliability *following* burn-in (i.e. the number of burn-in failures). These models, however, have not addressed the issue of failures occuring during burn-in. It is the purpose of this work to provide such an extension. In particular, it will be shown that the yield-reliability model can predict not only the number of burn-in failures that occur, but also the time at which they occur during burn-in. It will be shown that chip populations with different reliability indicators, (i.e. local region yield or repair count), have different failure rates during burn-in. For example, die with 10 repairs have a different failure rate than die with 0 repairs. Similarly, die with 8 faulty neighbors have a different failure rate than die with 0 bad neighbors. The significance of the present work lies in the fact that the yield-reliability model allows one to quantify the difference between these failure rates using information obtained from wafer probe testing. This information then allows semiconductor manufacturers to optimize burn-in durations for a given reliability requirement. Such an optimization procedure can often result in a significant reduction in capital expenditures (i.e. BIBs and/or burn-in ovens).

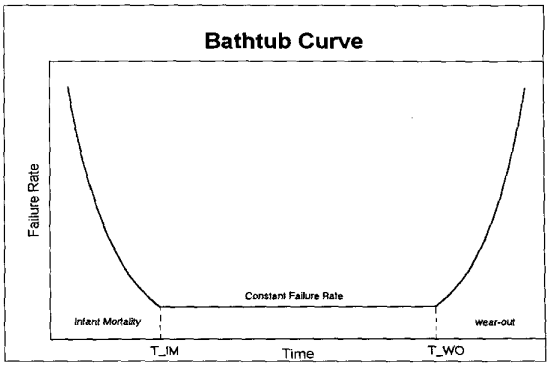

Figure 1: Ideal Bathtub curve. The infant mortality or early-life period goes up to time T_IM. Times where $T_IM < T < T_{WO}$ fall into the useful life region. Times greater than T_{WO} correspond to the wear out period.

2 Reliability Metrics and the Bathtub Curve

The reliability function, $R(t)$, denotes the probability that a chip survives on the time interval $[0, t)$. Conversely, the cumulative failure probability, $F(t)$, is defined as the probability that a chip fails on the interval $[0, t]$. Since a chip must either survive or fail on a given interval, it must be that $R(t) + F(t) = 1$ for all values of t. Thus, knowledge of $R(t)$ implies knowledge of $F(t)$ and vice versa. Both of these functions will be used in the following sections; the choice is simply one of convenience.

Closely related to reliability is the failure rate or hazard function. This is denoted by $h(t)$ and is defined such that $h(t)\Delta t$ is the probability a chip fails on $[t, t+\Delta t]$, given that it has survived until t. Mathematically, this can be written as

$$h(t) = \frac{-1}{R(t)}\frac{dR(t)}{dt} = -\frac{d\ln R(t)}{dt} \qquad (1)$$

The failure rate for many systems follows what is known as the bathtub curve. This curve is shown in Figure 1. The bathtub curve is characterized by three distinct regions. The first region is known as the early-life or infant mortality period. In this region the failure rate is high, corresponding to the failure of the "weak" members of the overall population. For integrated circuits this region is associated with the vast majority of reliability failures. Integrated circuit failures occuring at this stage are the result of flaws acquired during manufacturing. It is the purpose of stress tests such as

burn-in to precipitate these early-life failures before the product is shipped, and thereby maximize reliability in the field.

At the end of the infant mortality period, when most of the defective parts have failed, the failure rate decreases slowly, reaching a fairly constant value. This region corresponds to the operational life of the product. Integrated circuits appearing in applications should be operating in this portion of the failure rate curve. Finally, following the constant portion of the bathtub curve, the product enters the wearout stage. For integrated circuits, where the operational life is expected to be approximately 40 years [6], few products will be in use at the beginning of the wearout phase. In what follows, the early-life portion of the bathtub curve will be under consideration; no attempt is made to address the other portions of the bathtub curve.

3 Application to the Reliability of Integrated Circuits

The yield-reliability model uses wafer test information to estimate the number of latent manufacturing defects in a given population of die. These defects are precisely the ones causing early-life or infant mortality failures. This suggests that the yield-reliability model can be further extended to predict not only the number of latent defects, but also the time at which these failures occur. The following subsections, in conjunction with the appendix, will show how this can be done.

3.1 The Negative Binomial Distribution

The negative binomial distribution is a two-parameter distribution that is often used in projecting the wafer probe yield of integrated circuits. Its significance lies in its ability to describe the clustering of defects over the semiconductor wafer. The fact that defects cluster has been observed in the semiconductor industry for decades, and simply implies that defects are more likely to be found in groups than by themselves. The negative binomial distribution is characterized by the average number of defects per chip, denoted by λ, and the clustering parameter α. As the name implies, α describes the degree to which defects cluster over the wafer. The value of α typically ranges from 0.5 to 5 for different fabrication processes; the smaller values indicate increased clustering. As $\alpha \to \infty$ the negative binomial distribution becomes a Poisson distribution, which is characterized by no clustering [7].

For integrating yield-reliability modeling $\lambda = \lambda_K + \lambda_L$, where λ_K is the average number of killer defects per chip and λ_L is the average number of latent defects per chip. λ_K and λ_L are related through the parameter γ. That is, $\lambda_L = \gamma\lambda_K$, with $\gamma \approx 0.01 - 0.02$. Thus, for every 100 killer defects present, one expects, on average, $1 - 2$ latent defects.

Once wafer test has been performed, the parameters α and λ_K can be determined with standard statistical techniques [5, 8]. The value of λ_L (observed in burn-in), however, will depend on the burn-in duration. In particular, before burn-in, $\lambda_L = 0$, as these defects are not yet significant enough to reveal themselves. However, when burn-in is performed, these defects begin to "grow". Moreover, as the burn-in duration increases, more latent defects become severe enough to cause a failure. Thus, unlike killer defects, that will, with proper testing, reveal themselves immediately, latent defects are time-dependent. This suggests that λ_L should be replaced by $\lambda_L(t)$.

As shown in the appendix, when defects follow a negative binomial distribution, the reliability function $R(t)$ can be written as

$$R(t) = \left[1 + \frac{\lambda_L(t)}{\alpha}\right]^{-\alpha} \tag{2}$$

with the limiting values

$$\begin{aligned} R(t=0) &= 1 \\ R(t=\tau) &= \left[1 + \frac{\lambda_{L,max}}{\alpha}\right]^{-\alpha} \end{aligned} \tag{3}$$

The time $t = 0$ corresponds to the beginning of burn-in testing, where latent defects have yet to reveal themselves. $t = \tau$, on the other hand, corresponds to the end of the infant mortality period. Thus, at $t = \tau$, all latent defects have grown to failure, and $\lambda_L(t)$ reaches its maximum value, denoted above as $\lambda_{L,max}$. Of course, complete specification of $R(t)$ requires knowledge of the amplitude and time components of $\lambda_L(t)$. This will be discussed in the next section.

3.2 The Function $\lambda_L(t)$

Taking the logarithm of equation (2) gives

$$\ln \frac{1}{R(t)} = \alpha \ln \left[1 + \frac{\lambda_L(t)}{\alpha}\right] \tag{4}$$

Note that, the average number of latent defects per chip, $\lambda_L(t)$, will generally be a small number, even at its maximum value. Moreover, since the value of α typically ranges from 1 to 5, the ratio $\frac{\lambda_L(t)}{\alpha} << 1$.

Thus, with $\ln(1+x) \approx x$, for $x << 1$, equation (4) may be approximated as

$$\ln \frac{1}{R(t)} \approx \lambda_L(t) \tag{5}$$

Thus,

$$R(t) = \left[1 + \frac{\lambda_L(t)}{\alpha}\right]^{-\alpha} \approx \exp\left[-\lambda_L(t)\right] \tag{6}$$

which is only strictly true in the limit as $\alpha \to \infty$. However, since $\lambda_L(t) << 1$, this serves as a useful approximation.

In the semiconductor industry today it is often assumed that $R(t)$ follows a Weibull distribution. This is a two-parameter distribution with one parameter known as the shape parameter and the other known as the scale parameter. In this work the shape parameter will be denoted by β and the scale parameter will be denoted by ζ. The Weibull distribution then gives $R(t) = \exp\left[-(\zeta t^{\beta})\right]$. Note that this is equivalent to equation (6), as long as $\lambda_L(t) = \zeta t^{\beta}$. The boundary conditions at $t = 0$, where $\lambda_L(t=0) = 0$, and at some time $t = \tau$, where $\lambda_L(t=\tau) = \lambda_{L,max}$ reaches its maximum value, then imply that $\zeta = \frac{\lambda_{L,max}}{\tau^{\beta}}$. Thus, $\lambda_L(t) = \lambda_{L,max}\left(\frac{t}{\tau}\right)^{\beta}$ and equation (5) becomes

$$\ln \frac{1}{R(t)} \approx \lambda_L(t) = \lambda_{L,max}\left(\frac{t}{\tau}\right)^{\beta} \tag{7}$$

Taking the logarithm of both sides gives

$$\ln\ln \frac{1}{R(t)} = \beta \ln\left(\frac{t}{\tau}\right) + \ln \lambda_{L,max} \tag{8}$$

Thus, plotting the left hand side of equation (8) (obtained from stress data) versus $\ln t$ allows one to obtain the parameters β and τ.

One might argue that nothing new has been presented here since the plotting procedure described above has long been used in industry. Indeed, equation (8) amounts to assuming $R(t)$ follows a Weibull distribution and obtaining the associated shape and scale parameters from data. Note, however, that the formulation presented here has a very significant advantage in that it contains wafer probe information through the term $\lambda_{L,max}$. Indeed, as shown in the appendix, $\lambda_{L,max} = \alpha\gamma(1 - Y_K^{1/\alpha})$, where Y_K denotes the wafer test yield, α is the clustering parameter, and γ relates latent defects to killer defects. The fact that $\lambda_{L,max}$ depends on wafer test parameters provides the key to burn-in reduction strategies. This is addressed in the remaining sections of this paper.

4 Applications

4.1 Repaired Memory

The ability to repair integrated circuits can significantly increase chip yields. However, it has been demonstrated that the early-life reliability of repaired chips is degraded in the process [5]. This is a direct result of defect clustering; the more defects that are present on a chip (repaired or otherwise), the more likely it is to contain an additional early-life reliability defect.

To quantify the reliability impact of repairs, suppose the functional chips are separated into sub-groups based on the number of repairs performed following wafer test. The reliability function for functional chips with i repairs is then

$$R_i(t) = \left[1 + \frac{\lambda_L(t)}{\alpha}\right]^{-(\alpha+i)} \tag{9}$$

with $\lambda_L(t)$ given in the previous section.

In practice, one is often more interested in calculating the hazard function (i.e. failure rate) rather than the reliability function. This can be obtained directly from $R_i(t)$. In particular, it can be shown that the hazard function for chips with i repairs is simply related to the hazard function for chips with j repairs. That is,

$$h_j(t) = \left(\frac{\alpha+j}{\alpha+i}\right) h_i(t) \tag{10}$$

Equation (10) show that knowledge of any $h_i(t)$ allows one to obtain the remaining $h_j(t)$ for $i \neq j$. In particular, for $i = 0$

$$h_j(t) = \left[\frac{j}{\alpha} + 1\right] h_0(t) \tag{11}$$

Thus, the ratio of $h_j(t)$ and $h_0(t)$ is a line with slope $1/\alpha$ and intercept 1. Moreover, this ratio is independent of time and can be obtained once the clustering parameter is known; that is, following wafer test.

4.2 Numerical Results: Memory

The implications of equation (10) can now be considered. Suppose that a product containing repairable memory is to be subjected to burn-in. It is desired to determine the hazard function as a function of the number of repairs. Assume that the perfect yield is $Y_K = 0.20$ and the clustering parameter is $\alpha = 2$. Further, suppose $\gamma = 0.01$. Thus, for every 100 killer defects, one expects, on average, 1 latent defect. Moreover, latent defects are assumed to have the time dependence $\lambda_L(t) = \lambda_{L,max}\left(\frac{t}{\tau}\right)^{\beta} = \alpha\gamma(1 - Y_K^{1/\alpha})\left(\frac{t}{\tau}\right)^{\beta}$, with $\beta = 0.3$ and $\tau = 50$ burn-in hours.

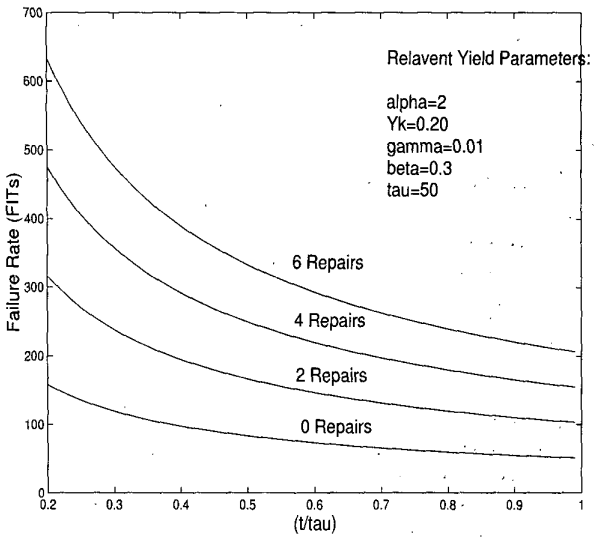

Figure 2: Hazard function versus normalized time for chips with various number of repairs. The relevant yield parameters are $\alpha = 2$, $\gamma = 0.01$ and $Y_k = 0.20$. The Weibull shape parameter is $\beta = 0.3$ and $\tau = 50$ burn-in hours.

Figure 2 shows the resulting hazard function in FITs versus $(\frac{t}{\tau})$ for chips with 0, 2, 4 and 6 repairs. 1 FIT is equal to 1 failure per million per 1000 hours of use. Further, the acceleration factor in burn-in is assumed to be 20,000. Thus, each burn-in hour is equal to 20,000 hours of use. Note that the failure rate at a given time can be significantly different depending on the number of repairs performed. For example, at $(\frac{t}{\tau}) = 0.6$, the failure rates are 73, 147, 220, and 293 FITs for chips with 0, 2, 4, and 6 repairs, respectively.

Moreover, as can be seen from equation (11), the clustering parameter α plays the key role in determining the relative failure rate for chips with a different number of repairs. As this equation shows, the greater the clustering (smaller α), the greater the relative hazard rate for chips with a different number of repairs. As $\alpha \rightarrow \infty$, the hazard curves are the same, regardless of the number of repairs.

4.3 Local Region Yield

The yield-reliability model can also be used predict the reliability of functional die from wafer regions with various local yields. A simple definition of local region yield corresponds to the wafer probe yield of a die's adjacent neighbors. One can then separate or bin functional die based on the number of faulty neighbors: die in bin 0 have 0 faulty neighbors, die in bin 1 have one faulty, and so on up to bin 8, where all neighbors are faulty. It has been demonstrated in [4] that the yield-reliability model based on defect clustering can then accurately predict the number of burn-in failures in each of the bins. The development in this subsection extends the model to predict the time at which these burn-in failures occur. As in the case of repaired memory, chips in different bins will be shown to have different failure rates.

Modeling the burn-in fall-out in time must incorporate the fact that the average number of latent defects is a time dependent quantity. As the mathematical development has shown, this means $\lambda_L \rightarrow \lambda_L(t)$, with $\lambda_L(t)$ chosen so as to satisfy the proper boundary conditions on the reliability function $R(t)$. Thus, all the equations derived in [4] pertaining to local region yield remain valid; one need only replace λ_L with the proper $\lambda_L(t)$.

4.4 Numerical Results: Local Region Yield

Figure 3 shows the hazard function in FITs for chips in different neighborhood bins. Again, 1 FIT is equivalent to 1 fail per million per 1000 hours of use. The wafer probe yield $Y_K = 0.50$, $\gamma = 0.01$, and the clustering parameter is $\alpha = 2$. The Weibull parameters are $\beta = 0.3$ and $\tau = 50$ burn-in hours. Moreover, it is assumed that the acceleration factor in burn-in is 20,000, so that each hour of burn-in corresponds to 20,000 hours of use. Note that, as with the case of chips with different repair counts, the hazard function can be quite different for die in different bins. In particular, die in bin 0 show very little fall-out in time, while bin 8 shows the largest fall-out for all times. For example, at $(\frac{t}{\tau}) = 0.4$ the hazard function in bin 8 is 140 FITs, while that in bin 0 is 15 FITs; a ratio of 9.3.

The hazard function for different values of α gives results similar to that for memory. In particular, the greater the clustering (smaller value of α), the greater the burn-in fall-out for chips in the higher numbered bins (i.e. surrounded by many faulty neighbors) compared to chips in the lower-numbered bins. When $\alpha \rightarrow \infty$, there will be no hazard difference between the bins.

5 Conclusions

This work has presented an early-life reliability model that allows one to relate wafer probe yield to burn-in

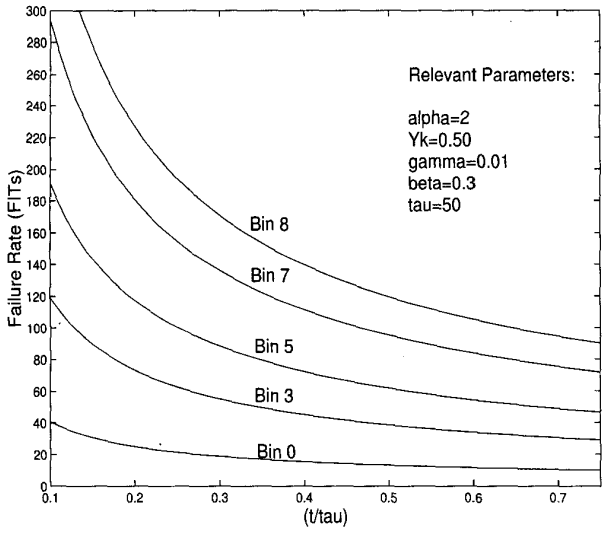

Figure 3: Hazard function versus $\left(\frac{t}{\tau}\right)$ for chips from various neighborhood bins. The relevant yield parameters are $\alpha = 2$, $\gamma = 0.01$ and $Y_k = 0.50$. The Weibull shape parameter is $\beta = 0.3$ and $\tau = 50$ burn-in hours.

fall-out in time. It was demonstrated that die that have been repaired following wafer test fall-out at a higher rate in burn-in when compared to chips with fewer repairs. Similarly, chips from regions of low local yield (i.e. many bad neighbors) were also shown to have higher failure rates than chips from local regions with higher yield (i.e. few bad neighbors). This information allows one to optimize stress tests such as burn-in by identifying failure rate curves for different populations and adjusting durations so as to meet outgoing reliability requirements; populations containing a larger number of latent defects will generally require longer burn-in durations to meet the same level of reliability.

Acknowledgments

This work was supported in part by the National Science Foundation through grant number NSF-CCR-9912389.

References

[1] Hance H. Huston and C. Patrick Clarke, "Reliability Defect Detection and Screening During Processing-Theory and Implementation," *Proceedings International Reliability Physics Symposium*, 1992, pp. 268-275.

[2] W. Riordan, R. Miller, J. Sherman, J. Hicks, "Microprocessor Reliability Performance as a Function of Die Location for a 0.25μ, Five Layer Metal CMOS Logic Process", *Proceedings International Reliability Physics Symposium*, 1999, pp. 1-11.

[3] Russell Miller, Walter Riordan, "Unit Level Predicted Yield: A Method of Identifying High Defect Density Die at Wafer Sort", *Proceedings 2001 International Test Conference*, October 2001, pp. 1118-1127.

[4] T.S. Barnett, A.D. Singh, M. Grady, K.G. Purdy, "Yield-Reliability Modeling: Experimental Verification and Application to Burn-In Reduction", *Proceedings 2002 VLSI Test Symposium*, May 2002, pp. 75-80.

[5] T.S. Barnett, A.D. Singh, M. Grady, K.G. Purdy, "Redundancy Implications for Product Reliability: Experimental Verification of an Integrated Yield-Reliability Model", *Proceedings 2002 International Test Conference*, October 2002.

[6] Way Kuo, Wei-Ting Kary Chien, Taeho Kim, "Reliability, Yield, and Stress Burn-in" *Kluwer Academic Publishers*, 1998.

[7] C.H.Stapper, F.M. Armstrong, and K. Saji, "Integrated Circuit Yield Statistics", *Proceedings of IEEE*, April 1983, pp. 453-470.

[8] I. Koren and C.H. Stapper, "Yield Models for Defect Tolerant VLSI Circuits: A Review," Defect and Fault Tolerance in VLSI Systems, Vol. 1, I. Koren (ed.), pp. 1-21, Plenum, 1989.

[9] I. Koren, Z. Koren and C.H. Stapper, "A Unified Negative Binomial Distribution for Yield Analysis of Defect Tolerant Circuits," IEEE Trans. on Computers, Vol. 42, June 1993, pp. 724-437.

[10] T.S. Barnett, A.D. Singh and V.P. Nelson, "Estimating Burn-In Fall-Out for Redundant Memory", *Proceedings 2001 International Test Conference*, October 2001, pp. 340-347.

[11] Way Kuo and Taeho Kim, "An Overview of Manufacturing Yield and Reliability Modeling for Semiconductor Products", *Proceedings of the IEEE*, Vol. 87, No. 8, August, 1999, pp. 1329-1344.

[12] R. Madge, M. Rehani, K. Cota, W.R. Daasch, "Statistical Post-Processing at Wafersort: An Alternative to Burn-In and a Manufacturable Solution to Test Limit Setting for Sub-micron Technologies", *Proceedings 2002 VLSI Test Symposium*, May 2002, pp. 69-74.

Mathematical Appendix

This appendix provides the mathematical details required to obtain the main results of the paper.

The Reliability Function

Let $R[t, K(0)]$ denote the probability that a chip survives burn-in on the interval $[0, t)$ and contains 0 killer defects. These chips are to be subjected to burn-in. $R[t, K(0)]$ can be written as

$$R[t, K(0)] = \sum_{\ell=0}^{\infty} R[\,t|L(\ell), K(0)]\, P[L(\ell), K(0)] \quad \text{(A.1)}$$

where $R[\,t|L(\ell), K(0)]$ is the probability a chip survives stress testing on $[0, t)$, given that it contains exactly ℓ latent (i.e. early-life reliability) defects, and 0 killer defects. Using Bayes' Rule $P[L(\ell), K(0)] = P[L(\ell)|K(0)]P[K(0)] = P[L(\ell)|K(0)]Y_K$, this can be written

$$R[t, K(0)] = Y_K \sum_{\ell=0}^{\infty} R[\,t|L(\ell), K(0)]\, P[L(\ell)|K(0)] \quad \text{(A.2)}$$

where $P[L(\ell)|K(0)]$ is the probability of exactly ℓ latent defects, given 0 killer defects, and $Y_K = P[K(0)]$ is the wafer probe yield (i.e. the probability of 0 killer defects).

To calculate $R[t|L(\ell), K(0)]$ note that, if the chip is to survive until time t, then all of the ℓ latent defects must cause a failure after time t. Thus, if τ_j denotes the time to failure due to the j^{th} defect, where $j = 1, 2, \ldots, \ell$, then one can write

$$\begin{aligned} R[\,t|L(\ell), K(0)] &= P[\tau_1 \geq t, \tau_2 \geq t, \ldots, \tau_\ell \geq t] \\ &= P[\tau_1 \geq t]\, P[\tau_2 \geq t] \ldots P[\tau_\ell \geq t] \end{aligned} \quad \text{(A.3)}$$

The last equality follows from the fact that the time to failure of the i^{th} defect is independent of the time to failure of the j^{th} defect, for all i and j. Moreover, if each one of the ℓ defects is described by the same reliability function, then $P[\tau_1 \geq t] = P[\tau_2 \geq t] = \ldots = P[\tau_\ell \geq t]$. But $P[\tau_i \geq t]$ is just the reliability function for a chip with a single latent defect. Hence, $P[\tau_1 \geq t] = P[\tau_2 \geq t] = \ldots = P[\tau_\ell \geq t] = R[t|L(1), K(0)]$, and equation (A.3) can now be written

$$R[\,t|L(\ell), K(0)] = R^{\ell}[\,t|L(1), K(0)] \quad \text{(A.4)}$$

Substituting this into equation (A.2) gives

$$\begin{aligned} R[t, K(0)] &= Y_K \textstyle\sum_{\ell=0}^{\infty} R^{\ell}[\,t|L(1), K(0)]\, P[L(\ell)|K(0)] \\ &= Y_K\, [T(z)]_{z=R[\,t|L(1),K(0)]} \end{aligned} \quad \text{(A.5)}$$

where

$$T(z) = \sum_{\ell=0}^{\infty} z^{\ell}\, P[L(\ell)|K(0)] \quad \text{(A.6)}$$

is the probability generating function for $P[L(\ell)|K(0)]$.

It was shown in [5] that $P[L(\ell)|K(0)]$ follows a negative binomial distribution with an average number of latent defects per chip $\lambda_{L,max} = \gamma\lambda_K/(1 + \frac{\lambda_K}{\alpha}) = \alpha\gamma(1 - Y_K^{1/\alpha})$ and clustering parameter α. This has the generating function [9]

$$T(z) = \left[1 + (1 - z)\frac{\lambda_{L,max}}{\alpha}\right]^{-\alpha} \quad \text{(A.7)}$$

The reliability function for chips with 0 killer defects is therefore

$$R[t, K(0)] = Y_K \left[1 + F[t|L(1), K(0)]\frac{\lambda_{L,max}}{\alpha}\right]^{-\alpha} \quad \text{(A.8)}$$

Here $F[t|L(1), K(0)] = 1 - R[t|L(1), K(0)]$ is the time to failure distribution for a chip with a single latent defect. Thus, $F[t|L(1), K(0)]$ represents the probability that a chip with a single latent defect fails on $[0, t)$.

Note that the term $F[t|L(1), K(0)]\lambda_{L,max}$ indicates that the average number of latent defects is a time dependent quantity. This simply means that, as an integrated circuit is operated or stressed in burn-in, more latent defects will manifest themselves. One can therefore define $\lambda_L(t) = F[t|L(1), K(0)]\lambda_{L,max}$. With this notation equation (A.8) may be rewritten as

$$R[t, K(0)] = Y_K \left[1 + \frac{\lambda_L(t)}{\alpha}\right]^{-\alpha} \quad \text{(A.9)}$$

If one is only interested in those chips that will subjected to burn-in, one can simply use Bayes' rule to eliminate Y_K. That is, since $R[t|K(0)] = R[t, K(0)]/P[K(0)] = R[t, K(0)]/Y_K$, dividing equation (A.9) by Y_K refers only to those chips subjected to burn-in. For simplicity of notation, it is convenient to define $R(t) \equiv R[t|K(0)]$. With this notation

$$R(t) = \left[1 + \frac{\lambda_L(t)}{\alpha}\right]^{-\alpha} \quad \text{(A.10)}$$

Memory Calculations

The development in the previous section will now be applied to integrated circuits containing repairable memory circuits. This means that $P[L(\ell)|K(0)]$ must be replaced by $P[L(\ell)|G(i)]$, where $G(i)$ denotes the event that a chip is functional following wafer probe and has been repaired exactly i times. As shown in [5, 10], $P[L(\ell)|G(i)]$ follows a negative binomial distribution with parameters $[\lambda_L(i), \alpha + i]$, where, $\lambda_L(i) = (\frac{\alpha+i}{\alpha})\lambda_{L,max}$ is the average number of latent defects given that there are i repairs, and $\lambda_{L,max} = \alpha\gamma(1 - Y_K^{1/\alpha})$. Y_K denotes the perfect wafer probe yield, that is, the fraction of functional chips with 0 repairs. Specifically, one can write

$$P[L(\ell)|G(i)] = \frac{\Gamma(\alpha+i+\ell)}{\ell!\,\Gamma(\alpha+i)} \frac{\left[\frac{\lambda_L(i)}{\alpha+i}\right]^{\ell}}{\left[1+\frac{\lambda_L(i)}{\alpha+i}\right]^{\alpha+i+\ell}} \tag{A.11}$$

Now assume that chips are separated into sub-groups based on the number of repairs performed following wafer test. Thus, the i^{th} sub-group denotes all those chips with exactly i repairs. Then according to equation (A.8), the reliability for chips with i repairs, $R[t|G(i)]$, will be different for chips with a different number of repairs. In particular, $R[t, G(i)]$ is given by

$$R[t, G(i)] = P[G(i)]\left[1 + F[t|L(1), K(0)]\frac{\lambda_L(i)}{\alpha+i}\right]^{-(\alpha+i)}$$

$$= P[G(i)]\left[1 + F[t|L(1), K(0)]\frac{\lambda_{L,max}}{\alpha}\right]^{-(\alpha+i)} \tag{A.12}$$

The last equality follows from the fact that $\frac{\lambda_L(i)}{\alpha+i} = \frac{\lambda_{L,max}}{\alpha}$ for all values of i. Defining $R_i(t) \equiv R[t|G(i)] = \frac{R[t,G(i)]}{P[G(i)]}$ as the reliability function for functional chips with i repairs, one can write

$$R_i(t) = \left[1 + \frac{\lambda_L(t)}{\alpha}\right]^{-(\alpha+i)} \tag{A.13}$$

The reliability of the chip population taken as a whole is obtained be calculating the sum

$$R(t) = \sum_{i=0}^{\infty} c_i R_i(t) \tag{A.14}$$

where c_i is the fraction of good die with exactly i repairs. The fraction c_i may be written as

$$c_i = \frac{p_R^i P[K(i)]}{Y_{Keff}} \tag{A.15}$$

where p_R is the fraction of chip area that is repairable, Y_{Keff} is the wafer yield with repair, and $P[K(i)]$ is the probability of exactly i killer defects. The sum of equation (A.14) can be easily evaluated, resulting in

$$R(t) = \left[1 + \left(\frac{Y_{Keff}}{Y_K}\right)^{\frac{1}{\alpha}} \frac{\lambda_L(t)}{\alpha}\right]^{-\alpha} \tag{A.16}$$

Thus, repairability has the effect of increasing $\lambda_L(t)$ by the factor $\left(\frac{Y_{Keff}}{Y_K}\right)^{\frac{1}{\alpha}}$. Note also that, when there is no repair capability, $Y_{Keff} = Y_K$ and equation (A.16) reduces to equation (A.10).

The hazard function, defined as $h(t) = -\frac{\partial \ln R(t)}{\partial t}$, can now be written individually for chips with i repairs.

$$h_j(t) = -\frac{\partial \ln R_j(t)}{\partial t} = \left(\frac{\alpha+j}{\alpha+i}\right) h_i(t) \tag{A.17}$$

For the population taken as a whole, $h(t) = -\frac{\partial \ln R(t)}{\partial t}$, with $R(t)$ is given in equation (A.16). With $\lambda_L(t) = \lambda_{L,max}\left(\frac{t}{\tau}\right)^{\beta}$, and after some simplification, this gives

$$h(t) = \frac{\beta\alpha}{t}[1 - R^{\frac{1}{\alpha}}(t)] \tag{A.18}$$

Calculations for Local Region Yield

Define $R_i(t)$ as the probability that a chip with i faulty neighbors survives on the interval $[0, t)$. This chip has been determined functional following wafer probe testing and is to be subjected to burn-in. $R_i(t)$ is then obtained by incorporating the proper time-dependence into the neighborhood equations derived in [4]. Define $P[X_K(k), X_L(\ell); t]$ as the probability that exactly k chips are free of killer defects and ℓ chips are free of latent defects. Then

$$R_i(t) = \frac{1}{N}\sum_{\ell=0}^{N} \ell\, \frac{P[X_K(N-i), X_L(\ell); t]}{P[X_K(N-i)]} \tag{A.19}$$

where $i = 0, 1, \ldots, N-1$, and N is the number of die in the neighborhood. The sum over ℓ may be evaluated. The result is

$$R_i(t) = \sum_{q=0}^{i} a(N, i, q; t) / \sum_{q=0}^{i} b(N, i, q) \tag{A.20}$$

with

$$a(N, i, q; t) = (-1)^q \binom{i}{q} \left[1 + \frac{(N-i+q)\lambda_K}{\alpha} + \frac{\lambda_L'(t)}{\alpha}\right]^{-\alpha}$$

$$b(N, i, q) = (-1)^q \binom{i}{q} \left[1 + \frac{(N-i+q)\lambda_K}{\alpha}\right]^{-\alpha} \tag{A.21}$$

Here $\lambda_L'(t) = \gamma\lambda_K\left(\frac{t}{\tau}\right)^{\beta} = \gamma\alpha(Y_K^{-1/\alpha} - 1)\left(\frac{t}{\tau}\right)^{\beta}$. The failure rate or hazard function for die in each bin can be obtained through differentiation via the equation $h_i(t) = -\frac{1}{R_i(t)}\frac{\partial R_i(t)}{\partial t}$.