ALPHA & OMEGA
SEMICONDUCTOR

Power Semiconductor Reliability Handbook

Alpha and Omega Semiconductor

475 Oakmead Pkwy
Sunnyvale, CA 94085
U.S.A.

Rev. 1.0 • 5/20/10

Table of Contents

1 The AOS Reliability Program

In a power device application, high power is usually encountered. AOS strives to make power devices reliable for their intended application. In order to achieve this goal, the reliability activities are spread throughout all phases of a product's lifetime.

1.1 Design-in of Reliability

During the design phase, especially when a new platform for new technology is involved, thorough calculations and simulations are carried out to ensure the designed electrical parameters and other reliability characteristics are optimized. For example, when the breakdown voltage needs to be changed, the simulation involves all the structure changes to optimize the breakdown voltage, threshold voltage, channel resistance, various parasitic capacitances, and the trade-off with UIS robustness, etc.

For a new package design, thermal resistance is very important. A thorough simulation using 3-D finite element analysis has to be carried out.

1.2 Technology Development

Once a design is done and reviewed by the cross-functional team, new silicon wafers and packages need to be built to verify the design. Based on the design and simulations, a thorough design of experiment (DOE) needs to be designed. Split runs and some unit process experiments need to be carried out. Electrical and physical analyses need to be performed.

Then, the best leg of the DOE is chosen as the baseline process. During this period of development, some of the device or test structures are chosen, either in the wafer-level form or the final package form, to do the reliability stress test to verify the robustness of the new technology. This initial reliability evaluation helps to further improve the design and process technology, if any weakness is found.

If some fundamental process modules or new devices need to be studied, quite often some specially designed test structures have to be run through the process and get characterized (e.g., hot-carrier lifetime, electromigration, and oxide integrity, etc.).

1.3 Qualification and Product Development

During this phase of development, the wafer process and package technology are defined and frozen. A product vehicle for qualification needs to be chosen. A formal qualification plan is generated. Full qualification tests need to be carried out, including life tests and environmental tests. The lot requirement and sample size have to follow an industry standard such as JEDEC (JESD47D). The qualification lot has to be run by production personnel without special attention or treatment, in order to reflect the true manufacturing process in the future.

1.4 Pre-production

Once the formal qualification is done on the frozen process and it passes all the stress tests, a larger quantity of the device is run to check the manufacturing process. Yield and some of the reliability monitoring data are reviewed closely to ensure there is no mass-production problem.

During this period of time, an initial production management program may need to be implemented, such as increasing sample size for in-line process monitoring. A reliability monitoring program and even some reliability screening programs need to be installed to ensure the product is reliable.

1.5 Mass Production

A formal monitoring program needs to be implemented during the lifetime of the product to ensure the reliability of the product remains the same. Requalification has to be done if any material or process changes that will impact the form and fit or quality and reliability. Customers will be informed of major changes to ensure there is no problem with their system production.

1.6 Customer Feedback and Failure Analysis

AOS provides customers with reliable product throughout the lifetime they use the device in their system. During the design-in period for the system, if any issue occurs, AOS dispatches a field application engineer (FAE) and system application engineer (AE) to work with customers to solve the problem.

Quite often, the problem is an application issue that can be solved by changing to a part with the right electrical parameters, slightly changing the system circuit topology, or using different values for some of the peripheral components. If the problem identified belongs to the manufacturing of the parts, AOS will implement corrective action to prevent future recurrences.

If a customer encounters device malfunction during their system production or from field failure, this is a very serious matter to AOS. Detailed, speedy failure analysis should be carried out to prevent the customer from having to shut down their production line. However, sometimes we prove the real root cause to be other component problem causing the part to overstress (e.g., noise spike from the MOSFET driver or the magnetic saturation of the coil occurred). But before we can prove it is not our issue, we have to continue searching for the real root cause until the customer is satisfied.

1.7 Continuous Improvement

Customer requirements and harsh market demands for improved reliability will never end. AOS realizes that continuous improvement is a must to stay in the power business. New product development and technology development should always put reliability as the first priority. The continuous cycle of design, qualification, mass production, evaluation, and customer feedback will enable us to further improve the reliability of our product.

2 Fundamentals of Reliability

2.1 Definition of Reliability

There are many words or phrases that can describe the concept of reliability, such as durability, a product's quality over time, a measure of future and ability to function normally over time, etc. But the most rigorous definition is as follows:

> *"The probability that an item will perform a required function under specified conditions for a specified period of time."*

This definition identifies three important independent concepts:

1. Duration of time
2. Environmental conditions of use
3. Functioning parameter value or range.

The duration of time is usually closely related to the customer's requirement for their system warrantee period. Environmental conditions are usually defined by the maximum ratings of the data sheet. Functional parameters are also specified in the data sheet.

2.2 Bathtub Curve

The bathtub curve was first used by life insurance companies to describe life expectancy. It is divided into three distinct regions (**Figure 1**). Semiconductor reliability behaves the same way.

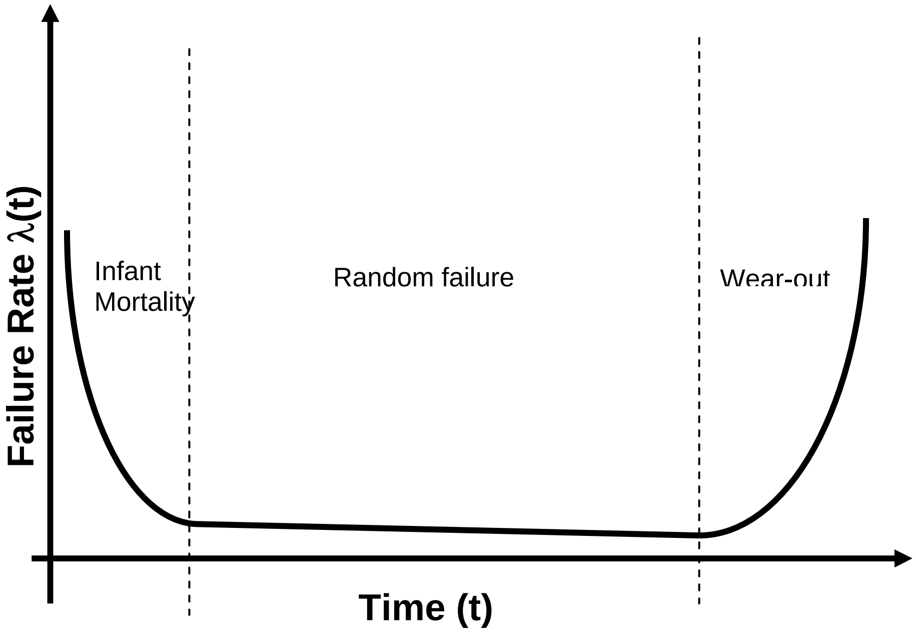

Figure 1. The bathtub curve identifies three stages of product failure over the life of the product: infant mortality, random failure, and wear-out.

2.2.1 Infant-Mortality Region

This period of time can also be called the initial failure period. During this period, the failure rate is in the decreasing mode. Usually, the failure is caused by a manufacturing defect, such as contaminating particles and imperfection of the circuit image from the lithography process.

Nowadays, in a modern wafer fab with good environmental and process controls, the defect density can be very low. Therefore, the infant mortality can be very low. If the infant mortality rate is higher than the customer wants, a screening test can be implemented to weed out the unreliable parts.

In this period of time, the failure rate can be fitted using a Weibull distribution function, which will be discussed later.

2.2.2 Random-Failure Region

In this period, most of the manufacturing defects have already attenuated, so the failure rate will be low and almost constant. However, in this period, the parts can also encounter some external stress, such as a system voltage spike, ESD attack, etc., which can happen randomly. The failure rate in this period can be expressed by an exponential distribution function, which is a special case of the Weibull distribution with shape parameter m = 1

2.2.3 Wear-Out Region

There are certain wear-out mechanisms in a semiconductor device, such as electromigration, hot carrier degradation of a short-channel device, and gate-oxide wear-out. These failure mechanisms were harmful back in the 1980s, when semiconductor devices were very aggressively scaled down.

Since then, the IC industry has put a lot of effort into understanding and characterizing failure mechanisms. Nowadays, most of the problems can be addressed by design-in of reliability with the device lifetime goal set at several decades. Most of the time, if the design rules given by a reputable silicon foundry are followed, the wear-out region will not be a concern for our customers.

3 Reliability Test Methods

3.1 Accelerated Life Tests

Reliability is a prediction of the future. How can the future be predicted? The accelerated method has to be used in order to get answers in a reasonable time. Using higher-than-normal stress, failure can be induced earlier than usual. Once the lifetime under accelerated stress is obtained, the known accelerated model can be used to predict the product lifetime under normal application stress. But care must be employed to ensure the stress level is reasonable. Over-stress should not occur. This concept is explained in **Figure 2**.

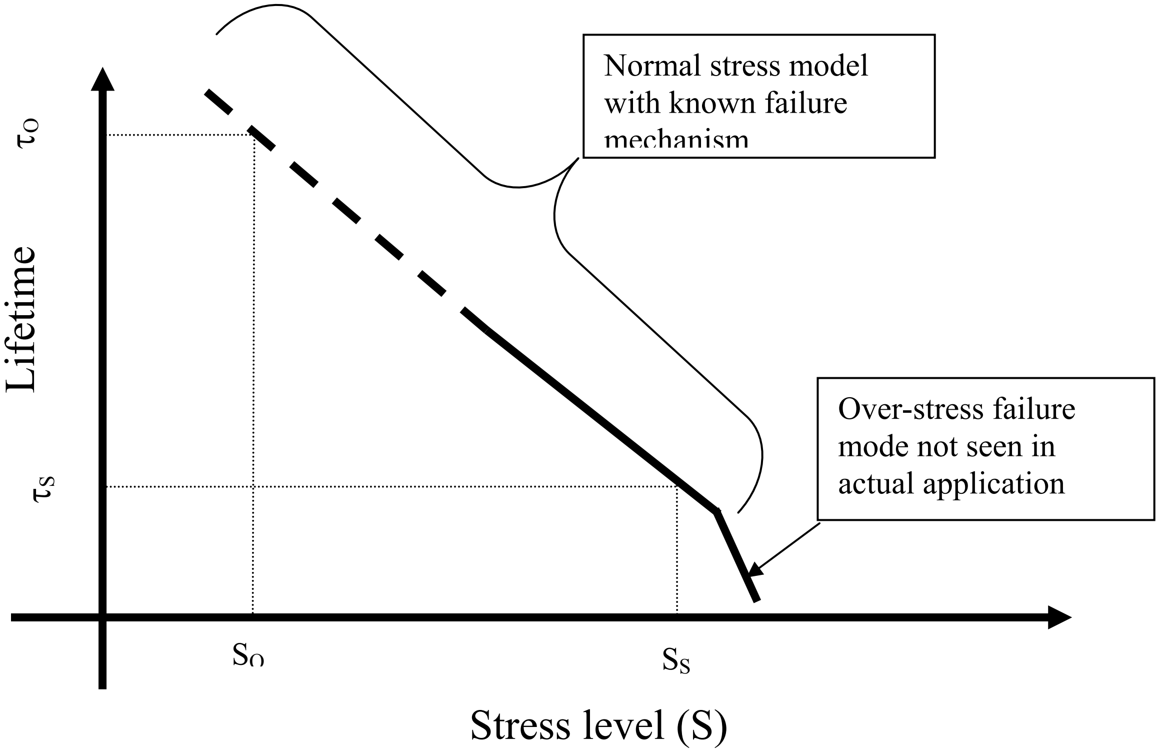
Normal stress model with known failure mechanism
Over-stress failure mode not seen in actual application
τ_O
Lifetime
τ_S
S_O
S_S
Stress level (S)

Figure 2. Once τ_S is obtained experimentally, the product lifetime under normal operation (τ_O) can be predicted using the known stress model.

This diagram is just for illustrating the concept. Quite often, both axes are in logarithm scale. S_S stands for the stress level during the reliability test that results in a product lifetime of τ_S. Using the known stress model and the operating stress (S_O) expected in the application, the product lifetime under normal operation (τ_O) can be predicted.

The solid line in this model represents some historical data that is available to support this model, and the dashed line represents the extrapolation of this model. Over-stress will also create failures, but the mode of failure may not be seen in normal operation. One cannot use the data from this region to make the prediction.

3.1.1 Temperature Acceleration

Most of the defect-related failures and wear-out mechanisms can be accelerated by using temperatures higher than the normal application temperature to induce the failure, so the reliability test and assessment can be predicted within a reasonable time. The induced failure usually involves some kind of chemical or physical reaction. The famous Arrhenius model can be derived as follows:

Reaction rate $\propto$ exp $\{-E_a/kT\}$

Time to failure $\tau \propto$ 1/Reaction rate $\propto$ exp $\{ E_a/kT\}$; therefore:

$$\tau = A \times \exp\left(\frac{E_a}{kT}\right)$$

where E_A is the activation energy, k is the Boltzmann constant = $8.617 \text{x} 10^{-5}$ ev, and A is a constant.

Suppose the stress temperature is T_S, the lifetime obtained is τ_S, the operating temperature is T_O, and the predicted lifetime is τ_O. In that case, the Arrhenius equation can be expressed with a more practical form as a temperature acceleration factor (AF_T)

$$AF_T = \tau_O / \tau_S = e^{\frac{E_a}{k}\left(\frac{1}{T_o} - \frac{1}{T_s}\right)}$$

Let’s try an example. A 1000-hr high-temperature life test at 150°C is equivalent to how many hours of operating life at 55°C?

Assuming the activation energy is 0.7 eV,

$$AF = \exp\{(0.7/8.617 \text{ x } 10^{-5}) \text{ x } [1/(273+55)-1/(273+150)]\} = 260$$

The total equivalent operating time = test time x AF
= 1000 hr x 260
= 21.6 years

3.1.2 Activation Energy

The activation energy for different failure mechanisms have been well characterized in the semiconductor industry. The following table shows some of the typical values.

Mechanism	Temperature Ea (eV)
Gate-oxide defect	0.3
Intermetallic defect	0.3
Poly to metal defect	0.3
Silicon junction defect	0.8
Masking defect	0.5
Electromigration	0.5
Contamination	1.0
Assembly	0.5
Hot carrier	-1.0
Intermetallic growth	1.0
Corrosion	0.3 to 1.1

All the E_a are positive except the hot carrier, E_a = -1 eV. At a lower temperature, the hot carrier has less scattering from lattice vibration; therefore, the degradation is faster. So, for the hot carrier, the low temperature needs to be used for the accelerated test.

3.1.3 Voltage Acceleration

The voltage-stress failure mechanism can be complicated. Depending on the device structure and type, the model can be very different. However, for MOS devices, especially for gate-oxide resistance to voltage stress, the Eyring-exponential model works well. The lifetime τ can be expressed as a function of stress voltage (V_S):

$$\tau = A \text{ x } \exp\{- V_S \text{ x } \beta\},$$

where A is a constant and β is the voltage acceleration coefficient for a given failure mechanism. If the lifetime τ_S is obtained from V_s, the operating lifetime is τ_O, and the operating voltage is $V_{O,}$ then the voltage acceleration factor (AF_V) can be calculated as follows:

$$AF_V = \exp\{ \beta \text{ x } (V_S - V_O)\}$$

Some of the voltage accelerations for different failure mechanisms are reported in the table below.

Mechanism	Voltage β (1/v)
Thin-gate-oxide defect	Tox/100
Intermetallic defect	1.5 to 3.0
Poly to metal defect	1.5 to 3.0
Silicon junction defect	0.0 to 0.5

3.1.4 Temperature-Cycle Acceleration

During the operation of a system, the semiconductor device encounters the temperature-cycle stress (e.g., before the power is on, the device is at a room temperature like 25°C; but when the system is on, usually with a good cooling design, the junction temperature will not exceed 55°C).

However, in some of the power devices, the junction temperature can reach as high as 85°C to 100°C. The semiconductor device consists of materials with different thermal expansion coefficients. Therefore, the temperature cycling induces certain failures, such as package cracking, die cracking, wire-bond opening, and a rise in contact resistance.

In order to evaluate the robustness of the device against these failures, a much-higher temperature difference for the cycling is used to evaluate the reliability of the device. The standard temperature cycling applied to a power device alternates the ambient? temperature between -65°C and +150°C. To pass for qualification, the power device must survive 1000 cycles.

The Coffin-Manson model is very popular for this acceleration. The lifetime τ is proportional to the temperature difference ΔT to the nth power.

$$\tau = A \text{ x } (\Delta T)^n,$$

where A is a constant. The acceleration factor can be expressed as follows:

$$AF_{TC} = (\Delta T_S/\Delta T_O)^n,$$

where ΔT_S is the stress temperature difference, ΔT_O is the operation temperature difference, and n is the exponent coefficient for the model.

Let’s look at an example. Thirty temperature cycles from -65ºC to 150ºC is equivalent to how many years of operation in a computer environment having a junction temperature of 55ºC? Assume the computer turns on and off three times a day and the ambient temperature (T_{ROOM}) is 25ºC.

Solution: Using the Coffin-Manson model, the acceleration factor (AF) can be obtained as follows by using n = 6 (for most of the package failures, n ~ 6 was established):

$$AF = \{(150+65)/(55-25)\}^6 = 135489.$$

The equivalent total cycle = 30 x 135489 = 4,064,670 cycles.

Since there are three cycles per day and 365 days per year,

Equivalent years of operation = 4,064,670/3/365 = 3712 years.

3.1.5 Humidity Acceleration

For economical reasons, plastic packages are widely used for semiconductor devices. A plastic package is considered non-hermetic. That means moisture gets into the package via a diffusion process through the plastic molding compound or through the interface between the molding compound and the lead frame. A simple model works well as follows:

$$\tau \ \infty \ (1/RH)^n,$$

where RH is the relative humidity. The acceleration factor can be expressed as

$$AF_H = (RH_S / RH_O)^n$$

Let’s look at an example for a 100-hr highly accelerated stress test (HAST) at 130°C and 85 percent RH. What are the equivalent years of operation? Assume an operating condition with a temperature of 55°C and relative humidity at 65 percent.

The total acceleration can be calculated as follows:

$$AF = AF_H \text{ x } AF_T$$

$$= (85/65)^3 \text{ x } \exp\{(0.9/8.617 \text{ x } 10^{-5}) \text{ x } [1/(273+55)-1/(273+130)]\} = 838,$$

where $n = 3$ and $E_a = 0.9$ eV .

The equivalent year will be 838 x 100 /365/24 = 9.57 years.

Let’s try another example. What will the acceleration factor be when the HAST (130ºC and 85 percent RH) is used instead of the temperature humidity bias (85ºC and 85 percent RH) stress? (HAST and THB are described more at the end of this chapter.)

$$AF = (85/85)^3 \text{ x } \exp\{(0.9/8.617 \text{ x } 10^{-5}) \text{ x } [1/(273+85)-1/(273+130)]\} = 26.$$

This means the 100-hr HAST is equivalent to 2600 hr of 85/85.

4 Reliability Function for Data Analysis

It is very hard to understand the physical meaning of the reliability calculation. The probability statistic was developed a long time ago in other fields. In this author’s opinion, the calculation is a curve-fitting job for the given data. As long as you can choose the function that fits the data and can do the right extrapolation, then it is all right. However, as a reliability engineer, you need to get familiar with all the probability density functions and their applications.

4.1 Definition of Reliability Function

The reliability function indicates the probability or ratio of a population that will survive during the time from 0 up to t. Mathematically, reliability can be specified as follows:

$$R(t) = \frac{n - r(t)}{n}, \qquad \textit{(Eq. 1)}$$

where n is the total number of good samples at time = 0 and r(t) is the cumulative number of samples that fail during the period from time = 0 to time = t.

The boundary conditions are R(0) = 1 (100%) and R(∞) = 0.

4.2 Failure Distribution Function

The failure distribution function can also be called the unreliability function. This indicates the probability or ratio of a population that will fail during the time from 0 up to t. This function also can be called the cumulative failure distribution.

$$F(t) = \frac{r(t)}{n}. \qquad \textit{(Eq. 2)}$$

The boundary conditions are F(0) = 0 and F(∞) = 1

The relationship of the reliability and unreliability functions is as follows:

$$R(t) + F(t) = 1. \qquad \textit{(Eq. 3)}$$

This can be shown in **Figure 3**.

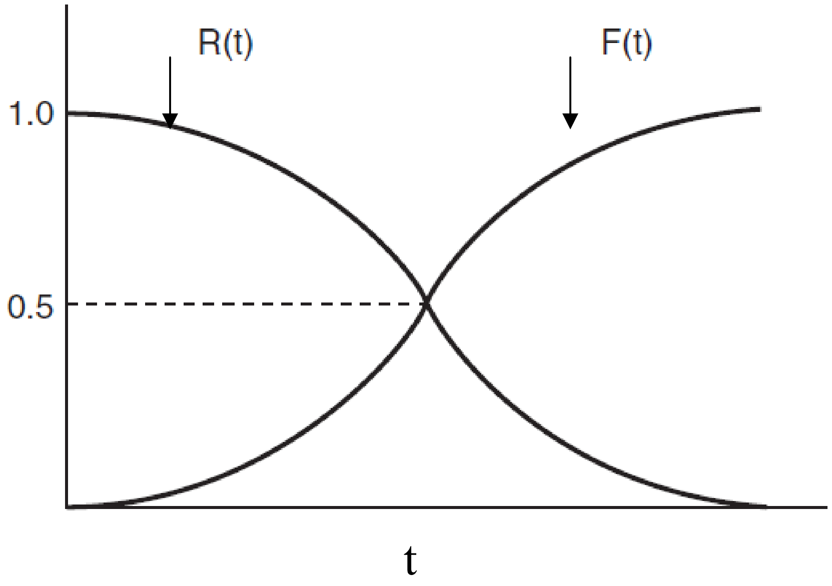

Figure 3. The relationship of the reliability and unreliability functions.

4.3 Failure Probability Density Function

The aforementioned two functions are cumulative functions throughout the entire time span for the total population. The more fundamental characteristics can be expressed by the rate of change of the probability versus time, which is the probability density function. This indicates the ratio of failure occurring during the time t+Δt against the total population:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dR(t)}{dt}. \qquad \textit{(Eq. 4)}$$

All fundamental statistics functions are expressed by the probability density function. All other functions can be derived from this.

The three functions can be further clarified by **Figure 4**.

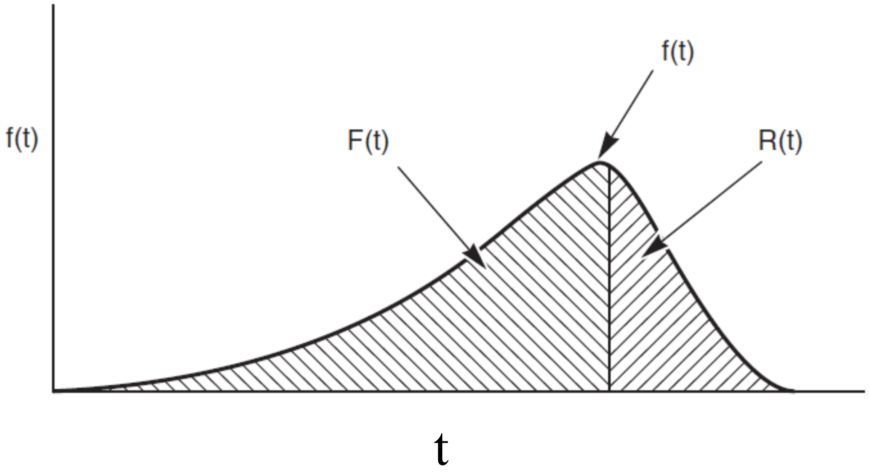

Figure 4. Reliability function (right), failure distribution function (left), and failure probability density function (center).

Once the probability density is chosen, the other functions can be derived from it as follows.

F(t) is the area under function f(t) from t = 0 to t = t . It can be expressed by the following integration:

$$F(t) = \int_0^t f(t)dt. \qquad \textit{(Eq. 5)}$$

R(t) is the area under function f(t) from time t = t to t = ∞. It is expressed by the following integration.

$$R(t) = \int_t^\infty f(t)dt. \qquad \textit{(Eq. 6)}$$

4.4 Failure-Rate Function (Hazard Function)

This is the function we usually use for reliability assessment. The failure rate function is like the bathtub curve mentioned previously. It measures the instantaneous failure rate during the time t+Δt against the remaining good parts that survive through time t.

$$\lambda(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1-F(t)}. \qquad \textit{(Eq. 7)}$$

This is the definition from many textbooks, but the meaning is not readily understood. Let's redefine this function from a different starting point.

The number of failures generated during the time t+Δt interval can be calculated as n x f(t), where n is the total sample size at t = 0 and f(t) is the probability the parts will fail at t+Δt interval. So, this term will be the numerator of λ(t). For the denominator, of course, it is the number of surviving parts at time t, which is [n-r(t)]. The ratio of these two terms is λ(t).

$$\lambda(t) = \frac{n * f(t)}{n - r(t)} = \frac{f(t)}{\frac{n - r(t)}{n}}. \qquad \textit{(Eq. 8)}$$

Eq. 7 is the same as Eq. 8 if you replace the R(t) in Eq. 1.

It is easy to confuse λ(t) vs. f(t), as both are the failure rate at time t, except λ(t) is a measure against the remaining population [n-r(t)] and f(t) is measured against the total original population of [n].

4.5 Unit for Failure Rate

The unit for failure rate is usually expressed in failures per hour. Since this is such a small number, customarily the unit of failure in time (FIT) is used:

1 FIT = One failure per 1 billion device hours

= 1 parts per million (ppm) per 1000 device hours

= 10^{-9} 1/hr.

4.6 Application of Different Probability Functions

4.6.1 Exponential Distribution

This distribution can describe the random failure region very well. It is commonly used for the semiconductor failure rate in the field. The failure probability density function f(t) and reliability function R(t) can be expressed as follows:

$f(t) = \lambda \text{ x } e^{-\lambda \text{ x } t}$

$R(t) = e^{-\lambda \text{ x } t}$.

The mean time to failure (MTTF) is used with non-repairable devices and parts, and the mean time between failure (MTBF) is used for repairable ones. Semiconductor devices are not repairable, so usually MTTF is used:

$$MTTF = \int_t^\infty t \times f(t) dt.$$

By substituting for f(t) and carrying out the integration, it can be proved that for an exponential distribution, MTTF is the inversion of λ.

MTTF = 1/ λ.

This relationship is only true for an exponential distribution; it cannot be used for other distributions.

The meaning of MTTF can also be understood as follows. By substituting t with MTTF in the reliability function R(t), one can obtain the following:

$R(MTTF) = e^{-\lambda \text{ x } (1/\lambda)} = e^{-1} = 0.368.$

This means that at t = MTTF, 36.8 percent of the population is still good or MTTF is the time when 63.2 percent of the population has failed.

4.6.2 Simple-Failure-Rate Calculation

For a random failure region, the simple failure rate can usually be calculated just by averaging the failure rate without going through the complicated distribution function analysis. The average constant failure rate can be defined as follows:

$$\lambda = \frac{r}{n \times t \times AF},$$

where n is the total number of good samples at time = 0, r is the cumulative number of samples failed up to t, and AF is the acceleration factor.

Sometimes this is also called the observed failure rate. If no failure is found up to time t, then the failure rate is zero. However, strictly speaking, this is not right statistically. Because the sample size is always limited, it cannot represent the total population. The χ^2 distribution with an upper-limit confidence level needs to be used to estimate the failure rate:

$$\lambda = \frac{\chi^2[CL,(2r+2)]}{2} \times \frac{1}{n \times t \times AF},$$

where CL is the confidence level in percentage. Some of the $\chi^2/2$ values are given in the following table.

r number of failure	Confidence levels (CL)				
	60%	**80%**	**90%**	**95%**	**99%**
0	0.92	1.6	2.3	3.0	4.6
1	2.0	3.0	3.9	4.7	6.6
2	3.1	4.3	5.3	6.3	8.4
3	4.2	5.5	6.7	1.8	10.0
4	5.2	6.7	8.0	9.2	11.6
5	6.3	1.9	9.3	10.5	13.1
6	1.3	9.1	10.5	11.8	14.6
7	8.4	10.2	11.8	13.1	16.0
8	9.4	11.4	13.0	14.4	11.4
9	10.5	12.5	14.2	15.7	18.8
10	11.5	13.7	15.4	11.0	20.1
11	12.6	14.8	16.6	18.2	21.5
12	13.6	15.9	11.8	19.4	22.8
13	14.6	11.0	19.0	20.7	24.1
14	15.7	18.1	20.1	21.9	25.4
15	16.7	19.2	21.3	23.1	26.7

The general practice in the semiconductor industry is to use a 60 percent confidence level to do the failure-rate calculation.

Let's try an example. During a qualification, 240 samples were used to do the high-temperature operating life (HTOL) stress test at 150ºC. After 500 hr, no failures were found. What is the predicted failure-rate upper limit with 60 percent confidence?

Since no failures were found, the average activation energy of 0.7 eV is used to do the acceleration calculation. A computer environment temperature of 55ºC is chosen to do the de-rating.

$$\begin{aligned} \mathrm{AF} &= \exp\{(E_a/K) \text{ x } [1/(273+T_a) - 1/(273+T_s)]\} \\ &= \exp\{(0.7/8.617 \text{ x } 10^{-5}) \text{ x } [1/(273+55)-1/(273+150)]\} = 260. \end{aligned}$$

By looking up the chi-square half table for zero failures, the value is 0.92; therefore, the failure rate can be estimated as follows:

λ = 0.92 x {1/(240 x 500 x 260)} = 29E-9 = 29 FIT.

4.6.3 Weibull Distribution

The Weibull distribution is the most difficult function to use. However, it is a more general function that can cover most cases. It has three parameters to do the data fitting. The basic probability density function takes the following form:

$$f(t) = \frac{m}{\eta}\left[\frac{t-\gamma}{\eta}\right]^{m-1} \times \exp\left[-\left(\frac{t-\gamma}{\eta}\right)^{m}\right],$$

where m is the shaping parameter, η is the scaling parameter and γ is the position parameter.

The position parameter γ determines when the failures can start to occur. In most cases, there’s a probability of failure at t=0, so it can be set at γ=0 to simplify the calculation. The f(t) function can be expressed simply as:

$$f(t) = \frac{m}{\eta}\left[\frac{t}{\eta}\right]^{m-1} \times \exp\left[-\left(\frac{t}{\eta}\right)^{m}\right].$$

The cumulative failure distribution function F(t) can be obtained by performing integration on f(t):

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^{m}\right].$$

This equation can be rearranged as follows:

ln ln {1/[1-F(t)]} = m(ln t) – m(ln η)

Let y = ln ln {1/[1-F(t)]}

x = ln t

b = -m (ln η)

The equation can reduce to a linear one. It is easy to do the data fitting

Y = mx + b

The failure rate λ(t) can also be derived:

$$\lambda(t) = \frac{m(t)^{m-1}}{\eta^{m}}.$$

The characteristics of the Weibull function can be further understood by looking at **Figure 5**.

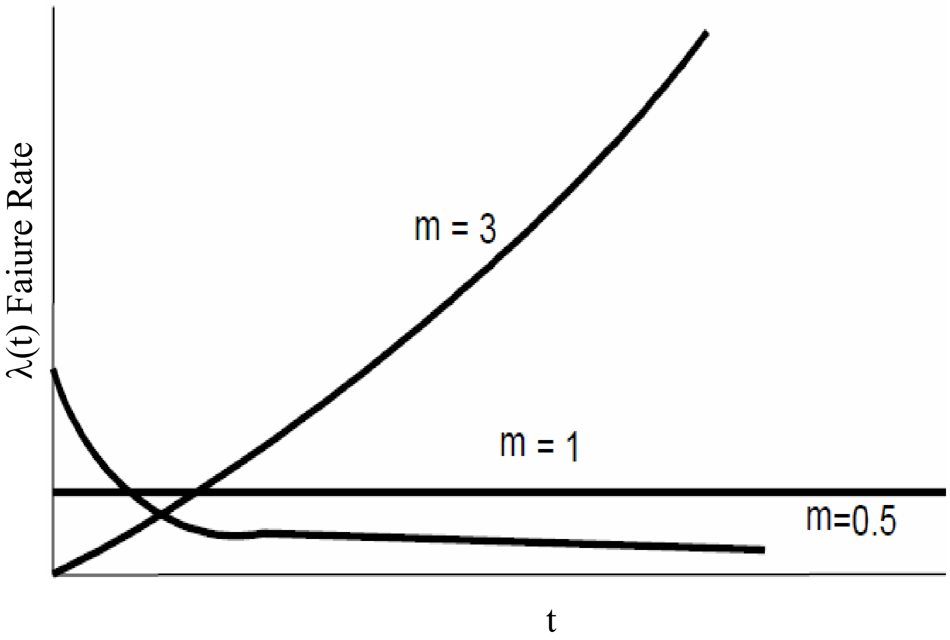

Figure 5. Failure rate plotted for different values of shaping parameter (m).

For the shaping parameter m = 1, the failure rate becomes constant. Therefore, Weibull is reduced to the exponential distribution.

For m < 1, it is a decreasing failure rate. It is useful for doing analysis of the infant mortality region. It is also useful for determining the screening time required to achieve the reliability goal for the customer.

For m >1, it is an increasing failure rate. It is useful for doing the analysis for the wear-out failure mechanism, such as oxide integrity, electromigration and hot carrier degradation. However, in those cases, usually the log-normal distribution is used, which is much simpler to deal with.

4.6.4 Log-Normal Distribution

This distribution can be used when the logarithm of the lifetime t fits the normal distribution. The probability density function f(t) is given as follows:

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t}\exp\left[-\frac{1}{2}\left(\frac{\ln t-\mu}{\sigma}\right)^2\right].$$

The cumulative failure function takes this form after integration:

$$F(t) = \frac{1}{\sqrt{2\pi}\sigma}\int_0^t \frac{1}{x}\exp\left[-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right]dx.$$

MTTF and t_{50} are expressed as follows:

$$\text{MTTF} = \exp(\mu + \sigma^2/2)$$

$$t_{50} = \exp(\mu)$$

In a log-normal analysis, quite often t_{50} is used instead of MTTF. t_{50} is the time when 50 percent of the population has failed. The F(t) function can be expressed closer to the form of a normal distribution:

$$F(t) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\ln t} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \ln t_{50}}{\sigma}\right)^2\right] d\ln t.$$

This means if the logarithm of time t is used as a variable, then it is exactly a normal distribution function.

4.6.5 Application of Log-Normal Distribution

For a wear-out failure mechanism, such as electromigration, hot-carrier-induced degradation or oxide wear-out, the data can be fitted to the log-normal function well, because this function describes an increasing hazard rate. In practice, it is very simple to use with the help of a statistical table or even the function in Excel (NORMSINV) to deal with the F(t) function.

This function can be further reduced to the following form:

$$F(t) = \Phi\{\ln(t/t_{50})/\sigma\},$$

where σ is called the shape factor for the log-normal distribution and function Φ is the area just under the standard normal curve for value "Z." By using the reverse function of Φ, you can obtain the linear equation for curve fitting:

$$Z = \Phi^{-1}[F(t)] = \frac{1}{\sigma}\ln t - \frac{\ln t_{50}}{\sigma}.$$

From the slope and intersect σ and t_{50}, Z can be determined.

Let's try an example. A highly accelerated reliability stress test is done on 2000 samples up to 700 hr. A pull point is at every 100 hr. The number of failures and their log-normal calculations are shown in the following table.

What are the log-normal shape factor σ and t_{50} values for this set of data? The acceleration factor is known, AF = 260. After de-rating to the application condition, what will the total failures be after three years of operation assuming power is on all the time?

Total sample n = 2000					
t (hr)	Number of failures	Cumulative failure i	$F(t) = \frac{i-0.3}{n+0.4}$	ln(t)	$Z = \Phi^{-1}[F(t)]$
100	10	10	0.005	4.61	-2.59
200	25	35	0.017	5.30	-2.11
300	52	87	0.043	5.70	-1.71
400	63	150	0.075	5.99	-1.44
500	70	220	0.110	6.21	-1.23
600	80	300	0.150	6.40	-1.04
700	96	396	0.198	6.55	-0.85

Table 1. Number of failures and their log-normal calculations.

F(t) can be simply expressed as F(t) = i/n. However, to be statistically correct, one needs to use either mean ranking i / (n+1) or median ranking (i-0.3) / (n+0.4). In the previous table the F(t) is calculated by using the median ranking because the log-normal probability density function is skewed. For symmetrical distribution the mean ranking should be applied. The Z value or sigma number is obtained by using Excel function NORMSINV() putting the value in the F(t) column in the argument.

The least-square fit is done by Excel as plotted in **Figure 6**.

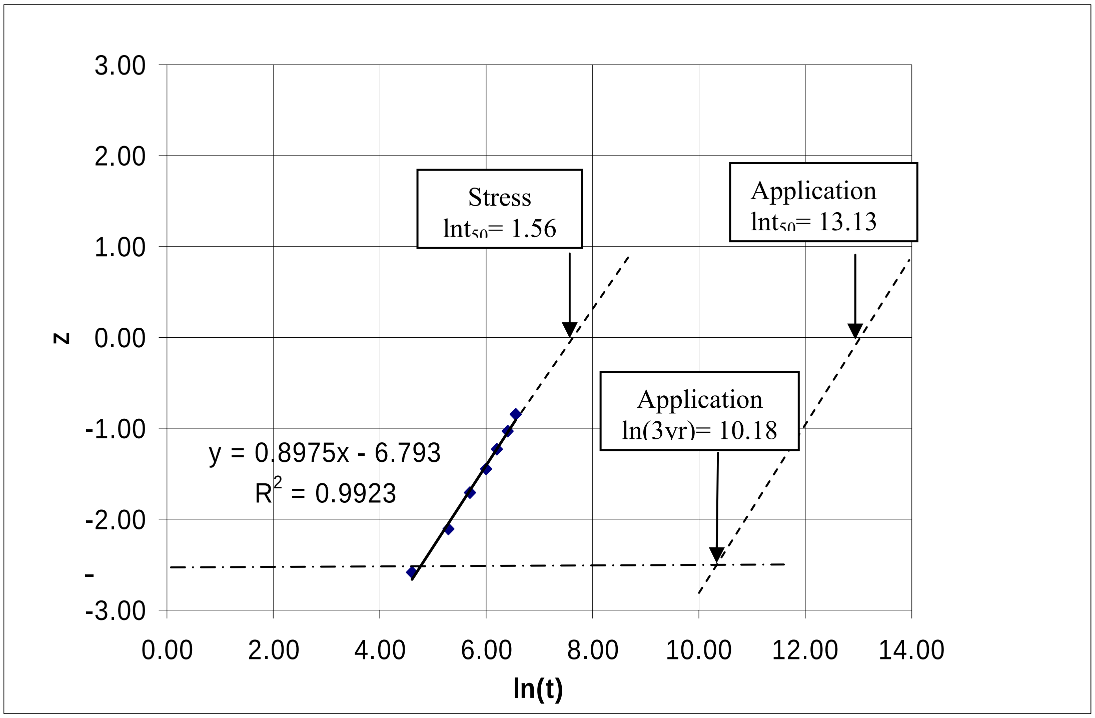

Figure 6. Least-square fit.

By comparing the linear equation with the fitted results:

$1/\sigma = 0.8975$

$\ln t_{50} / \sigma = 6.793.$

Therefore, the log-normal shape factor σ and t_{50} can be obtained:

$\sigma = 1.11$ and $t_{50} = 1937$ hr.

This means if the same stress is continuously applied, it will take 1937 hr to get 50 percent or 1000 parts to fail cumulatively.

The Z equation to describe this set of data can take the form as follows:

$Z = (1/1.11) \times (\ln t - \ln 1937).$

For estimating the total failure in three years, first the t_{50} for the unaccelerated situation needs to be determined. Since the acceleration factor is 260, by definition t_{50} will be

$1937 \times 260 = 503620$ hr

$\ln t_{50} = \ln 503620 = 13.13.$

Once this point is determined, one can draw a line in Z vs. lnt plot through (0,13.13) and parallel to the stressed cumulative line. This line will represent the cumulative failure distribution for the application condition. Now the Z equation can take this form:

$Z = (1/1.11) \times (\ln t - 13.13).$

For three years of power on all the time, t = 24 x 365 x 3 = 26280 hr.

Using this value of t, one can obtain the Z value to be -2.66 by looking up the standard normal table. One can determine the area under the normal curve at Z = -2.66 to be 0.0039, or in three years, the total failure will be 3900 ppm. It is much easier to look at the Excel plot to understand this whole procedure.

5 Technology Stress Tests

During the new wafer process technology development, it is essential to check all the wear-out mechanisms. Once this is characterized and the lifetime for each mechanism is measured quantitatively, then proper design rules can be generated so that design-in of reliability can be achieved.

This type of work is also called process reliability. Much research has been done and papers have been published by the semiconductor industry to improve the intrinsic reliability of the technology against wear-out. Only a brief introduction to these topics is given here.

5.1 Hot-Carrier Lifetime

It is well known that when a device is scaled down, the electric field inside the device is going to be very high. The high field can accelerate the carriers to gain kinetic energy much higher than 3/2kT. These are referred to as hot carriers.

When a hot carrier has enough energy to enter the gate oxide, it will change the V_T as well as mobility. The mechanism can be understood by looking at **Figure 7**, which depicts an n-channel MOSFET in saturation.

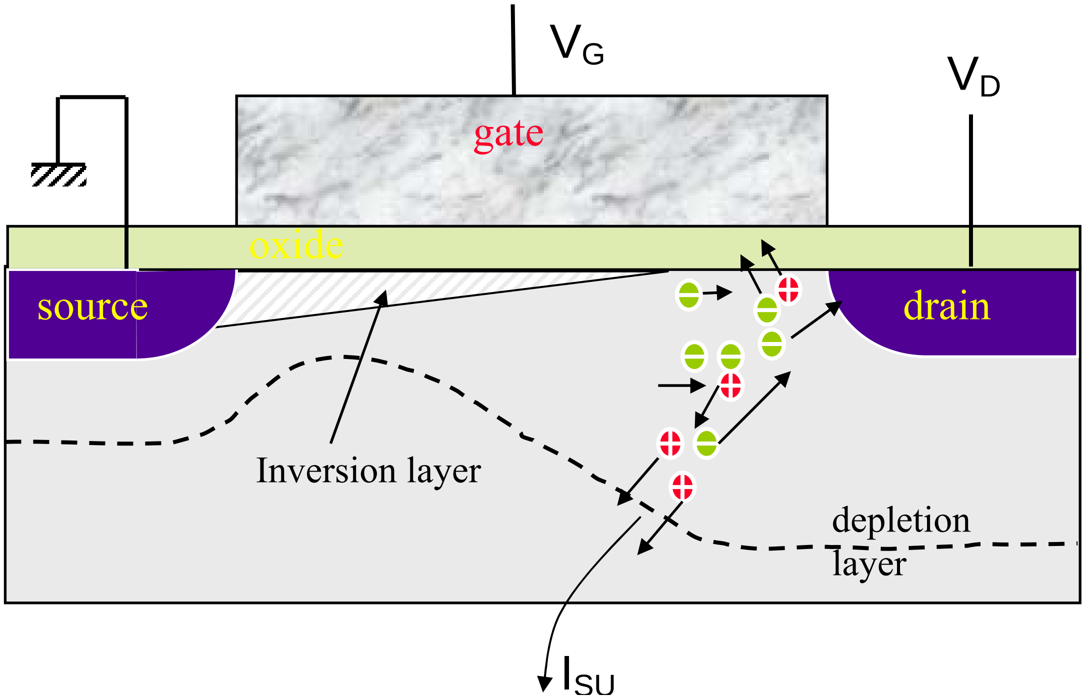

Figure 7. An n-channel MOSFET in saturation.

This diagram shows the MOSFET is in a saturation-biased condition ($V_{DS} > V_{GS} - V_T$) and the inversion layer is pinched off by high V_{DS}. The channel electrons can be accelerated to gain high energy. Some of the hot electrons will cause impact ionization to generate more electron-hole pairs. Most of the holes will go through the substrate and form I_{SUB}. A popular hot electron lifetime model is given as follows:

$$\tau \propto \{ I_{SUB}{}^{-m}/I_{DS}{}^{(1-m)}\}.$$

For an n-channel MOSFET, when $V_{GS} \sim ½\, V_{DS}$, it will generate the highest $I_{SUB,}$ thus maximum degradation will occur.

A p-channel MOSFET behaves differently. Because the energy barrier for holes to enter the oxide is much higher than that for electrons, the hot-carrier degradation for the p-channel MOSFET is less of an issue. It was reported that the maximum p-channel degradation is at low V_{GS} and the lifetime is a function of I_G rather than I_{DS}.

$$\tau \propto I_{GS}{}^{-n}.$$

The hot-carrier lifetime can be improved by a device structure change, such as using a lightly doped drain (LDD) near the drain edge to reduce the electric field. Different gate-oxide formations can also impact the lifetime. Any new technology or related process change needs to re-characterize the lifetime for hot carriers.

Hot-carrier degradation has negative activation energy. At low temperature, the crystal-lattice vibration amplitude is smaller; therefore, the hot carrier has a higher probability of traveling to the gate-oxide and silicon interface with energy higher than the barrier height.

For most power MOSFETs, the device structures are quite different than the regular MOSFET in a CMOS digital application. The drain is always in the lightly doped region and the electrical field at the drain edge is not high; therefore, hot carrier-degradation is not an issue for power MOSFETs.

5.2 Electromigration

This phenomenon also becomes more pronounced during the device scale-down, where the line width reduces and current increases. At such high current density, metallization such as the aluminum interconnect will develop voids and can develop protrusions on the sides of the lines as well. Both will cause circuit functional failures (**Figure 8**).

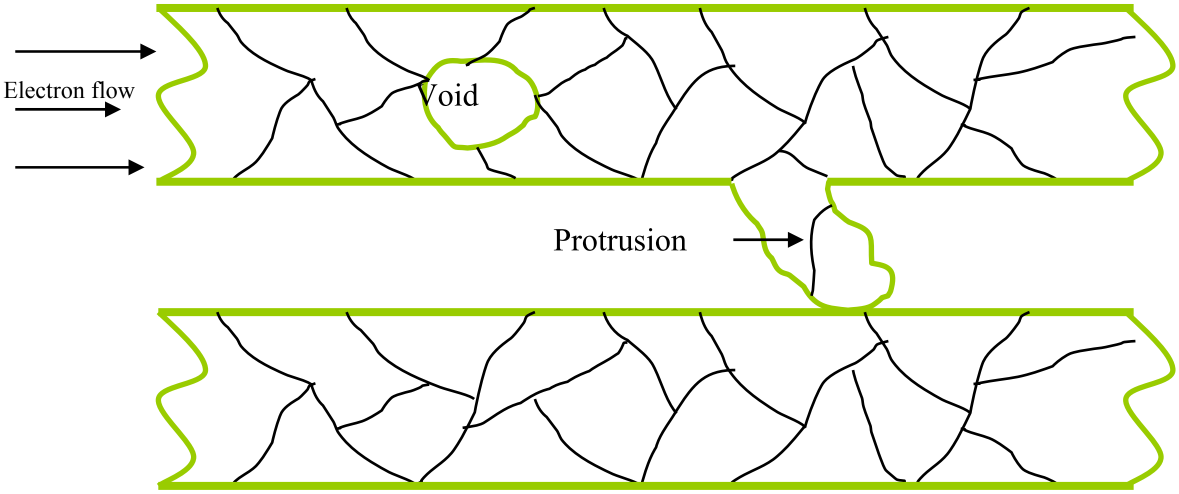

Figure 8. Voids and protrusions will cause circuit functional failures.

The atoms inside a metal line can be impacted by a high electron flow caused by high current density. The atoms can be moved more easily along the grain boundary. A triple-point intersection, with one line going in the electron flow direction and two lines going out, will have the chance to develop voids. The displaced atoms can also move outside the original line boundary to form a protrusion.

In the past, the most popular form of metallization was aluminum with 1 percent doped silicon. Nowadays, an additional 0.5 percent copper dopant is added to increase the metal's resistance to eletromigration.

Any time a new metallization system is introduced, it is necessary to fully characterize the system's electromigration capability using a high-stress method. Often, high-temperature and high-current stresses are used simultaneously. The popular Black's equation is used to calculate the lifetime:

$$\tau \propto J^{-n} \text{ x } \exp\{ -E_a/kT\}.$$

Once the lifetime is characterized, the related current density design can be implemented. As long as the layout design obeys the rules, actual failure in the field due to electromigration is not likely to occur.

5.3 Time-Dependent Dielectric Breakdown (TDDB)

The highest electric field encountered in a MOS semiconductor is in the gate-oxide region. Nowadays, the gate oxide is becoming thinner and thinner; therefore, the electrical field is getting higher and higher.

A good gate oxide can stand up to 10 MV/cm. However, during manufacturing, a certain defect density will be unavoidably introduced that will degrade the TDDB lifetime. With a high electric field, hole injection can occur on the anode side that will generate traps that gradually degrade the oxide for TDDB. Many different models have been reported, but the two most popular models are called E model and 1/E model:

$$\tau \propto \exp(-\beta \text{ x } E)$$

$$\tau \propto \exp(\gamma/E).$$

For any new process development and process change, if the gate-oxide integrity may change, the full TDDB characterization has to be done to ensure the reliability of the MOS device.

Usually this is done by using test structures and doing the accelerated test by using a high electric field and high temperature. However, the formal TDDB tests still take a long time. The voltage ramping, often called VRAMP, can be used to do a quick assessment between different process conditions. The VRAMP method uses a small voltage step to stress the oxide for a short time at each step until the oxide breakdown.

6 Product Stress Tests for Qualification

Although the test chip has been used extensively during technology development, for final qualification a real product must be selected as the vehicle for reliability stress tests. This applies to final qualification of a new wafer process technology, a new package or a new component.

It is impossible to do all the stress tests on every new product. Therefore, the so-called generic family and extension rule need to be defined, to help choose the proper stress test to ensure reliability.

There are two different categories of product stress tests for commercial plastic-package products: high-temperature life tests and environmental tests.

6.1 High-Temperature Life Tests

These tests are used to evaluate the overall semiconductor die reliability against temperature stress. Quite often, after analysis of the induced failure, some design and manufacturing problem is unveiled. The failure rate FIT value can be estimated by using the data from these tests.

6.1.1 High-Temperature Gate Bias (HTGB)

The high-temperature gate bias (HTGB) test for the MOSFET tests its gate-oxide integrity and ionic-contamination level. It is done by applying the maximum gate voltage allowed by the data sheet while stressing the device at a 150°C ambient up to 1000 hr.

6.1.2 High-Temperature Reverse Bias (HTRB)

In the high-temperature reverse bias (HTRB) test for the MOSFET, eighty percent of the BV_{DSS} allowed by the data sheet is applied to the body diode with the gate grounded to the source while stressing the device at a 150°C ambient up to 1000 hr. The combined electrical and thermal stress can test the junction integrity of the body diode, crystal defects and ionic-contamination level.

6.1.3 High-Temperature Operating Life (HTOL)

The high-temperature operating life (HTOL) test for an IC device can be performed with the chip under either static bias or dynamic bias. This term is often mistakenly called burn-in. Actually, burn-in is a term for production screening of products with a high infant-mortality rate. The stress condition may be the same as HTOL, but usually the time is much shorter for burn-in. HTOL is performed at a T_J in the range of 125°C to 150°C up to 1000 hr. The power consumption for each device needs to be calculated to adjust the oven temperature, so the junction temperature does not exceed this range.

6.2 Environmental Tests

These tests usually are used to evaluate package integrity. However, sometimes the interaction between die and package reliability cannot be separated (for example, the mobile ion migration to the device surface due to imperfect passivation).

6.2.1 1.621 Precondition (Accounting for the Popcorn Effect)

Because a surface-mount plastic device is non-hermetic, moisture will penetrate into the package. During surface-mount operations, such as the solder reflow process, the temperature excursion can generate very high steam pressure to break the package. This is called the popcorn effect.

In order to simulate this effect before any environmental test, precondition needs to be performed on plastic package prior to performing those tests. The standard procedures for precondition involve bake out, control moisture absorption and solder reflowing. For detailed procedures, refer to the JEDEC standard (IPC/JEDEC J-STD-020).

6.2.2 High-Temperature Storage (HTS)

High-temperature storage (HTS) usually is done at a 150°C ambient without bias. This test can unveil the contamination introduced during the package assembly process, which may cause the corrosion accelerated by high temperature. It can also accelerate intermetallic growth in the wire bond, which may weaken the ball bond in the assembly.

6.2.3 Temperature Cycling (TC)

Temperature cycling (TC) refers to cycling the device between extreme high and low ambient temperatures. Usually this is done between -65°C and 150°C for power devices. This one is primarily to test package endurance since the package brings together many materials with different thermal expansion coefficients. Temperature cycling can also reveal interface adhesion integrity and material cracking resistance.

6.2.4 Temperature Humidity Bias (THB)

Historically, the temperature humidity bias (THB) test is conducted at an ambient of 85°C and 85 percent relative humidity. The test is also called 85/85. It tests the moisture penetration resistance of a plastic package. Because it is also biased, it can accelerate the electrolytic-induced corrosion, if the contamination level of the molding compound is too high. The biasing rule is to bias adjacent pins with different potential. However, if the biasing scheme causes the device to draw high current and high heat, then the scheme needs to be modified. The reason for this is the high heat will prevent moisture penetration, which may defeat the purpose of the test. The test is usually done for 1000 hr.

6.2.5 Highly Accelerated Stress Test (HAST)

Basically, the highly accelerated stress test (HAST) is another biased temperature humidity test at higher temperature than THB. Usually, HAST is done at an ambient of 130°C while the humidity remains at 85 percent. Since the acceleration is very high, usually only 100 hr is needed to do this test. Therefore, it saves a great deal of time for qualification. However, not all the failures are valid. It’s possible for this test to create a highly accelerated situation failure that will never occur in a real-life application. For example, some of the package molding compound can decompose under HAST conditions.

6.2.6 Pressure-Cooker Test (PCT)

The pressure-cooker test (PCT), which can also be called Autoclave, is performed at 121°C and 2-atm pressure (1 atm above the atmosphere pressure). Under these conditions, the relative humidity is 100 percent. Quite often, condensation occurs inside the chamber. Same as HAST, one needs to be very careful in validating the failure data. Some of the failure mechanisms will not occur in the real world. Some semiconductor companies have already stopped using this type of saturated moisture test. Unfortunately, some of the standards still call for this test. The main purpose is to test the package moisture resistance and corrosion resistance induced by a high concentration of certain contaminations. The more severe corrosion can be considered as valid failure, but some minor leakage after bake or decap can be considered invalid failure. Most of the standards use 96 hr for this test.

6.2.7 Intermittent Operational Life (IOL)

The intermittent operational life test is sometimes called the power cycle test. Usually automotive applications require this test. Unlike the temperature cycle, the heat is generated from the device within itself, rather than from the ambient heat. The device needs to be heated up by applying power to it. The temperature difference is set for greater than 100°C. The test system will apply power to each device automatically. Once the temperature reaches the high point, the fan will automatically turn on for rapid cooling. This will speed up the process of power cycling. The total number of cycles required is defined in the spec from the Automotive Electronic Council (AEC-Q101). (For example, the T0-220 package qualification needs 8572 cycles for ΔT_J = 100°C. For higher acceleration at ΔT_J = 125°C, only 4286 cycles are required.).