A dual burn-in policy for defect-tolerant memory products using the number of repairs as a quality indicator

SEUNG-HOON TONG  
Memory Division, Semiconductor Business  
Samsung Electronics Co., Ltd.  
San16 Banwol-dong, Hwasung-city  
Gyeonggi-do 445-701, Korea  
tel. +82-31-208-4626  
fax. +82-31-208-6698  
e-mail. shtong@samsung.com

BONG-JIN YUM*  
Department of Industrial Engineering  
Korea Advanced Institute of Science and Technology  
373-1 Guseong-dong, Yuseong-gu  
Daejeon 305-701, Korea  
tel. +82-42-869-3116  
fax. +82-42-869-3110  
e-mail. bjyum@kaist.ac.kr

* Corresponding Author
A dual burn-in policy for defect-tolerant memory products using the number of repairs as a quality indicator

SEUNG-HOON TONG a and BONG-JIN YUM b,*

a Memory Division, Semiconductor Business, Samsung Electronics Co., Ltd., San16, Banwol-dong, Hwasung-city, Gyeonggi-do 445-701, Korea

b Department of Industrial Engineering, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea

In most existing studies on the optimization of burn-in for semiconductor products, all chips are treated equally and subjected to burn-in of the same duration (i.e. a single burn-in (SBI) policy is employed). However, the quality levels of chips before burn-in are not the same in general, and therefore, it may be more advantageous to treat chips differently at the burn-in process based on appropriate quality indicators. This paper considers defect-tolerant memory products and develops a dual burn-in (DBI) policy in which the chips submitted to burn-in are classified into two groups according to the number of repairs, a quality indicator that can be obtained from the wafer probe test results, and different burn-in durations are applied to different groups of chips. Then, cost models are developed for the SBI and DBI policies, and their relative performances are compared in terms of the expected total cost per chip. The effectiveness of the proposed DBI policy is demonstrated using the actual data for a certain type of 256M DRAM products.

Keywords: Single burn-in policy; Dual burn-in policy; Defect-tolerant memory, Cost model

* Corresponding author. e-mail: bjyum@kaist.ac.kr
1. Introduction

Semiconductor manufacturing consists of two major steps: fabrication of integrated circuits onto wafers to form dies, and assembly of each die into a finished product (i.e., a chip). Immediately after fabrication, wafer probe (WP) test is conducted on each die to identify good and defective ones. In addition, defect-tolerant memory products are laser-repaired if possible by replacing defective memory cells with the provided redundancies. Then, dies which pass the WP test, i.e. good and repaired ones, are sent to the assembly process, at the end of which burn-in is conducted to screen chips for infant mortality failures.

While burn-in is beneficial to improving the reliability of products shipped to customers, it is costly since burn-in test equipment is highly expensive and burn-in boards are expendable materials that should be replaced regularly. Moreover, since burn-in is usually applied to all packaged chips, it affects the productivity and could prevent timely shipment of product. Therefore, it is imperative for a semiconductor manufacturer to devise an efficient burn-in policy in terms of cost and product reliability. Burn-in efficiency can be achieved by establishing a rational policy concerning such factors as burn-in duration, burn-in conditions, etc. such that the required reliability level in the field is secured and the related cost is minimized. In this paper, we focus on the burn-in duration as a major factor for achieving burn-in efficiency.

If the burn-in duration is extended, then the direct burn-in and burn-in failure costs increase while the field failure cost during the warranty period decreases. Considering these trade-offs, many authors [1-6] developed burn-in policies in which the burn-in duration is optimized with respect to the total cost involved. However, most of the
previous works treated the products submitted to burn-in equally (i.e. the same burn-in duration is applied to all products). On the other hand, it is well known for semiconductor products that the quality and eventually the reliability levels of chips submitted to burn-in are not the same, and, for defect-tolerant memory products, may depend on the number of repairs made at the WP test [7,8].

Defects from semiconductor manufacturing can be classified into killer and latent defects [7]. A killer defect affects the electrical functioning of a die and can be detected and subsequently repaired at the WP test. The latent defect is basically the same as the killer defect in nature, but may pass the WP test without being detected since it is small in size and/or occurs in undetectable locations. It is well known that manufacturing defects on a wafer are clustered rather than randomly distributed [9-12]. This implies that a chip which passes the WP test after a large number of repairs due to killer defects is likely to have more latent defects than those with a small number of repairs, and therefore, it becomes more susceptible to failure during burn-in and/or in the field. In other words, the number of repairs \( (r) \) at the WP test can be used as an indicator for the quality and eventually the reliability level of a chip submitted to burn-in [7,8,13].

Developed in this paper is a dual burn-in (DBI) policy in which the chips submitted to burn-in are classified into two groups according to \( r \), and shorter burn-in duration is employed for the group of chips with \( r \) smaller than a boundary value \( r_c \), and vice versa. The justification of this policy is as follows. A chip with a larger value of \( r \) than others is more likely to fail early in the field, and therefore, the burn-in duration for such a chip is extended to increase the chance of screening, and thereby, to meet the reliability requirement in the field. On the other hand, when \( r \) is small, the burn-in duration is reduced to enhance productivity and to make economic benefits.
Fig. 1 shows the overall procedures for developing a DBI policy for defect-tolerant memory products. In the data collection step, $r$ and the burn-in result (i.e. failed or unfailed) for each chip are collected at the WP and burn-in test, respectively. Before formulating a cost model, two probability models need to be developed. One is the distribution of $r$, and the other is the reliability of a chip as a function of $r$ and burn-in duration. In the present investigation, a negative-binomial distribution is adopted for the former, and an extended proportional hazard (PH) model [14] in which $r$ is considered as an explanatory variable is used for the latter. For the proposed DBI policy, burn-in test results from two different burn-in durations are sufficient to estimate unknown parameters in the extended PH model. The proposed DBI policy is then optimized by determining $r_c$ and the burn-in duration for each group such that the expected total cost per chip is minimized under the constraint on the average failure rate (AFR) in the field.

For a certain type of 256M DRAM products, the effectiveness of the proposed DBI policy is demonstrated by comparing its expected total cost per chip with that of the single burn-in (SBI) policy in which the same burn-in duration is applied to all chips regardless of $r$. In addition, a sensitivity analysis is conducted to assess the effect of the uncertainties in the assumed burn-in acceleration factor (AF) and cost parameters on the effectiveness of the proposed DBI policy.

The rest of this paper is organized as follows. Section 2 presents a review of the literature relevant to the present investigation. Estimation procedures for the unknown parameters in the repair distribution and the PH model are introduced in sections 3 and 4, respectively. Section 5 develops cost models for the DBI and SBI policies based on the results from sections 3 and 4. Then, for a certain type of 256M DRAM products,
each policy is optimized and the relative performances of the two policies are compared in terms of the expected total cost per chip. Finally, conclusions, cautionary remarks, and future research directions are presented in section 6.

2. Related works

Early works [15,16] on the relationship between the WP and burn-in test results are mainly concerned with correlating WP yield with burn-in yield (or reliability yield as termed in Kuo and Kim [12]). In these early yield-reliability approaches, however, only the yield data are considered among the information that can be obtained from the WP test, and in addition, time variable is not taken into consideration although the reliability yield depends on burn-in duration. Later studies [5,17-19] extend the previous works to include time variable in their yield-reliability models. Nevertheless, these later works also do not consider such information as $r$ at the WP test.

Barnett and Singh [13] propose a theoretical yield-reliability model in which time variable and $r$ are explicitly considered. In developing their model, the number of defects of a die is assumed to follow a negative binomial distribution based on the well-known fact that defects are more likely to be clustered than randomly distributed on a wafer. An important advantage of the Barnett and Singh approach is that the model parameters (e.g. average number of killer defects, degree of defect clustering, ratio of the average number of latent defects to that of killer defects, etc.), if properly estimated, can be effectively used for the purpose of monitoring and controlling a semiconductor manufacturing process. However, two assumptions need to be met to utilize their model. That is, 1) either the time point at which all latent defects are precipitated to failure or the ratio of the average number of latent defects to that of killer defects per chip is assumed to be known, and 2) the average number of latent defects are assumed to be
much smaller than the clustering parameter. In practice, however, it may not be easy to specify such a time point or ratio as in the first assumption. In addition, the second assumption may not be valid for immature products which are more prone to defects than mature products. Therefore, in the present investigation, an empirical approach is developed using a PH model for describing the product reliability during and after burn-in as a function of $r$ and burn-in duration.

3. Distribution of the number of repairs

3.1 Negative binomial distribution

Let $r$ be a random variable which represents the number of repairs of a die which passes the WP test, and $N$ be the maximum number of possible repairs per die. A realization of $r$ will be denoted by $i$. Under the assumption that $N = \infty$, Barnett et al. [8] derived the probability mass function of $r$, i.e. pmf($r$). After some algebraic manipulation of their result, it can be shown that $u(i)$, the pmf($r$), can be expressed as a negative binomial distribution with parameters $\alpha$ ($> 0$, clustering parameter) and $w$ $(0 < w < 1)$ as follows.

$$u(i) = \Pr(r = i \mid \text{WP passed}) = \frac{\Gamma(\alpha + i)}{i!\Gamma(\alpha)} w^i (1 - w)^\alpha, \quad i = 0, 1, \ldots$$

(1)

The pmf in (1) can be used as a good approximation to the repair distribution when $N$ is finite but large, which is the case for today’s memory chips which have hundreds of redundancies (e.g. word and bit lines) to improve WP yield.

3.2 Fitting negative-binomial distribution to repair data

To fit a negative binomial distribution to actual repair data, an estimate of $u(i)$ is first calculated as:
\[ u'(i) = \frac{n_{wi}}{wN}, \ i = 0, 1, \cdots, N \]

where \( n_{wi} \) is the number of chips, each of which passes the WP test and is repaired \( i \) times, and \( wN \) is the total number of chips which pass the WP test. Then, the following nonlinear regression model can be formulated.

\[
u'(i) = u(i) + \varepsilon_i = \frac{\Gamma(\alpha + i)}{\Gamma(\alpha)} w^i (1 - w)^\alpha + \varepsilon_i, \quad i = 0, 1, \cdots, N \tag{2} \]

where \( \varepsilon_i \) is an error term.

Since the variance of \( u'(i) \) in (2) is not constant across \( i \), parameters \( \alpha \) and \( w \) are estimated using the weighted nonlinear least squares method. The weight is defined as the reciprocal of the variance of \( u'(i) \), and can be approximated as follows.

\[ w_i \approx \frac{N_w}{\left\{ u'(i) [1 - u'(i)] \right\}}, \quad i = 0, 1, \cdots, N. \]

3.3 Application to 256M DRAM repair data

The above nonlinear regression is performed on the 256M DRAM repair data using the ‘weight’ option in the NLIN procedure of SAS [20]. For the data, \( N_w \approx 19,800. \) The value of \( N \) is not provided in this paper for proprietary reasons.

The corresponding ANOVA table and the estimates of \( \alpha \) and \( w \) (\( \hat{\alpha} = 10.7759 \) and \( \hat{w} = 0.7282 \)) are shown in Tables 1 and 2, respectively. Fig. 2 shows \( u'(i) \) and fitted value \( \hat{u}(i) \) for the 256M DRAM repair data. Both the ANOVA results in Table 1 and Fig. 2 indicate that the model in (2) fits the repair data reasonably well.

4. PH model for reliability

In order to formulate a cost model for developing a burn-in policy, the reliability of
the product must be described as a function of burn-in duration and \( r \). The PH model proposed by Cox [21] has been frequently used to describe the relationship between the lifetime of a product and its covariates or explanatory variables [22,23]. The present study employs an extended PH model in which \( r \) is used as an explanatory variable to describe the reliability of a chip that passes the WP test and is subjected to burn-in for certain duration.

4.1 PH model

The Cox PH model [21] is given by

\[
   h(t; \mathbf{x}) = h_0(t) \exp \left( \sum_{k=1}^{P} \gamma_k x_k \right)
\]

where \( t \) is a time variable, \( x_k \) is an explanatory variable, and \( h_0(t) \) is the baseline hazard function. In the present study, the following extended Cox PH model [14] is employed to allow for a more flexible fit to the actual data.

\[
   h(t; \mathbf{x}) = h_0(t) \exp \left[ \sum_{k=1}^{P} \gamma_k x_k + \sum_{k=1}^{P} \delta_k x_k g_k(t) \right]
\]

where \( g_k(t) \) is a known function of \( t \). In terms of the reliability function, (3) can be re-expressed as [24]:

\[
   R(t; \mathbf{x}) = \left[ R_0(t) \right] \exp \left[ \sum_{k=1}^{P} \gamma_k x_k + \sum_{k=1}^{P} \delta_k x_k g_k(t) \right].
\]

One of the most frequently used baseline hazard functions in parametric PH models is the Weibull hazard function. That is,

\[
   h_0(t) = \left[ \left( \beta/\eta \right) \left( t/\eta \right)^{\beta-1} \right],
\]

or equivalently,
\[ R_0(t) = \exp\left[-\left(\frac{t}{\eta}\right)^\beta\right] \]

where \( \beta \) and \( \eta \) are the shape and scale parameters of a Weibull distribution, respectively. Since \( r \) is the only explanatory variable in the present study, (4) reduces to

\[ R(t; i) = \left\{ \exp\left[-\left(\frac{t}{\eta}\right)^\beta\right] \right\} \mathbb{E}^{\theta_r i + \theta_{rt} i g(t)} \]

(5)

where \( t \) represents the burn-in duration and equals zero at the start of burn-in, and \( R(t; i) \) is the probability that a chip, which passes the WP test and is repaired \( i \) times, survives \( t \) hours of burn-in. Taking natural logarithm of both sides of (5) twice yields

\[ \ln\left[\ln R(t; i)\right] = -\beta \ln \eta + \beta \left(\ln t\right) + \theta_r i + \theta_{rt} i g(t) \]

\[ \Rightarrow y = \theta_0 + \theta_r \left(\ln t\right) + \theta_{rt} i + \theta_{rt} i g(t) \]

where \( y = \ln\left[\ln R(t; i)\right], \theta_0 = -\beta \ln \eta, \) and \( \theta_r = \beta \).

When actual data are given, \( R(t; i) \) can be estimated as follows.

\[ R_{ij} = R'(t = t_j; i) = n_{wij} / N_{wij}, \ i = 0, 1, \ldots, N, \ j = 1, 2, \ldots, m \]

where \( n_{wij} \) is the number of chips, each of which passes the WP test, is repaired \( i \) times, and survives \( t_j \) hours of burn-in, \( N_{wij} \) is the total number of chips, each of which passes the WP test, is repaired \( i \) times, and is subjected to \( t_j \) hours of burn-in, and \( m \) represents the number of different burn-in durations.

### 4.2 Weighted least squares estimation

Let \( y_{ij}' = \ln\left(\ln R_{ij}'\right) \). Then, the following linear regression model can be formulated.

\[ y_{ij}' = \theta_0 + \theta_r \left(\ln t_j\right) + \theta_{rt} i + \theta_{rt} i g(t_j) + \epsilon_{ij}, \ i = 0, 1, \ldots, N, \ j = 1, 2, \ldots, m \]  

(6)
where $\varepsilon_{ij}$ is an error term and $g(\cdot)$ is a known function. The unknown parameters $\theta_0$, $\theta_t$, $\theta_r$, and $\theta_{ri}$ in (6) can be estimated using the usual linear regression technique. Since the variance of $y'_{ij}$ (or equivalently the variance of $\varepsilon_{ij}$) is not constant across $i$ and $j$, the weighted least squares method should be used with weight $w_{ij}$ for $y'_{ij}$ being given by the reciprocal of the variance of $y'_{ij}$ as follows (see appendix A for derivation).

$$w_{ij} \approx \left[ N w_{ij} R_{ij}' \left( \ln R_{ij}' \right)^2 \right] \left( 1 - R_{ij}' \right).$$  (7)

### 4.3 Application to 256M DRAM burn-in data

In Fig. 3, $y'$ is plotted against $i$ for the 256M DRAM product subjected to two burn-in durations. The cases where $n_{wij}$ is below 20 are excluded since the corresponding $y'$'s show fluctuating behavior due to small sample size. Most of the excluded cases are located over the range of large $i$. Fig. 3 shows that, for given $t$, $y'$ is approximately linearly related to $i$ with a slope being dependent on $t$. This can be accommodated by taking $g(t_j)$ as $\ln t_j$ in (6). That is,

$$y'_{ij} = \theta_0 + \theta_t \left( \ln t_j \right) + \theta_r i + \theta_{ri} i \left( \ln t_j \right) + \varepsilon_{ij}, \quad i = 0, 1, \cdots, N, \quad j = 1, 2$$  (8)

Note that $m = 2$ in (8). That is, unknown parameters in (8) can be estimated if burn-in data exist for at least two different burn-in durations, $t = t_1$ and $t_2$.

Table 3 shows the ANOVA results for the above regression analysis, and Table 4 shows the estimates of the unknown parameters in (8) for the 256M DRAM data in Fig. 3. The P-values in Table 4 indicate that $\ln t$ and possibly $i$ term are not statistically significant. On the other hand, the hierarchy principle [25,26] states that if a model
contains a high-order term, then it should also contain lower-order terms that are part of the high-order term. In the present study, the hierarchy principle is followed and \( \ln t \) and \( i \) terms are also included in the final model since \( i(\ln t) \) term is statistically significant.

Based on the above analysis, the regression equation for the 256M DRAM product is estimated as

\[
\hat{y}' = -2.5468 + 0.0181(\ln t) - 0.008438 i + 0.020981 i(\ln t) .
\] (9)

In terms of the reliability function (see (5)),

\[
\hat{R}(t;i) = \left\{ \exp \left[ - \left( \hat{\theta}_1 e^{\hat{\theta}_0} \right) \right] \right\} \exp \left[ \hat{\theta}_1 i + \hat{\theta}_2 i (\ln t) \right] \\
= \left[ \exp \left( -0.078332 t^{0.0181} \right) \right] \exp \left[ -0.008438 i + 0.020981 i (\ln t) \right]. \quad (10)
\]

5. Burn-in cost models

5.1 SBI policy

In the SBI cost model, three types of costs are considered per chip. They include direct burn-in cost \( c_1 \) (per hour of burn-in), burn-in failure cost \( c_2 \), and field failure cost \( c_3 \). The burn-in failure cost is incurred if a chip fails during burn-in, while the field failure cost is incurred if a chip shipped to a customer fails within the warranty period. The manufacturing cost per chip is regarded the same for all chips which pass the WP test, and therefore, not included in the cost model. In general, the three types of costs are related as \( c_1 < c_2 < c_3 \).

When the burn-in duration is \( t_b \) and the warranty period is \( t_f \) as shown in Fig. 4, the expected total cost per chip can be expressed as follows.
\[ C_{SBI}(t_b) = (\text{direct burn-in cost}) + (\text{burn-in failure cost}) + (\text{field failure cost}) \]
\[ = c_1 t_b + c_2 \Pr(T < t_b | \text{WP passed}) \]
\[ + c_3 \Pr(T < t_b + t_f, T > t_b | \text{WP passed}). \]

where \( T \) represents the lifetime of a chip.

\[ \text{Derivation of } \Pr(T < t_b | \text{WP passed}) : \] This is the probability that a chip which has passed the WP test fails during burn-in of duration \( t_b \), and is given by
\[ \sum_{i=0}^{\infty} u(i) \left[ 1 - R(t_b; i) \right] \text{ where } R(\cdot; i) \text{ is defined in (5) (see appendix B for derivation).} \]

\[ \text{Derivation of } \Pr(T < t_b + t_f, T > t_b | \text{WP passed}) : \] This is the probability that a chip which has passed the WP test survives the burn-in of duration \( t_b \) but fails in the field within the warranty period \( t_f \), and is given by
\[ \sum_{i=0}^{\infty} u(i) \left[ R(t_b; i) - R(t_b + t_f; i) \right]. \] This can be proved in a similar manner as in Appendix B.

Based on the above derivations, (11) can be rewritten as (12).
\[ C_{SBI}(t_b) = c_1 t_b + c_2 \sum_{i=0}^{\infty} u(i) \left[ 1 - R(t_b; i) \right] \]
\[ + c_3 \sum_{i=0}^{\infty} u(i) \left[ R(t_b; i) - R(t_b + t_f; i) \right]. \]

In addition, there usually exists a requirement on the reliability level of the product in the field. Such a requirement is often expressed as the AFR during \( t_f \) [27], which is defined as
\[ AFR(t_f) = \left( \frac{1}{t_f} \right) \int_0^{t_f} h(x) dx = \left[ -\ln R(t_f) \right] / t_f \]
where $h(\cdot)$ and $R(\cdot)$ are the hazard and reliability functions in the field, respectively.

*Derivation of $R(t_f)$*: Since only the chips that survive the burn-in test are shipped to the customer, $R(t_f)$ is defined as $\Pr(T > t_b + t_f | T > t_b, \text{WP passed})$. In a similar manner as in appendix B, it can be shown that

$$R(t_f) = \sum_{i=0}^{\infty} u(i) R(t_b + t_f; i) \left/ \sum_{i=0}^{\infty} u(i) R(t_b; i) \right..$$

(14)

The product reliability at the field time $t_f$ can be obtained using (14). In this paper, the field time is translated into the time in the burn-in time scale using AF. In addition, it is assumed that warranty period $t_f$ is within the infant mortality period.

Based on the above, the optimization problem for determining the burn-in duration $t_b$ for the SBI policy can be formulated as follows.

$$\begin{align*}
\text{Minimize} & \quad C_{SBI}(t_b) \\
\text{subject to} & \quad AFR(t_f) \leq AFR_{req}
\end{align*}$$

(15)

where $AFR_{req}$ is the upper limit of $AFR(t_f)$. In the above optimization, $u(i)$ and $R(t_b; i)$ are replaced with their estimates $\hat{u}(i)$ and $\hat{R}(t_b; i)$, respectively. In addition, the upper limit of each summation with respect to $i$ is replaced with $N$ as an approximation.

### 5.2 DBI policy

In the proposed DBI policy, the chips which pass the WP test are first classified into two groups using the boundary value $r_c$ of $r$. Then, the chips whose $r$ is less than or equal to $r_c$ are subjected to burn-in of duration $t_{b1}$, while the chips whose $r$ is greater than $r_c$ are subjected to burn-in of duration $t_{b2}$ where $t_{b1} < t_{b2}$.
Under the proposed DBI policy, the expected total cost per chip consists of the expected direct burn-in, burn-in failure, and field failure costs per chip for each group and the handling cost per chip \( c_4 \) for classifying the chips into two groups. That is,

\[
C_{DBI}(t_{b1}, t_{b2}, r_c) = \left[ \left( \text{direct burn-in cost} \right)_{1} + \left( \text{burn-in failure cost} \right)_{1} + \left( \text{field failure cost} \right)_{1} \right] \\
+ \left[ \left( \text{direct burn-in cost} \right)_{2} + \left( \text{burn-in failure cost} \right)_{2} + \left( \text{field failure cost} \right)_{2} \right] + c_4
\]

\[
= c_1 t_{b1} \sum_{i=0}^{r} u(i) + c_2 \sum_{i=0}^{r} u(i) \left[ 1 - R(t_{b1}; i) \right] + c_3 \sum_{i=0}^{r} u(i) \left[ R(t_{b1}; i) - R(t_{b1} + t_f; i) \right] \\
+ c_1 t_{b2} \sum_{i=r_c+1}^{\infty} u(i) + c_2 \sum_{i=r_c+1}^{\infty} u(i) \left[ 1 - R(t_{b2}; i) \right] + c_3 \sum_{i=r_c+1}^{\infty} u(i) \left[ R(t_{b2}; i) - R(t_{b2} + t_f; i) \right] \\
+ c_4
\]

In general, \( c_4 \), the cost per chip for sorting and handling with respect to \( r \), is smaller than \( c_1 \) since sorting chips according to their \( r \) values at the WP test and handling the sorted chips in the assembly process and at the start of burn-in can be mostly automated.

**Derivation of** \( R(t_f) \): It can be proved using a similar argument as in appendix B that the field reliability function \( R(t_f) \) under the DBI policy can be expressed as:

\[
R(t_f) = \frac{\sum_{i=0}^{r} u(i) R(t_{b1} + t_f; i) + \sum_{i=r_c+1}^{\infty} u(i) R(t_{b2} + t_f; i)}{\sum_{i=0}^{r} u(i) R(t_{b1}; i) + \sum_{i=r_c+1}^{\infty} u(i) R(t_{b2}; i)}.
\]

(16)

Then, \( AFR(t_f) \) can be calculated using (13).

The optimization problem for the DBI policy can be formulated as for the SBI policy (see (15)). Again, \( u(i) \) and \( R(\cdot; i) \) are replaced with their estimates \( \hat{u}(i) \) and
\( \hat{r}(\cdot; i) \), respectively, and the upper limit of each summation with respect to \( i \) is replaced with \( N \) as an approximation.

5.3 **Comparison of SBI and DBI policies for 256M DRAM product**

For the 256M DRAM product considered in the present study, the relevant costs are estimated as: \( c_1 = 1 \), \( c_2 = 20 \), and \( c_3 = 40 \). Note that \( c_1 \) is standardized and the other costs are estimated relative to \( c_1 \). It is assumed that the warranty period \( t_f \) is one year (i.e. approximately 9,000 hrs) and \( AFR_{req} \) for one year in the field is 1,000 FITs where one FIT corresponds to one failure in \( 10^9 \) hrs. In addition, the AF value for the 256M DRAM product under the current burn-in condition (specified by the applied temperature and voltage) is estimated as 20,000.

Since the burn-in duration is usually determined in an increment of one hour in practice, the optimal burn-in duration \( t_b \) for the SBI policy is also determined in one-hour unit using a grid search method over the search region \([0, 96]\). Similarly, a grid search method is used to determine optimal values of \( r_{b1} \), \( t_{b1} \) and \( t_{b2} \) for the DBI policy. Note that \( r_{c1} \) can take values from zero to \( N \). Optimal values of \( t_{b1} \) and \( t_{b2} \) are also determined in one-hour unit over the search region \([0, 96]\).

For a comparison of the SBI and DBI policies, each policy is optimized and the following \( \Delta \) is calculated.

\[
\Delta = (\text{optimal expected total cost per chip for the SBI policy}) - (\text{optimal expected total cost per chip except } c_4 \text{ for the DBI policy}).
\]

That is, if \( c_4 \) is smaller than \( \Delta \), then the proposed DBI policy is preferred to the SBI policy, and vice versa. Table 5 shows the optimal SBI and DBI policies and the corresponding expected costs. The optimal DBI policy is to classify the chips that pass
the WP test into two groups using the boundary value 49 of $r$, and to employ the burn-in durations of 6 and 29 hrs for the first and second groups, respectively. Since the proportion of the first group is 96.35%, most of the chips that pass the WP test are subjected to the shorter burn-in duration, 6 hrs, and the expected burn-in duration per chip is 6.84 hrs. For the SBI policy, the optimal burn-in duration turns out to be 8 hrs. Note that $\Delta = 1.401$. Since $c_4$ is usually smaller than $c_1 (= 1)$, the DBI policy is more cost-effective than the SBI policy for the product considered.

>> Table 5 <<

Since uncertainties may exist in the estimates of $c_2$, $c_3$, and AF, a sensitivity analysis is conducted with respect to these parameters. It is first assumed that the true values of $c_2$, $c_3$, and AF lie within $\pm 30\%$ of the estimated values. Then, for each parameter, three true values are selected as

(70\% of the estimate, 100\% of the estimate, 130\% of the estimate).

This results in 27 combinations of true parameter values, and for each combination, the true expected total cost per chip is calculated for each policy which was determined using the estimated parameter values. Table 6 shows the sensitivity analysis results. It is first noticed that $\Delta$ values change from 1.318 to 1.481, implying that the advantage of the proposed DBI policy is still maintained under the assumed uncertainties in the parameter estimates. On the other hand, the actual AFR values become greater than the required value ($= 1,000$ FITs) for both policies when the true AF is smaller than the estimated one. This implies that an overestimation of AF should be avoided as much as possible.

>> Table 6 <<
6. Discussions

In this paper, a DBI policy is proposed and its effectiveness is demonstrated for a certain type of defect-tolerant 256M DRAM products. A distinct and desirable feature of the proposed policy is that it explicitly utilizes the information on the quality level (i.e. the number of repairs) of a chip before burn-in. In addition, the proposed policy does not require that such parameters as ‘average number of killer defects’, ‘ratio of the average number of latent defects to that of killer defects’, etc. be known or estimated since it is based on empirically derived models. Finally, it can be applied to any semiconductor product as long as it can be repaired and the burn-in AF is known.

A cautionary remark is in order about the PH model employed in the present investigation. Inferences based on the PH model involve extrapolation in stress, and possibly, in time. That is, the test results from accelerated BI conditions with corresponding BI durations are used to predict the reliability at a certain point in time in the field. The danger of extrapolation with a statistical (or empirical) model is well known, and therefore, it should be properly dealt with when developing and using such a model. In the present investigation, extrapolation in stress is performed using the AF between the BI and field conditions, and the uncertainty involved in the assumed AF affects the accuracy of the predicted reliability in the field. One way of dealing with such uncertainty is to evaluate the sensitivity of the final results with respect to the AF as is done in Section 5.3. If the final results (e.g., superiority of one BI policy over the other) turns out to be sensitive, then every possible effort should be made to estimate the AF as precisely as possible. To avoid or alleviate the danger of extrapolation in time, it is recommended that, in the BI experiment to generate data for developing the model in (8), the BI durations $t_1$ and $t_2$ should be selected to cover a sufficiently wide time span (e.g., $t_1 = 3$ and $t_2 = 30$ hrs). Note that $t_f$ is the field time at which the reliability is to
be predicted, and when translated into the time in the BI time scale it becomes relatively small. For instance, in the present investigation, $t_f$ is one year in the field, and 0.45 hrs ($= 1 \text{ year} / AF = 9,000 \text{ hrs}/20,000$) in the BI time scale. Therefore, if the interval $[t_1, t_2]$ is wide enough, $t_{b_1} + t_f$ and $t_{b_2} + t_f$ for the DBI policy or $t_b + t_f$ for the SBI policy (see Fig. 4) can be made to lie within the interval, and subsequently, extrapolation can be avoided. In the present investigation, $t_1 = 3$ and $t_2 = 6$ hrs for the BI experiment, $t_{b_1} = 6$ and $t_{b_2} = 29$ hrs for the DBI policy, and $t_b = 8$ hrs for the SBI policy. Note that the degree of extrapolation in time for the first group of the DBI policy is not severe since $t_{b_1} + t_f$ ($= 6.45\text{hrs}$) is slightly greater than $t_2$, while the SBI policy and the second group of the DBI policy involve a mild and rather severe extrapolation in time, respectively.

This was inevitable due to the limited availability of experimental BI data at the time of analysis, and should be taken into consideration when interpreting the results of the present investigation.

There may exist other indicator variables for the quality level of chips than the number of repairs considered in this paper. For instance, for a given chip, the number of its neighborhood chips that fail to pass the WP test can be used as such an indicator [13]. In addition, suppose that chip ‘$a$’ is located in a certain XY position on wafer ‘$A$’ in a lot. Then, among the chips located in the same XY position as ‘$a$’ on the wafers other than ‘$A$’ in the same lot, the number of chips failed to pass the WP test may also serve the purpose of indicating the quality level of ‘$a$’ [28]. Our on-going research is focused on developing DBI policies considering such indicator variables either individually or in combination. In addition, it may be a fruitful area of future research to develop a policy with more than two groups and compare its effectiveness to the present one.
Appendix A: Derivation of $w_{ij}$ in (7).

$R_{ij}'$ is defined as $n_{w_{ij}} / N_{w_{ij}}$ where $n_{w_{ij}}$ is distributed as Binomial $\left(N_{w_{ij}} \cdot p_{ij}\right)$.

Therefore, $E\left(R_{ij}'\right) = p_{ij}$ and $\text{var}\left(R_{ij}'\right) = p_{ij} \left(1 - p_{ij}\right) / N_{w_{ij}}$. Since $y_{ij}' = \ln\left(-\ln R_{ij}'\right)$, a Taylor series expansion of $y_{ij}'$ around $p_{ij}$ yields

$$y_{ij}' = \ln\left(-\ln R_{ij}'\right) \approx \ln\left(-\ln p_{ij}\right) + \left[1 / \left(p_{ij} \ln p_{ij}\right)\right] \left(R_{ij}' - p_{ij}\right)$$

Therefore, $E\left(y_{ij}'\right) \approx \ln\left(-\ln p_{ij}\right)$ and

$$\text{var}\left(y_{ij}'\right) \approx \left(\frac{1}{p_{ij} \ln p_{ij}}\right)^2 \text{var}\left(R_{ij}'\right) = \frac{1}{p_{ij}^2 \left(\ln p_{ij}\right)^2} \frac{p_{ij} \left(1 - p_{ij}\right)}{N_{w_{ij}}}$$

$$= \frac{1}{p_{ij} \left(\ln p_{ij}\right)^2} \frac{1 - p_{ij}}{N_{w_{ij}}} \approx \frac{1}{R_{ij}' \left(\ln R_{ij}'\right)^2} \frac{1 - R_{ij}'}{N_{w_{ij}}}$$

Finally, $w_{ij}$ is defined as the reciprocal of $\text{var}\left(y_{ij}'\right)$.

Appendix B: Derivation of $\Pr\left(T < t_b \mid \text{WP passed}\right)$.

$$\Pr\left(T < t_b \mid \text{WP passed}\right)$$

$$= \sum_{i=0}^{\infty} \Pr\left(T < t_b \mid r = i \mid \text{WP passed}\right)$$

$$= \sum_{i=0}^{\infty} \Pr\left(T < t_b \mid r = i \mid \text{WP passed}\right) / \Pr\left(\text{WP passed}\right)$$

$$= \sum_{i=0}^{\infty} \frac{\Pr\left(r = i \mid \text{WP passed}\right) \Pr\left(T < t_b \mid r = i \mid \text{WP passed}\right)}{\Pr\left(\text{WP passed}\right)}$$

$$= \sum_{i=0}^{\infty} \Pr\left(r = i \mid \text{WP passed}\right) \Pr\left(T < t_b \mid r = i \mid \text{WP passed}\right)$$

$$= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a(i) \left[1 - R(t_b, i)\right]$$
References


For a given type of defect-tolerant memory product, collect repair data at the WP test and burn-in test data using two or more different burn-in durations.

Fit a negative binomial distribution to repair data. Fit an extended PH model to burn-in data to estimate the reliability function.

Develop a cost model considering direct burn-in, burn-in failure, field failure, and handling costs with a constraint on AFR.

Optimize the cost model with respect to $r_c$, and burn-in durations $t_{b1}$ and $t_{b2}$.

Compare the performance of the DBI policy with that of the SBI policy. Conduct sensitivity analysis, if necessary, with respect to the assumed parameters.

Fig. 1. Procedures for developing a DBI policy.

Fig. 2. A negative-binomial distribution fitted to the repair data for 256M DRAM (values of the number of repairs are not included for proprietary reasons. The upper tail portion is not shown).
Fig. 3. Log transformed burn-in yield vs. the number of repairs for 256M DRAM burn-in data (Values of the number of repairs are not included for proprietary reasons. The cases where $n_{wij}$ is below 20 are excluded).

Fig. 4. Burn-in and field times.
TABLES

Table 1
ANOVA table for model (2): Negative-binomial distribution fitted to 256M DRAM repair data.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Approx. Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>17695.7</td>
<td>8847.8</td>
<td>620.31</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>185</td>
<td>2638.8</td>
<td>14.2636</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncorrected Total</td>
<td>187</td>
<td>20334.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2
Estimates and confidence limits of parameters in model (2): Negative-binomial distribution fitted to 256M DRAM repair data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Approx. Std Error</th>
<th>Approximate 95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>10.7759</td>
<td>0.5043</td>
<td>9.7810 - 11.7707</td>
</tr>
<tr>
<td>$w$</td>
<td>0.7282</td>
<td>0.0101</td>
<td>0.7082 - 0.7481</td>
</tr>
</tbody>
</table>

Table 3
ANOVA table for model (8): 256M DRAM repair and burn-in data.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Approx. Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>527.55</td>
<td>175.90</td>
<td>209</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Residual Error</td>
<td>118</td>
<td>99.31</td>
<td>0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>121</td>
<td>626.86</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4
Estimates of parameters in model (8): 256M DRAM repair and burn-in data.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>Std Error of Coefficient</th>
<th>t</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-2.546800</td>
<td>0.21130</td>
<td>-12.10</td>
<td>0</td>
</tr>
<tr>
<td>$\ln t$</td>
<td>0.018100</td>
<td>0.13550</td>
<td>0.13</td>
<td>0.894</td>
</tr>
<tr>
<td>$i$</td>
<td>-0.008438</td>
<td>0.00612</td>
<td>-1.38</td>
<td>0.171</td>
</tr>
<tr>
<td>$i \times \ln t$</td>
<td>0.020981</td>
<td>0.00386</td>
<td>5.43</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 5
Optimal burn-in policies.

<table>
<thead>
<tr>
<th>r_c</th>
<th>t_{b1}</th>
<th>Proportion of chips with t_{b1}</th>
<th>t_{b2}</th>
<th>Cost</th>
<th>SBI Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>6</td>
<td>96.35%</td>
<td>29</td>
<td>11.085</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DBI Policy</th>
<th>SBI Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>t_b</td>
<td>Cost</td>
</tr>
<tr>
<td>8</td>
<td>12.486</td>
</tr>
</tbody>
</table>

Table 6
Sensitivity analysis results.

<table>
<thead>
<tr>
<th>AF</th>
<th>c_2</th>
<th>c_3</th>
<th>True AFR</th>
<th>Actual AFR</th>
<th>Actual Cost</th>
<th>Actual AFR</th>
<th>Actual Cost</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>14000</td>
<td>14</td>
<td>28</td>
<td>1420</td>
<td>9.895</td>
<td>1394</td>
<td>11.222</td>
<td>1.327</td>
<td></td>
</tr>
<tr>
<td>14000</td>
<td>14</td>
<td>40</td>
<td>1420</td>
<td>10.018</td>
<td>1394</td>
<td>11.340</td>
<td>1.322</td>
<td></td>
</tr>
<tr>
<td>14000</td>
<td>14</td>
<td>52</td>
<td>1420</td>
<td>10.140</td>
<td>1394</td>
<td>11.458</td>
<td>1.318</td>
<td></td>
</tr>
<tr>
<td>14000</td>
<td>20</td>
<td>28</td>
<td>1420</td>
<td>11.083</td>
<td>1394</td>
<td>12.485</td>
<td>1.402</td>
<td></td>
</tr>
<tr>
<td>14000</td>
<td>20</td>
<td>40</td>
<td>1420</td>
<td>11.206</td>
<td>1394</td>
<td>12.603</td>
<td>1.397</td>
<td></td>
</tr>
<tr>
<td>14000</td>
<td>20</td>
<td>52</td>
<td>1420</td>
<td>11.328</td>
<td>1394</td>
<td>12.721</td>
<td>1.393</td>
<td></td>
</tr>
<tr>
<td>14000</td>
<td>26</td>
<td>28</td>
<td>1420</td>
<td>12.271</td>
<td>1394</td>
<td>13.747</td>
<td>1.476</td>
<td></td>
</tr>
<tr>
<td>14000</td>
<td>26</td>
<td>40</td>
<td>1420</td>
<td>12.394</td>
<td>1394</td>
<td>13.866</td>
<td>1.472</td>
<td></td>
</tr>
<tr>
<td>14000</td>
<td>26</td>
<td>52</td>
<td>1420</td>
<td>12.516</td>
<td>1394</td>
<td>13.984</td>
<td>1.468</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>14</td>
<td>28</td>
<td>999</td>
<td>9.811</td>
<td>980</td>
<td>11.140</td>
<td>1.329</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>14</td>
<td>40</td>
<td>999</td>
<td>9.898</td>
<td>980</td>
<td>11.224</td>
<td>1.327</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>14</td>
<td>52</td>
<td>999</td>
<td>9.984</td>
<td>980</td>
<td>11.307</td>
<td>1.323</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>20</td>
<td>28</td>
<td>999</td>
<td>10.999</td>
<td>980</td>
<td>12.403</td>
<td>1.404</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>20</td>
<td>40</td>
<td>999</td>
<td>11.085</td>
<td>980</td>
<td>12.486</td>
<td>1.401</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>20</td>
<td>52</td>
<td>999</td>
<td>11.172</td>
<td>980</td>
<td>12.569</td>
<td>1.397</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>26</td>
<td>28</td>
<td>999</td>
<td>12.187</td>
<td>980</td>
<td>13.666</td>
<td>1.479</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>26</td>
<td>40</td>
<td>999</td>
<td>12.273</td>
<td>980</td>
<td>13.749</td>
<td>1.476</td>
<td></td>
</tr>
<tr>
<td>20000</td>
<td>26</td>
<td>52</td>
<td>999</td>
<td>12.359</td>
<td>980</td>
<td>13.832</td>
<td>1.473</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>14</td>
<td>28</td>
<td>771</td>
<td>9.766</td>
<td>755</td>
<td>11.096</td>
<td>1.331</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>14</td>
<td>40</td>
<td>771</td>
<td>9.832</td>
<td>755</td>
<td>11.160</td>
<td>1.328</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>14</td>
<td>52</td>
<td>771</td>
<td>9.899</td>
<td>755</td>
<td>11.224</td>
<td>1.325</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>20</td>
<td>28</td>
<td>771</td>
<td>10.953</td>
<td>755</td>
<td>12.359</td>
<td>1.406</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>20</td>
<td>40</td>
<td>771</td>
<td>11.020</td>
<td>755</td>
<td>12.423</td>
<td>1.403</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>20</td>
<td>52</td>
<td>771</td>
<td>11.087</td>
<td>755</td>
<td>12.487</td>
<td>1.400</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>26</td>
<td>28</td>
<td>771</td>
<td>12.141</td>
<td>755</td>
<td>13.622</td>
<td>1.481</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>26</td>
<td>40</td>
<td>771</td>
<td>12.208</td>
<td>755</td>
<td>13.686</td>
<td>1.478</td>
<td></td>
</tr>
<tr>
<td>26000</td>
<td>26</td>
<td>52</td>
<td>771</td>
<td>12.274</td>
<td>755</td>
<td>13.750</td>
<td>1.476</td>
<td></td>
</tr>
</tbody>
</table>

a: c_4 not included.