DESIGN FOR MICROELECTRONICS RELIABILITY

Product reliability is important because it affects the resources of manufacturers and customers alike. These resources may include production facilities, manpower, money, and time. Poor product reliability may increase production costs due to low yield, warranty costs, and possible business loss if units fail during service in the customer's environment. Likewise, the customer may incur resources losses. Therefore, it is necessary to understand product reliability and its detractors to minimize wasted resources and to maximize customer satisfaction and loyalty.

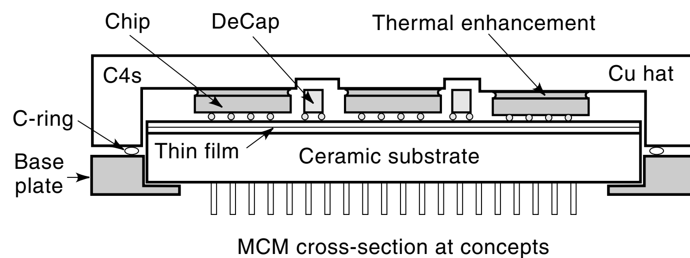
This article reviews the basics of microelectronic component reliability analysis: reliability concepts, failure distributions, accelerated testing, and reliability modeling. In particular, this article deals with the reliability methods and modeling applied to circuitry and interconnections of electronic packages (i.e., encapsulated chips attached to a substrate) and wiring to a board system. Figure 1 shows a high-end module with the various parts and components, which illustrates some of the interconnection features for a particular IBM-produced multilayer ceramic (MLC) module. This is a SCIMITAR module used in the IBM S/390 G4 system, a high-end parallel processor computer. Reliability analysis must deal with high-end and the myriad of low-end and intermediate-size multichip modules (MCM). Many factors affect a microelectronic component's robustness in terms of manufacturing yield, service reliability, and usage latitude. Following are some of the factors:

1. Machine and device architecture
   - Redundancy
   - Error correction
2. Circuit design
   - Signal margins
   - Immunity to external electric stimuli (e.g., ESD, RF noise)
3. Materials compatibility
4. Physical design tolerance
   - Susceptibility to particles, misalignment
   - Dimensional stability
5. Fabrication
   - Tool control
   - Fabrication environment
6. Reliability and quality screens
   - Inspections
   - Electrical and physical testing
   - Burn-in
7. Application
   - Environmental aspects (e.g., electrical, mechanical, acoustical)

Many of these factors may interact synergistically, may cancel each other, or they may not be effective under field conditions. For example, materials compatibility may pose little or no problem if the application service temperatures are low enough so that materials interactive rates are extremely low. Consequenty, robustness must be addressed from the viewpoints of both manufacturability and serviceability.

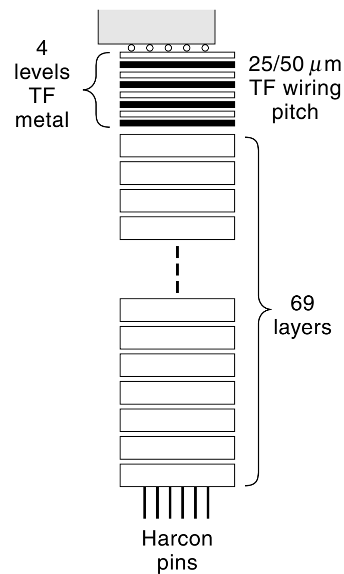
Electronic package circuitry and the solder interconnects at the die and package level can fail in terms of a number of mechanisms, such as fatigue, creep, corrosion, metal migration, electromigration, and thermomigration (1). Under normal field conditions, the solder joints and circuit wiring have a nonzero probability of failing. Failure may range from less than one part per million to several percent, depending on the stress level and the vulnerability of the metallization, which is largely a function of the process and design. It is the job of

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 1999 John Wiley & Sons, Inc.



(a)

(b)



(c)

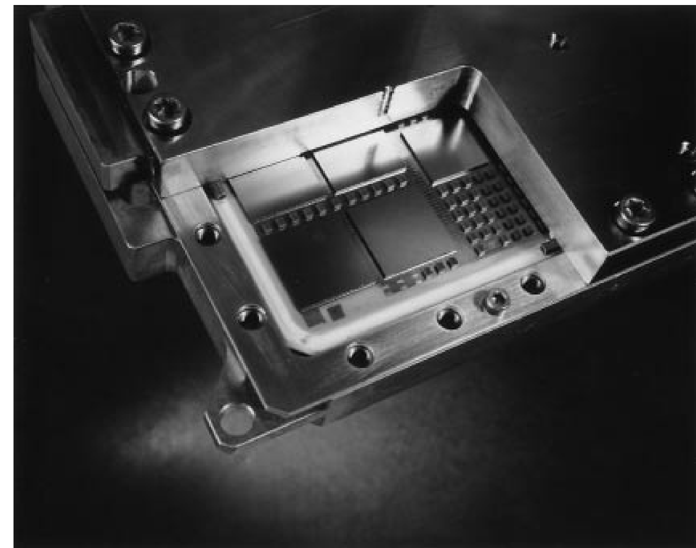

(d)

Figure 1. High-end Scimitar module with its various components.

reliability engineering to develop accelerated tests whereby one can predict the relative life of components in the field on the basis of short laboratory tests. Because microelectronic components are designed to last 100,000 h and the test times are usually limited to few hundred hours, the acceleration factor must approach one thousand, considering a safety factor. Physically, the acceleration factor represents the number of times the prevailing failure mechanism kinetics are accelerated by test condition variables, such as temperature, applied potential, impurity concentration, relative humidity, mechanical stress, and electric current density. It is key, therefore, to study and understand the mechanism which predominates and is responsible for the observed failures. An important aspect of reliability modeling is to ensure that the mechanism applies consistently across the various conditions and also to know the limits beyond which the model fails to predict the field lifetime. Also, one should not extrapolate test results to conditions under which the mechanism may not be operating or conditions outside the mechanism that prevails under test.

The scope and objective of this article is to discuss the means to predict the field cumulative failures at the end of life with a certain level of confidence, based on (1) the statistical distribution of the data and the acceleration factor as a function of stress conditions, design, and materials properties, and on (2) physical models developed from failure mecha-

nisms and verified by empirical data as a function of material properties and package design. The article teaches how to develop reliability modeling and optimize design and to achieve accurate failure predictions under field conditions with a high confidence level.

RELIABILITY CONCEPTS

Reliability here is defined as the probability that a device, component, or assembly will perform the required functions under service or operational conditions for any given period of time. The reliability of an electronic system, over the period of its expected lifetime, should be at least high enough to meet the manufacturer's commitments to the users. Using a memory component as an example, the required function could be the proper storage and output of a very specific set of word patterns for a number of years of continuous operation under certain electrical and environmental conditions. For a metal-to-metal connector contact, the required function may be that the interface resistance does not increase above some threshold value. A failure results when the unit does not perform its required function. Hence, the manufacturer must understand the user's needs and the proper tests and failure criteria to ensure the required performance. Almost all microelectronic products cease to perform satisfactorily at some point. The goal is to eliminate failures within the lifetime of the system or to reduce them to a tolerable level under operating conditions.

The first concept in our discussion of reliability is that of components' lifetimes as a continuous random variable that can be described in mathematical terms, such as the probability density function (pdf). This is a statistical way of describing the histogram of all lifetimes for a population of units. Because histograms are almost always based on samples drawn from a much larger population, they are used to compute estimates of the parent population's life characteristics. The mathematics of statistics endeavors to infer estimates of the behavior of the whole population from subpopulation behavior. Figure 2 is a histogram of failure times fitted by the Weibull probability density function $f(t)$. The product $f(t)dt$ is the fraction of all of the lifetimes contained in the time interval dt beginning at time t.

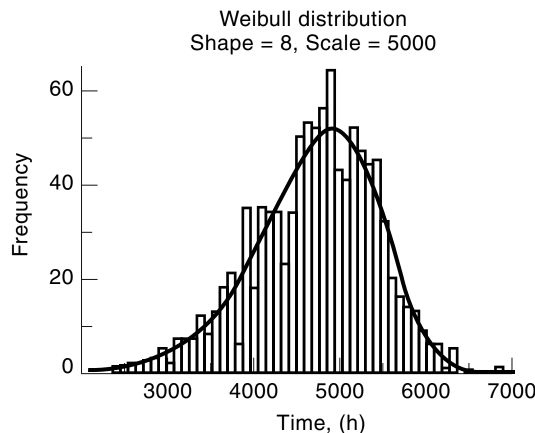


Figure 2. Failure histogram fitted with the probability density function (pdf) for the Weibull distribution. Courtesy D. Scheider.

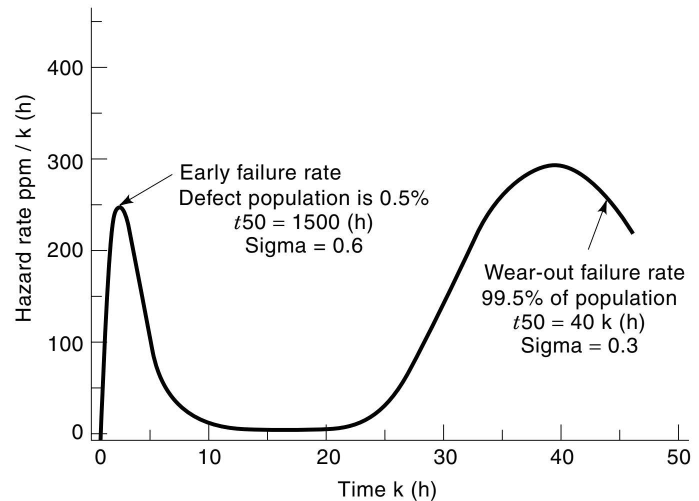


Figure 3. Bathtub curve featuring early life and wear-out mechanisms.

Most often, components have several failure mechanisms operating during service which contribute to lessening their reliability. Then it becomes necessary to describe the reliability of the component in terms of the individual mechanisms. Combining these individual mechanism lifetime distributions often yields a probability function from which the familiar bathtub curve is derived (Fig. 3). This curve illustrates the instantaneous failure rate, which is the rate of failure in the interval δt of those units that have survived up to that time. The figure illustrates the early failures (initial peak near the origin) and the wear-out failure distribution beginning at the right. The early and wear-out distributions are further illustrated in Fig. 4. The cumulative density function (cdf), $F(t)$, and the reliability function $R(t)$ are used most often to derive information about time-to-fail distributions. Many reliabilty analyses utilize continuous distributions as a function of time. If $f(t)$ is the pdf, then the cdf or $F(t)$ is given by

$$F(t) = \int_0^t f(y)\,dy \tag{1}$$

where $0 \le F(t) \le 1$.

There are two practical interpretations for the cdf of lifetime data: (1) $F(t)$ is the probability of a unit failure by time t, and (2) $F(t)$ is the fraction of all units of a population which fail by time t. The reliability function $R(t)$ is the complement of $F(t)$ and represents the probability of a unit's survival at time t: $R(t) = 1 - F(t)$. The hazard rate or instantaneous failure rate is given by $h(t) = f(t)/R(t) = f(t)/[1 - F(t)]$. If the hazard rate $h(t)$ represents the fraction of failures per hour, then when multiplied by 10^5, the units are percent failures per kilohour (%/kh), whereas multiplying by 10^9 yields ppm/kh. The fractions $F(t)$ and $R(t)$ are also usually expressed in percent. There are two rules used to develop the formulas for computing the cdf and R of a system composed of n identical components with a single failure mode and of units with multiple failure modes (2). They are (1) the multiplicative rule, which states that the probability that several independent events will all occur is the product of the individual event probabilities, and (2) the complementary rule, which states

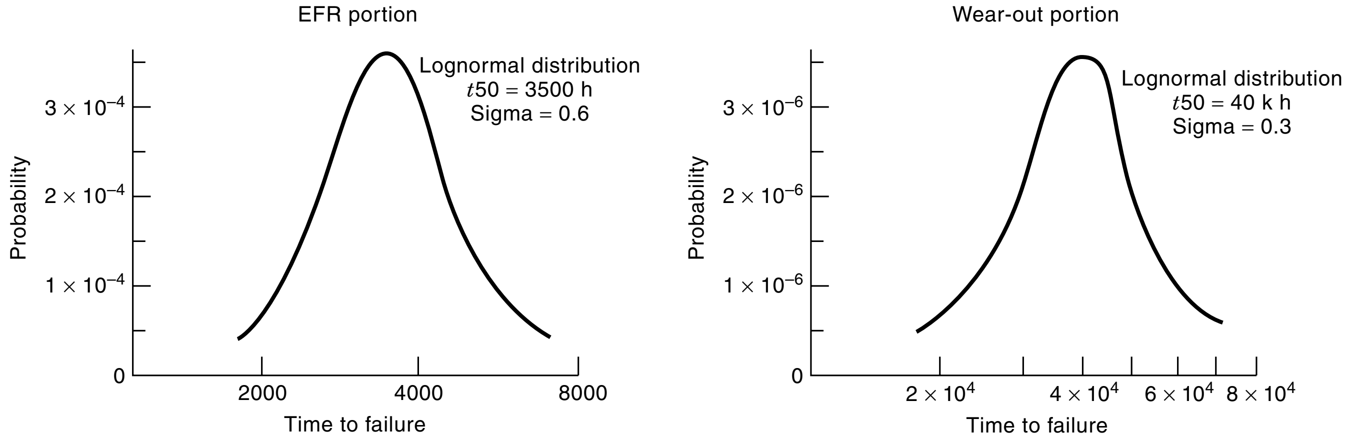


Figure 4. Probability distribution functions that result in the bathtub curve in Fig. 3.

that the probability that an event will not occur is one minus the probability that the event will occur. For n identical units, each with the same $R(t)$, $R(t)^n$ is the probability that all units survive to time t. The probability of a system failure by time t (system fails when first of n units fails) is given by $F_s(t) = [R(t)]^n$. The reliability function for a component with multiple failure modes is given by the product of the reliability function for each mode:

$$R_{\text{total}}(t) = \prod_{i=1}^{n} R_i(t) \tag{2}$$

where $R_i(t)$ is descriptive of mechanism i. The total cdf is expressed as $F_{\text{total}}(t) = 1 - R_{\text{total}}(t)$.

TYPES OF FAILURE DISTRIBUTIONS

There are a number of failure distributions used in the reliability of interconnections. One can read Tobias and Trindade (2), Hahn and Shapiro (3), Hogg and Craig (4), and Montgomery (5) to obtain broader knowledge of their properties and their specific use. Tobias and Trindade point out some theoretical reasons for using one over the other. The lognormal is indicated when final failure results from a multiplicative degradation process, such as crack propagation, fatigue, or charge injection. The Weibull finds favor when a component has a multitude of flaws, each competing to be the first failure site: it is a smallest extreme value type distribution. Often the microphysics and chemistry of failure mechanisms is only partially understood, and a choice on theoretical grounds is subjective. Then analysts use the distribution that best fits the data within the experimental time range or the one with which they have the most familiarity. If only a small fraction of components' lifetimes are observed, the choice is even more difficult.

Again, when dealing with failure mechanisms in general, the most suitable and applied distribution is the lognormal distribution, though in some applications the 3-parameter Weibull fits the data better because of an added degree of freedom, which allows one to tailor the fit with more precision. There is also the normal distribution, which in general finds its use in nonmechanistic applications. When dealing with failure mechanisms, it rarely fits the empirical failure distribution. If t is the random variable, the normal distribution probability density for t is given by

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{t-t_{\rm m}}{\sigma}\right)^2} \tag{3}$$

The distribution is defined over the range $-\infty$ to $+\infty$, and generally it is not employed for reliability failure projections as already indicated, but rather it forms the basis for the more often used lognormal distribution. The characteristic parameters are the mean, $t_{\rm m}$, and the standard deviation σ, whose square is called the variance σ^2. The cumulative density function (cdf) is the probability that a normal variable t is equal to or less than a given value b,

$$P(t \le b) = \int_{-\infty}^{b} \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{t-t_{\rm m}}{\sigma}\right)^2} dt \tag{4}$$

$$z = \frac{t - t_{\rm m}}{\sigma} \tag{5}$$

Because Eq. (4) cannot be integrated in a closed form, we change variables, which transforms the normal distribution into the unit normal distribution, such that the probability is expressed in terms of the number of sigmas independently of the mean and the value of σ:

$$P(t \le b) = \Phi\left(\frac{t - t_{\rm m}}{\sigma}\right) \tag{6}$$

which has a mean equal to zero and a sigma equal to one. The probability density function thus becomes

$$\phi(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{7}$$

$$\Phi(z) = \int_{-\infty}^{z} \phi(z)\, dz \tag{8}$$

which can be evaluated using tabulated $\Phi(z)$ values for given z-values. This is called standardization, converting a normal $N(t_{\rm m}, \sigma)$ to a unit normal distribution $N(0, 1)$. In a lognormal

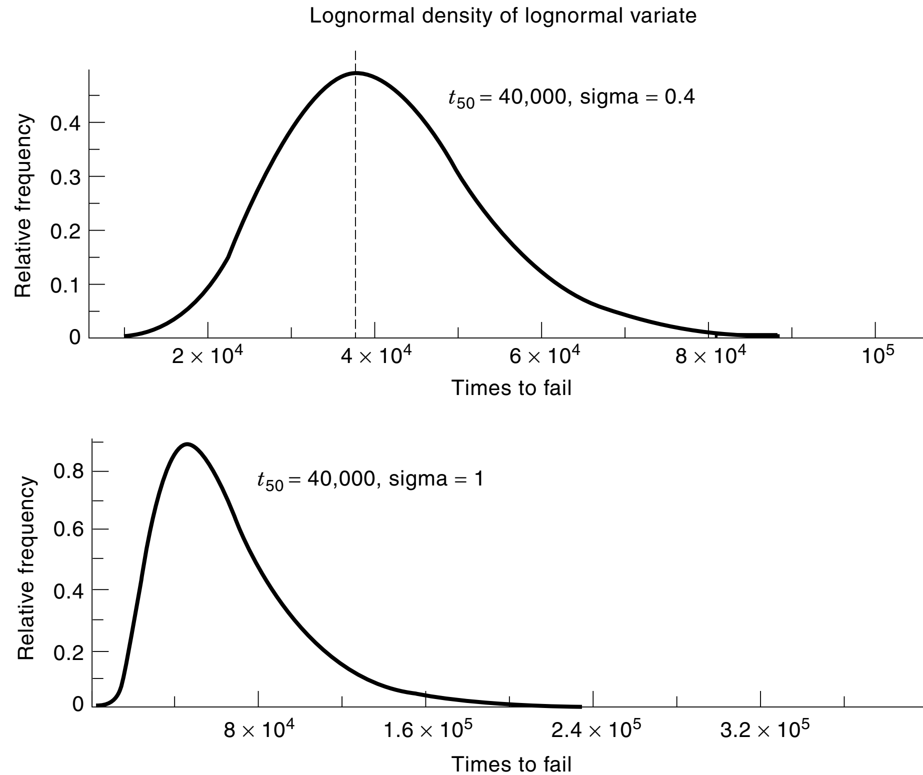


Figure 5. Lognormal pdf with two sigma values and constant t_{50} plotted vs. time.

distribution, the random variables have their natural logarithms normally distributed when they fit this type of distribution. The probability density function $f(t)$ is expressed as

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} e \frac{-(\ln t - \ln t_{50})^2}{2\sigma^2} \tag{9}$$

and the cumulative density function $F(t)$ is given by

$$F(t) = \Phi\left[\frac{\ln(t/t_{50})}{\sigma}\right] \tag{10}$$

where Φ refers to the unit normal distribution. In practice, the empirical cdf can be plotted on normal probability paper against the logarithm of time in order to estimate the median time to failure t_{50}, and the slope σ:

$$\ln t = \ln t_{50} + \sigma\Phi^{-1}[F(t)] \tag{11}$$

Next, the Weibull distribution is written as

$$f(t) = \frac{\beta}{C}\left(\frac{t-\mu}{C}\right)^{\beta-1} \exp\left[-\left(\frac{t-\mu}{C}\right)^{\beta}\right] \tag{12}$$

where

$t \geq \mu$
$-\infty < \mu < +\infty$, location parameter
$C > 0$, scale parameter
$\beta > 0$, shape parameter

The Weibull distribution has great flexibility and assumes a variety of shapes by selecting different values of the parameters μ, C, and β. The cumulative Weibull distribution for $\mu = 0$ is given by

$$F(t) = 1 - \exp\left[-\left(\frac{t}{C}\right)^{\beta}\right] \tag{13}$$

Figures 5 and 6 illustrate how the lognormal probability density function is affected by varying σ at a constant t_{50}. As σ increases, the left tail of the distribution spreads to earlier times whereas the right tail spreads to longer lifetimes. Therefore, one likes to keep σ small in the reliability field. Figure 5 is the lognormal density of lognormal variate, a linear time plot illustration of the pdf with two sigma (σ) values at constant t_{50}, and Fig. 6 is the normal density of the lognormal variate, illustrating a logarithm time plot. The latter shows the symmetry of the normal distribution which attests to the fact that the data are lognormally distributed over time. When such data are plotted on a linear time-scale as in Fig. 5, the distribution is characterized by a fast initial rise and a long tail to the right after the maximum, which is the manner in which failures occur most of the time.

FAILURE MECHANISM ACCELERATION

Under normal circumstances, one cannot test a component over its required lifetime to determine its reliability before deciding on its acceptability because the lifetime is too long (perhaps many years). Thus one resorts to accelerated-life stress tests to obtain the failure distribution much sooner. The idea is to increase the stress on a component to a level greater than in normal application to shorten the failure time. This leads to the concept of an acceleration factor, a

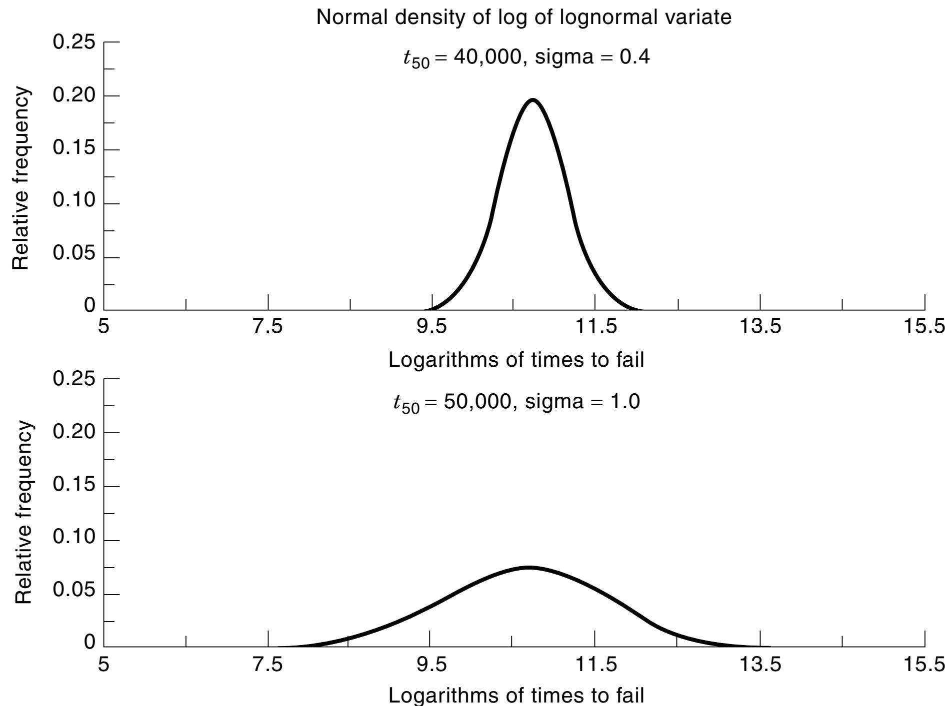


Figure 6. Lognormal pdf with two sigma values and constant t_{50} plotted vs. log time.

mathematical way of projecting distribution parameters developed at high stress levels to lower levels, which includes the application stress. The benefit is that one may decide about components' reliabilities in very short times, perhaps weeks, facilitating near real-time manufacturing, design, and marketing decisions regarding the components. But there is a price to pay: modeling, which allows one to interpret the results from the short tests in terms of field lifetime. Modeling entails a wide range of experiments to study how components behave and fail under a variety of stress conditions and which failure mode responds to which stresses. The discipline for this includes experimental design, statistics, failure analysis (including physical, chemical, mechanical, and metallurgical), and test methodology. Many stress variables are used to induce mechanism acceleration. Some of these are temperature; relative humidity (including steam and high pressure); voltage; current density; temperature and power cycling; frequency; amplitude; dwell; ramp rate; mechanical strain and stress (including vibration and shock); corrosive gas concentration; and radiation. The acceleration factor (AF) is often thought of as the linear relationship AF = t_{use}/t_{stress}, which are the times or cycles to failure under use and accelerated stress conditions for a given fraction or percentile of population failures. Figure 7 illustrates the acceleration factor with two lognormal cdf plots obtained under different stresses. The interconnect lifetime is predicted on the basis of the failure distribution and the acceleration factor. The latter is computed from the model and is the ratio of the field lifetime to the test lifetime, assuming that the operating mechanism remains unchanged under field and test conditions. If the mechanism and sigma remain unaltered, the test failure distribution can be converted to field-equivalent time in its entirety by multiplying the test time to failure by a single acceleration factor. This enables one to determine the cumulative percent failure for any given time or number of cycles under field conditions, as shown in Fig. 7. Because the two failure distributions are the result of two different temperatures, the acceleration is used to determine the activation energy of the mechanism. However, when the sigmas of the test and field failure distributions differ, there are virtually an infinite number of acceleration factors from one extremity of the distribution to the other, as shown in Fig. 8. In such a case, a common sigma is forced on the data with a loss of confidence level in the statistical parameters. The constancy of the sigma should be ascertained by testing the vehicles under different conditions. To determine a reliable sigma and acceleration factor, the test data must cover failures beyond 50% of the population. There is a wear-out and an early-failure distribu-

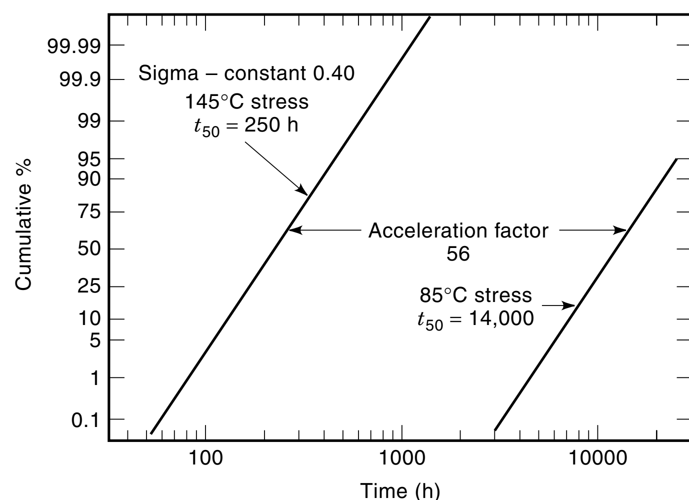


Figure 7. Illustration of acceleration factor between lognormal plots with the same sigma.

tion. Wear out is the degradation process inherent in the properties of the materials and causes the component to fail in a characteristically statistical manner often representable by a lognormal or Weibull distribution. Wear out is a property of the materials being tested unrelated to defects. Variation in the wear-out time to failure from component to component is ascribed to the heterogeneity and anisotropy that exist in the metallurgical structure prevailing in the circuitry and interconnects after the joining process. Heterogeneity and anisotropy affect every failure mechanism and are mostly responsible for the wear-out distribution. The early-failure distribution, if any, is due to defects, which cannot endure the applied stress for the component's lifetime. Failures are characterized by failure criteria, whatever the mechanism, in terms of electric resistance increase, e.g., 30 mΩ for a four-point resistance measurement, which consists of measuring the voltage drop across the solder joint upon sending a low current through the bump. The resistance change is obtained by dividing the voltage drop by the electric current and subtracting the initial resistance. The measurement process is repeated at intervals until the resistance change equals or exceeds the failure criterion. When enough components fail (i.e. $\geq 50\%$), one obtains a failure distribution as a function of time or cycles. Generally, fatigue failures follow a lognormal distribution. However, at times they fit the Weibull distribution. This is true for other failure mechanisms, which are likewise studied through electric resistance measurement. In the case of metal migration or current leakage, one determines the insulation resistance of the path, for which the failure criterion is commonly 10 MΩ, by measuring the voltage drop across a known large resistor in series with the path. This is the only measurement needed to calculate the insulation resistance of the gap, knowing the applied voltage. The failure distribution parameters are determined by distributions on paper or by computer program, which yields bounds for a given confidence level.

It must be emphasized that the failure mechanism must remain unchanged over the test and application stress range, otherwise the acceleration factor is incorrect. To avoid such a problem, one must stress devices only within a verified window which is determined empirically on the basis of failure data consistency within the range of test and application conditions, so that one can safely make field projections within those boundaries. But sometimes the window cannot be verified for the application conditions. In such a case the application/functionality window stands outside the experimentally verified window, which forces one to extrapolate the results to uncharted territory, assuming that the model, which was derived on the basis of higher stresses, is predictive with reasonable accuracy. Figure 9 illustrates a verified window for any failure mechanism in a plot having generalized coordinates. An application/functionality window is shown inside the verified window and another partly outside, indicating that no verification was made for the external portion. This involves some risk in projecting the data to the application stress. The window inside is well within empirical knowledge, which is where any application, functionality, and specification range should fall for minimal risk. Briefly, for a particular failure mechanism, for which lifetimes are characterized by lognormal distributions, one generally considers the valid stress range as that in which the physical failure mode is the same throughout the range and the lifetime distribution shape parameters are essentially invariant.

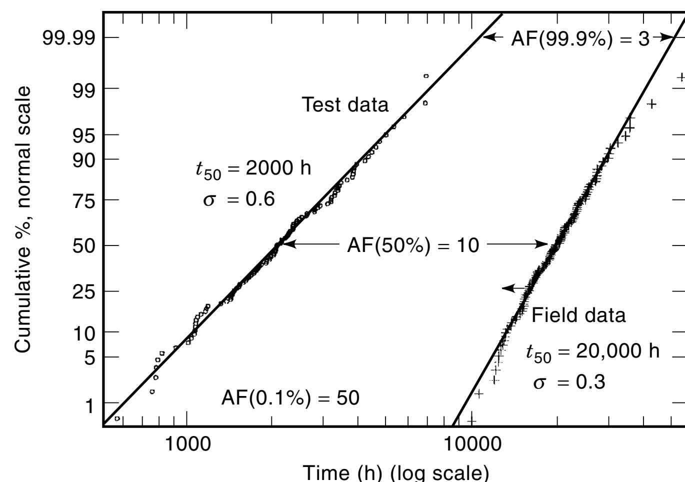


Figure 8. Illustration of dissimilar sigmas between test and field distributions.

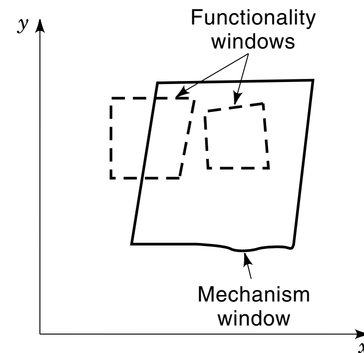


Figure 9. Illustration of verified mechanism window and functionality windows.

DEFECTS AND BURN-IN

Defects are physical anomalies which affect the failure mechanism and therefore the time to failure of a component or device. Some defects may serve as nuclei for fatigue crack initiation. Others may be the trigger sites for electromigration or metal migration. On the other hand, a defect may be cause for delamination at the joint interface, which becomes conducive to water condensation and ultimately the site of electrochemical or galvanic corrosion. Inclusions, abrupt variation in grain size, and other structural differences can also influence electromigration and fatigue failure. Defects are generally detectable by low-power microscopes and physically consist of missing material, obvious contamination, residues, inclusions, cracks, metallic stains, particles between joints, which are mostly generated during manufacture and can be screened out by close inspection. A defective population has a separate failure distribution from that of the wear out and its own statistical parameters. The final recourse is the use of burn-in, which imposes stresses for a short period much more

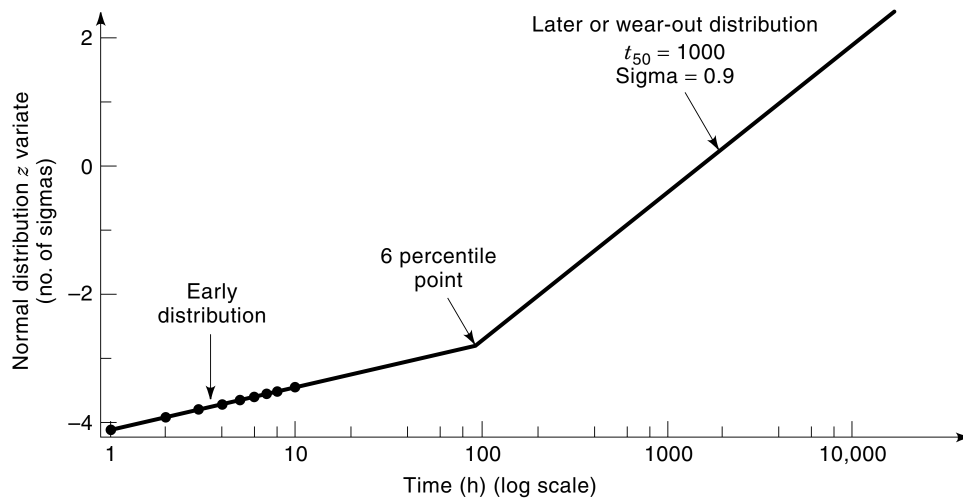


Figure 10. Lognormal cdf plots illustrating competing failure mechanisms.

severe than those experienced in the field and which weed out most of the defectives. But the stresses can also shorten the useful service life of a healthy component if there is an insufficient safety factor for the wear-out process. There are, however, cosmetic defects which are irrelevant to the operating failure mechanism, that is, chemically inert stains. Defects cause a failure distribution to become bimodal. The bimodality is characterized by a break in the failure distribution slope distinguishing early-failure and wear-out, each with its own sigma. The early failures extend and broaden the frontal tail of the distribution resulting in a larger sigma. When characterized by failure analysis, the early failures can be separated from the wear-out failures, and both can be treated separately as two distinct populations. Early failure is caused by a defect that cannot withstand the overwhelming effect of any of the failure mechanisms acting on the defect for the lifetime.

Early and wear-out failure distributions, illustrated together in Fig. 10, are often encountered when there is a considerable number of defects. The lognormal plots illustrate competing failure modes, one predominating for early times, and the other for later times. The first part of the curve has a large sigma and represents the early-failures distribution, whereas the second part has a smaller sigma and represents the wear-out distribution. In practice, if test measurements are not taken early enough, the early-failure distribution can be missed, and the failures are included with the wear-out distribution, thus missing the opportunity to discover a defect problem usually traceable to manufacturing.

For competing failure modes (j), where each mode would result in 100% failure if operating alone and given enough time, the cumulative distribution for the units is given by

$$F_{\mathrm{T}}(t) = 1 - \prod_{j=1}^{k}[1 - F_j(t)] \tag{14}$$

Such a composite model is often called a competing risk model. When a single mechanism is operating instead and only a fraction of the sample population is affected or can ever fail, the failure fraction in a lognormal cdf plot "saturates" or approaches some value asymtotically. An analysis of this type of data involves treating only the failure-prone units as the sample. Once the distribution parameters are estimated, and the cdf is computed for the defective population, one can describe the whole population through the relationship

$$\mathrm{cdf}(t)_{\mathrm{WP}} = p - \mathrm{cdf}(t)_{\mathrm{DP}} \tag{15}$$

where

WP = whole population
DP = defect population
p = fraction of defects

Figure 11 is an example of a lognormal plot of a "defect" population illustrating the saturating behavior. In this cdf plot, we can estimate the cdf value p, which the curve approaches asymptotically. If N is the test sample size, then pN is the approximate defect fraction. If the defect population is lognormally distributed, then redoing the cdf plot based on sample size pN produces a straight line on lognormal probability paper, thereby facilitating estimations of $t_{50\mathrm{defects}}$ and

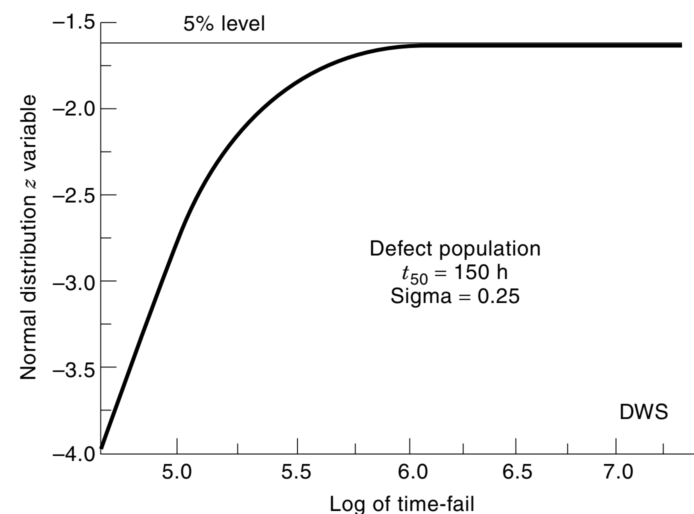


Figure 11. Cumulative failure distribution of a limited-defect population.

σ_{defects}. In the general case, when there are j mechanisms operating and each one affects only a small part of the population, the total pdf is given by

$$f_T(t) = \sum_{j=1}^{k} a_j f_j(t) \quad (16)$$

where

$f_j(t)$ = pdf of the jth mode
a_j = population fraction affected by mode j

An analogous formula describes $F_T(t)$. Then the hazard rate for the population is computed from $f_T(t)$ and $F_T(t)$. Such situations are described as mixed-mode models.

THERMOMECHANICAL MODELS

Fatigue

Thermomechanical fatigue is the predominant failure mechanism of first- and second-level interconnects. The first-level interconnects join a chip to a substrate (i.e., C4), while the second-level interconnects join a substrate to a card (i.e., BGA). Failure results from thermal cycling due to the thermal mismatch between chip and substrate, which causes shear strains in the solder joints of an amplitude proportional to the temperature-delta and the joint distance from the neutral point (DNP), which has a zero strain.

Thermal-Cycle Stress. Thermal cycling takes place in chambers programmed to take the test components through a prescribed thermal cycle, e.g. 0°C to 100°C which induces shear strain. The fatigue life of a joint is inversely proportional to the shear strain raised to the second power (Coffin–Manson), to the cycle frequency to the one-third power, and directly proportional to the thermal activation term, which is a function of the peak temperature. The dependence of N_{50} on frequency addresses the stress relaxation and creep at the dwell temperature and along the temperature profile. The degree of relaxation increases with cycle duration or dwell time, which in turn vary inversely with the frequency. The mechanism is controlled by grain boundary diffusion creep and plays an important role in solder fatigue. For high-lead Pb–Sn solder, stresses under field conditions relax to a threshold level in about one hour. Below the threshold stress level, no further fatigue damage occurs in the joint for that cycle, though the temperature cycle period may last many hours or days. Below the threshold, grain boundary diffusion creep and the fatigue process come practically to a halt, mainly because dislocation sources are not regenerated at such low stress. Solder bump fatigue is caused either by mechanical or thermomechanical cycling. In a mechanical cycle, the cyclic stress is applied isothermally at constant amplitude, which need not be symmetrical about zero stress. One can bias the stress so that the tensile and compressive stress amplitudes are asymmetrical, and one of them could be zero. This is not representative of the thermal fatigue cycles experienced by solder interconnects in electronic packages where the stress and strain result from thermal mismatch between chip and substrate. The fatigue results are different because stress relaxation and associated creep in the thermal cycle occur continuously over the temperature excursion. The thermal effects on fatigue dynamics are absent when the joints are mechanically cycled. Thermal effects cannot be quantitatively interpreted without an accurate knowledge of how strain and temperature interact. The temperature ramp rate is also important because the stress profile depends on this rate and so does the duration of the stress cycle above a threshold which dictates the dislocation dynamics and fatigue damage.

From a mechanistic aspect, the process of failure involves crack nucleation and propagation, which are inherently statistical in nature. Through the years electron transmission microscopy has shown complex three dimensional dislocation tangles which develop during cyclic stressing. It appears that cracks form and grow in each cycle by being fed dislocations that move in from the surrounding lattice. Thus, the intersection of moving dislocations is a principal cause of the formation of vacancies and fatigue damage. We know from the striations on the fracture surface that crack propagation takes place in each cycle. Forsyth (6) has advanced a mechanism for the formation of striations, which involves first, the extension of the crack front by simple slip on the primary glide plane and second, further extension by cross-slip as the stress increases. Oxygen aids crack propagation kinetics, which is believed to be caused by grain-boundary oxidation and the lowering of the surface energy for crack propagation.

The present "modified" fatigue model, introduced by Norris and Landzberg (7), accounts for the anelastic behavior of solder through a frequency and dwell temperature factor. But serious fatigue modeling goes back much further, when many attempts were made to fit S–N curves with a simple equation. It remained for Coffin and Manson to show that many polycrystalline materials, fatigued at constant plastic strain amplitude, obey the relationship, $C = \sqrt{N_f} \cdot \Delta\epsilon_p$, where N_f is the number of cycles to failure, $\Delta\epsilon_p$ the plastic strain range, and C is a constant of the material. Apparently the reason so many materials obey an equation like Coffin's is because of the fundamental nature of the mechanism of crack initiation and propagation. Therefore the "modified" fatigue model is an extension of the Coffin-Manson relationship that also enables one to relate fatigue to frequency and temperature, which are important variables in the solder fatigue encountered in electronic packages. In a thermal-cycle regimen, the number of cycles to failure N_f depends principally on the magnitude of the shear strain γ raised to the negative second power; the frequency f to the one-third power; and the maximum cycle or dwell temperature T_m. The effect of dwell at the off-temperature has much less importance and therefore the off-temperature is ignored as a variable. The three variables are the basis of the modified fatigue model:

$$N_f = A\gamma^{-2} f^{1/3} e^{1440/T_m} \quad (17)$$

where A is an experimental constant of the material and process.

Stress Threshold. The stress threshold is calculated on the basis of the threshold stress τ_{TH} temperature profile time con-

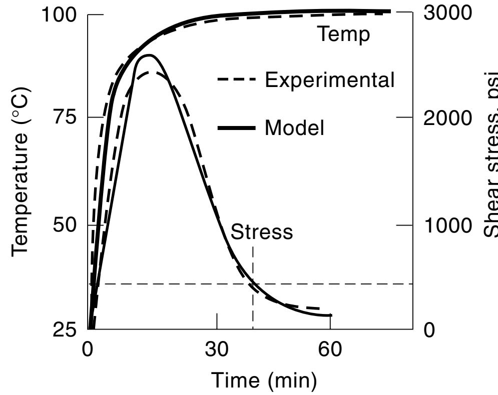


Figure 12. Stress and temperature profiles for solder bumps; model vs. test (dotted).

stant β^{-1}, and the maximum cycle temperature (Fig. 12). For high-lead Pb–Sn solders, it varies from 6 cycles per day (cpd) at ambient temperature to 36 cpd at 100°C. This means that after 40 min at 100°C, the stress is fully relaxed with respect to fatigue, and no additional fatigue damage occurs. Therefore, at 100°C, one must use 36 cpd instead of the acual field frequency, which is much lower. The frequency threshold expression, derived from Maxwell's viscoelastic element can be written as (8)

$$\tau_{\mathrm{th}} = G(T)\gamma(1 - e^{-\beta t_{\mathrm{th}}})\ln(2e^{\beta t_{\mathrm{L}}} - 1)\exp(-t_{\mathrm{th}}/t_{\mathrm{C}}) \tag{18}$$

where

$$t_{\mathrm{c}} = t_{\mathrm{cc}}\exp\left[-\frac{\Delta H}{K}\left(\frac{1}{T_0} - \frac{1}{T}\right)\right] \tag{19}$$

$T = T_{\mathrm{m}} - \Delta T\exp(-\beta t_{\mathrm{th}})$
$t_{\mathrm{th}} = 1/f_{\mathrm{th}}$, reciprocal frequency threshold
$G(T)$ = Solder shear modulus parameter
γ = Solder shear strain
β = Temperature reciprocal time constant
t_{co} = Solder stress relaxation time constant at T_0
T_0 = Reference temperature, i.e., ambient
T_{m} = Peak temperature
$\Delta T = T_{\mathrm{m}} - T_0$
ΔH = Activation energy for GB diffusion

In high-strain fatigue cycles much beyond the yield point, the product $G(T)\gamma$ becomes the flow stress, which remains reasonably constant. For a 0° to 100°C cycle, its value is 18.6 MPa. From Eq. (18), one determines f_{TH} based on known parameters, β, $G(T)\gamma$, and the relaxation constant t_{c}, as a function of temperature. Equation (18) predicts that for a 0° to 80°C temperature profile with a common β, the threshold frequency for high-lead Pb–Sn solder is 20 cpd. To calculate the fatigue life ratio between the field at 1 cpd and the test at 72 cpd, use the frequency ratio according to the model, and replace the field frequency by the threshold frequency:

$$\frac{N_{\mathrm{T}}}{N_{\mathrm{F}}} = \left(\frac{f_{\mathrm{T}}}{f_{\mathrm{F}}}\right)^{1/3} = \left(\frac{72}{20}\right)^{1/3} = 1.53 \tag{20}$$

This shows that the fatigue life at the test frequency is 53% longer than in the field.

Power-Cycle Transient Strain. Transient strain is caused by a temporary chip temperature rise above that of the substrate at the start of a power cycle. The transient occurs because the chip is the heat source and has a thermal mass much smaller than the substrate. Consequently, the chip temperature rises much faster than that of the substrate as the power is turned on. The time lag of the substrate temperature profile depends on heat diffusion through the interconnects. Eventually, at steady state, the temperature of the substrate approaches that of the chip, separated only by few degrees due to the product of the bump's thermal resistance and the power through it. The magnitude of the transient peak ΔT_{t} is proportional to the temperature excursion $T_{\mathrm{cm}} - T_{\mathrm{A}}$ and is a function of β (9):

$$\Delta T_{\mathrm{t}} = b\beta^{1/3}(T_{\mathrm{cm}} - T_{\mathrm{A}}) \tag{21}$$

$$\beta = -\frac{1}{t}\ln\frac{T_{\mathrm{cm}} - T_{\mathrm{c}}}{T_{\mathrm{cm}} - T_{\mathrm{A}}} \tag{22}$$

where

β = reciprocal of the temperature profile time constant
T_{c} = temperature at t along the profile
T_{cm} = peak chip temperature
T_{A} = ambient temperature
b = constant of materials, design, and thermal parameters

For glass ceramic-silicon, $b = 0.3$ min$^{1/3}$, and the number of cycles to failure N_{f} is expressed as

$$N_{\mathrm{ft}} = A(b\beta^{1/3}\gamma)^{-2}f_{\mathrm{t}}^{1/3}e^{\phi/[T_{\mathrm{A}} + b\beta^{1/3}(T_{\mathrm{cm}} - T_{\mathrm{A}})]} \tag{23}$$

where

$\phi \approx 1440$ K
γ = strain calculated from $T_{\mathrm{cm}} - T_{\mathrm{A}}$
$f_{\mathrm{t}} \approx 30$ cpd, transient strain frequency

to be compared to the N_{f} steady state for the alumina-silicon system, Eq. (17), which is repeated here:

$$N_{\mathrm{fs}} = A\gamma^{-2}f^{1/3}e^{\phi/T_{\mathrm{cm}}} \tag{24}$$

If the substrate thermal coefficient of expansion (TCE) is a factor of 2 or larger than that of the chip, the transient strain becomes negligible relative to the steady-state strain and of little consequence for any β. This parameter plays the role of the strain rate and is easily calculated from Eq. (22) by using a couple of points on the temperature profile.The number of cycles to failure, which includes both the effect of the steady state and the transient strain, is written as

$$N_{50_{\mathrm{t}}} = A/[\gamma^2 f^{1/3}e^{\phi/T_{\mathrm{cm}}} + (b\beta^{1/3}\gamma)^2 f_{\mathrm{t}}^{1/3}e^{\phi/[T_{\mathrm{A}} + b\beta^{1/3}(T_{\mathrm{cm}} - T_{\mathrm{A}})]}] \tag{25}$$

The first term in Eq. (25) dominates in the case of the alumina substrate, whereas the second term (after the plus sign) dominates when the substrate is glass ceramic. Both terms are important for ceramics, such as aluminum nitride, silicon carbide, and mullite which have intermediate TCEs between silicon and alumina.

Temperature Profile Shape. For a square profile, stress relaxation occurs mainly at the maximum temperature of the cycle, which is reached rather quickly. In such an extreme case, the stress relaxation time above threshold is shorter and therefore corresponds to a higher threshold frequency. Thus a square-cycle frequency has a greater effect on fatigue than a sinusoidal frequency, which is addressed in the model by normalizing the cycle frequency to f_{th}:

$$N_f = A\gamma^{-2}(f/f_{th})^{1/3}e^{\phi/T_m} \tag{26}$$

where f_{th} is determined by Eq. (18) in which the β value is from Eq. (22). Such normalization also accounts for the cycle's tail, where the square profile stress relaxation and creep linger longer with additional fatigue than in the sinusoidal case.

Interconnect Diameter. Another variable affecting fatigue is the joint diameter according to Satoh et al. (10), who studied crack propagation kinetics yielding

$$N_{50} = M(\gamma, f, T_m)\ln[(xL_f + B)/(xL_i + B)] \tag{27}$$

where

$x \simeq 0.0006$/cycle, crack propagation parameter
$B \simeq 0.070$ micrometers/cycle, crack nucleation parameter
L_i = Initial defect dimension, submicron
L_f = Final crack length at failure (L_f = bump diameter)
M = Basic fatigue model

which for $xL_i/B \ll 1$ becomes

$$N_{50} = M\ln\left(\frac{X}{B}L_f + 1\right) \tag{28}$$

Oxygen Partial Pressure. Another important factor in the fatigue life of solder joints is the effect of oxygen. The N_{50} increases with decreasing oxygen concentration in the module environment. It has been shown that oxygen has a strong, adverse effect on the Pb–Sn solder interconnection fatigue life. Oxygen reacts preferentially at the grain boundaries, weakening the fatigue resistance of the joint, so that crack propagation is enhanced during thermal or mechanical cycling. Snowden (11) studied lead fatigue behavior as a function of air partial pressure from atmospheric pressure to 0.0013 Pa. At plastic strains of 0.14% and 0.075%, he showed that the fatigue life of lead increases from a factor of 10 to 60 times as the oxygen concentration goes from 130 Pa to 13 Pa. Berriche et al. (12) showed that the fatigue life of high-lead Pb–Sn solder is much higher in vacuum (0.0013 Pa) and suggested that the reaction of oxygen with Sn to form stannous oxide degrades the fatigue life of the alloy. Later, Di Giacomo (13) investigated high-lead Pb–Sn solder bumps in flip chips, with different oxygen partial pressures within the module. For oxygen concentrations of 20, 200, 1000 ppm, and atmospheric pressure, the fatigue life was determined by thermal cycling between −10° to 100°C at 20 to 46 cycles per day (cpd). The cycles to failure N_f can be expressed by

$$N_f = M/(1 + c^m) \tag{29}$$

where

C = Oxygen concentration in μg/g
M = Basic fatigue model
m = 1/3

Equation (29) can be used to calculate the fatigue life of solder bumps in a hermetic module because C can be calculated as a function of the initial oxygen level inside the module C_0 and because the module leak rate L, internal volume V, and the time t are known. In addition, one needs to know the rate at which oxygen is consumed within the module to determine the actual oxygen concentration at any time. The oxidation kinetics are well known for the Pb–Sn eutectic and the Sn–Bi eutectic (14,15) and for high-lead Pb–Sn from various sources. When the rate Q at which the oxygen is being consumed inside the module is introduced, Eq. (29) is rewritten as

$$N_f = \frac{F(\gamma, f, T_m)}{1 + \left[\frac{L}{(Q+L)}\left(C_i + C_o\left[1 - \left(1 - \exp-\frac{(L+Q)t}{V}\right)\right]\frac{V}{(L+Q)t}\right)\right]^m} \tag{30}$$

where

$Q \simeq naNR(T)$, $m^3(O_2)$/s, rate in open module
$R(T) = R_0 \exp - 10{,}500\,(1/T - 1/T_0)$, $m^3/s \cdot m^2$
R_0 = Oxidation rate at T_0 = 358 K, $4.5 \times 10^{-12} m^3/s \cdot m^2$
n = Number of bumps per chip
a = Lateral area per bump
N = Number of chips per module
L = Leak rate (air), m^3(atm)/s

For $Q \ll L$, Eq. (30) reduces to Eq. (26) because the oxygen consumption rate within the module is negligible relative to the O_2 leak rate. However, if $Q \gg L$, the maximum oxygen level achievable within the module (steady state) is reduced by the factor L/Q. Such an equilibrium level depends on both the leak rate and oxidation rate. The larger the number of solder joints per chip and the number of chips per module, the lower the O_2 at steady state. For example, at 85°C, N = 64 chips, and n = 2500 bumps (5 mils diameter and 3 mils high), the oxygen consumption rate is $2.1 \times 10^{-14} m^3(O_2)$/s or $1.05 \times 10^{-13} m^3$(air)/s. For a leak rate of $L = 10^{-15} m^3$/s, the level of oxygen in the interior of the module reaches a steady-state concentration of 0.045 C_0. For a module with nine chips and 1500 bumps/chip, the steady-state level is 0.36 C_0. Thus, fatigue is significantly affected by the number of joints in a module.

Minicycle Superposition. Minicycles are thermal perturbations during CMOS operation or are caused by the computer going into a "dormant stage" to save energy. Their effect on solder-joint fatigue behavior, as they superimpose on the

main power-cycle profile of the chip, is incorporated in the fatigue model via a normalized minicycle factor which approaches one as the minicycle amplitude goes to zero. The effect depends on the amplitude and frequency of the minicycles relative to the main cycle. To achieve a correct model, an experiment was conducted with thermal conduction modules (TCMs), thermally cycled at 3 cph (cycles per hour), on which minicycles of various amplitudes and a frequency of 1000 cph were superimposed (16). The experiment also included finite-element analysis for many frequencies and amplitudes so that the fatigue model could be based on the relative strains and creep values whithin the realm of the experiment. It was found that the N_{50} for the superposition of minicycles on the main cycle can be expressed simply as

$$N_{\mathrm{Mm}} = A(\gamma_{\mathrm{M}} + \gamma_{\mathrm{m}})^{-2}(f_{\mathrm{m}}\sqrt{F})^{1/3}e^{1441/T_{\mathrm{m}}} \tag{31}$$

where

$$F = (\epsilon_{\mathrm{co}}/\epsilon_{\mathrm{c}}) = 1 + 5.5 \times 10^{4}\gamma_{\mathrm{m}}^{5/3}(1 - e^{-.4(f_{\mathrm{m}}/f_{\mathrm{M}})^{1/3}}) \tag{32}$$

In Eq. (32), ϵ_{co} is the bump creep without minicycles (just main cycles), and ϵ_{c} is the bump creep in the presence of minicycles. Equation (32) has been validated with the fatigue data produced in the experiments described, with and without minicycles, for the conditions under which the thermal cycle tests have been conducted. Equation (31) states that when minicycles are superimposed on main cycles, the two strain amplitudes are added, and the maximum cycle temperature is raised by the ΔT of the minicycles (both changes reduce the joint fatigue life), whereas the main cycle frequency is increased by a multiplier F which reduces to 1 when there are no minicycles. The effect of the minicycle frequency on the main cycles is benign because by shifting it to higher values it reduces fatigue. The reason for this behavior is that the creep rate and total creep accompanying stress relaxation during the thermal cycle are reduced by the higher frequency strain perturbations, which have been experimentally observed.

Chip Underfill. With an underfill, strain is cut by a factor of 5 in most cases. The reduction is caused by the fact that upon bending, the mechanical strain narrows the effect of the thermal mismatch, reducing the strain considerably. Briefly, as the temperature rises, the sustrate expands. However its surface facing the chip contracts as a result of the bending, so that the net surface displacement is much smaller than that calculated from the TCE. On the other hand, the chip surface facing the substrate expands because of the bending so that the total strain (thermal plus mechanical) is greater than that calculated from the TCE. The result is that the thermal mismatch between chip and substrate is effectively reduced by a factor of 5, which explains the improvement in the fatigue life by a factor of 20 to 25. The strain factor of 5 is predicted theoretically based on the thermomechanical properties and thicknesses of chip, substrate, card, and solder joints, with and without a polymer underfill, using fatigue data and finite element modeling (FEM) analysis.

Creep

Creep can affect solder bumps when the chip is subjected to a load, that is, spring or heat-sink attached to the chip, especially if the solder is low-melting. The creep is cyclic, consisting of two components: static creep and cyclic creep. The latter is caused by the shear strain imparted to the joint by thermal cycling. The cyclic component is the cumulative side effect of each fatigue cycle on static creep, which is the time-dependent deformation of the bump due to the perpendicular load. It is believed that the shear stress induced in each cycle regenerates dislocation sources and therefore enhances the static creep for a period until the stress is relaxed to the threshold level. Below the threshold level the cyclic effect vanishes as the dislocation sources are exhausted. Cyclic creep is often explained in terms of excessive vacancies which, according to Felter (17), are generated more easily during cyclic creep than during static creep. According to Hong et al. (18), if excess vacancies generated by cyclic stress form climb dislocations, plastic deformation is more conducive as the vacancy-generating process is accelerated.

Static Creep. Static creep in electronic package interconnects is quite different from that traditionally observed under a constant load, characterized by primary, secondary, and tertiary stages. In solder interconnects, the initially applied stress normal to the solder joint decreases as the cross-sectional area increases due to creep. Generally, the stresses involved in solder bump creep are low, which is in the realm of elastic strain and microcreep, controlled by grain boundary diffusion. Therefore we do not explore creep behavior outside the realm of microelectronics applications. Solders behave viscoelastically. However, there are other forms of creep. Plastic-viscous creep is characterized by stress power (19). Ashby and Werrall (20) combine the diffusion mechanism with the power law to get a viscoelastic relationship. Weertman (21) describes creep by a hyperbolic sine and power-law equation.

But we are concerned with microcreep in solder joints at stress much below the elastic limit, i.e., <1 MPa to 2 MPa, controlled by diffusion creep which tends to saturate with time, as determined experimentally. As mentioned, the bump cross-sectional area increases as the solder joint deforms under normal load. Thus the stress diminishes with time, as the static creep rate approaches zero. As a result, creep will appear to have a weaker dependence on the stress as it is allowed to remain fixed in the model formulation. In addition, solder joints have a host of intermetallics that form at the interface and also throughout the bulk, involving the thin-film metals that react with tin in the joining process. As the solder is compressed by the load, the intermetallics impede the solder flow more and more while the ductile solder is squeezed out. The result is that the force required to maintain active creep becomes greater and greater, leading to a low-power equation which states that solder joint creep, as measured in practice in the presence of intermetallics and with an expanding cross section, is not very sensitive to stress, as expected from the mechanisms that appear in the literature. It has been determined experimentally that for high-lead Pb–Sn solders, solder bumps can withstand up to 0.0015 kg per joint (0.000125 m diameter) or 1.25 MPa for the lifetime of the module at a maximum cycle temperature of 100°C. However, for low-melting solders, the safe limits of design and operation are more stringent because of the high creep rate associated with these solders. The failure criterion is the degree or percent bulging of the solder bump that one can safely allow before the probability of contact between two ad-

jacent bumps becomes unacceptable in terms of reliability. One can scale creep with respect to the thermal activation term by expressing the exponent as the ratio of melting point temperature to the actual creep temperature for different compositions of a solid solution alloy system:

$$\gamma_c = Me^{-c\left(\frac{T_M}{T}\right)} \tag{33}$$

where

C = 12.9 for Pb–5%Sn solder
T_M = Melting point temperature
M = Constant

Cyclic Creep. Under test conditions, the cyclic creep component is larger than the static creep component because the cycles expected in the field lifetime are crowded into a relatively short test period. Cyclic creep is a function of the number of cycles rather than time because each cycle induces a shear "burst" of stress (and strain) which relaxes in a finite time to a threshold level. Cyclic creep has been studied in depth with flip chips (22). The work describes the static and cyclic creep of Pb–5%Sn solder bumps as a function of applied stress, temperature, dwell time, relaxation characteristic constant, footprint, number of cycles, and creep time. A model based on static and cyclic creep test data was developed and verified. Each cycle adds a finite amount of creep which has the duration of a transient shear-stress cycle. The duration of the stress cycle is the time from the initiation of the stress to the point where it subsides below the stress threshold. Figure 7 illustrates creep enhancement superimposed on static creep. On the basis of the experimental results, the total creep ϵ_T, depends on the shear strain, temperature excursion, number of cycles, maximum temperature, and initial stress applied normal to the chip. Static creep can be written as

$$\epsilon_s = 0.11(\sigma_0 t)^{2/3} e^{\frac{-0.33}{KT_c}} \tag{34}$$

while cyclic creep is given by

$$\epsilon_c = 7.1\sigma_0^{4/5} N_c^{2/3} \gamma^{1/2} \left[1 - \exp\left(-\frac{t_d}{\tau}\right)\right] \exp(-0.33/KT_c) \tag{35}$$

and the total creep is the sum

$$\epsilon_T = \epsilon_s + \epsilon_c \tag{36}$$

where

α_s = substrate thermal coefficient of expansion
T_c = chip temperature
α_c = chip thermal coefficient of expansion

and

$$\gamma = \frac{\text{DNP}}{h}[(T_s T_A)\alpha_s - (T_c - T_A)\alpha_c] \tag{37}$$

where

T_a = ambient temperature
h = height of bump
γ = average shear strain
σ_0 = initial stress
N_c = number of thermal cycles
t_d = dwell time
$\tau = 15' \exp(-0.33/K)(1/T - 1/373\ \text{K})$, minutes
$t = N_c \times t_d$, static-creep time
K = 8.3143 J/mol · K

Equation (35) predicts that under stress of 1.27 MPa normal to the bump (0.0032 kg on a 0.000125 m diameter bump) for 2500 thermal cycles 0°C to 100°C, at a frequency of 3 cycles per hour, and 10 min dwell per cycle, the cyclic creep is 16%, whereas Eq. (34) predicts that the static component is only 0.7%. However in the field, under the long-dwell cycle which favors static creep (40 h/cycle), for a stress of 0.14 MPa, static creep is 3.9% and cyclic creep is 2.8%. The sharp difference in the static/cyclic creep ratio between test and field is caused by the short static creep time during test.

ELECTROCHEMICAL MODELS

Corrosion

Corrosion is a prevalent failure mechanism which causes open circuit failures. The speed of corrosion is controlled by rate-limiting factors, e.g., diffusion, and depends on the magnitude of the activation energy. The change in free energy associated with corrosion characterizes the metallic state upon reaction and determines whether a reaction is thermodynamically feasible. It does not necessarily determine the corrosion rate.

Potential-Driven Corrosion. Electrochemical corrosion occurs with the generation of electrolytic current through an electrolyte driven by an applied voltage. The electrolyte consists of adsorbed/condensed water films in which ionic species are dissolved. When no voltage is applied, the corrosion is galvanic, which operates under the driving force of local electrical potentials due to material heterogeneities. The corrosion rates are higher when a potential is applied. Because electronic packages involve polymer encapsulants and corrosion films, the reaction rates are diffusion-controlled. One must consider chemical and physical rate-limiting processes to arrive at a correct model formulation. Reliability modeling is based on a few key variables, which address the application conditions, process, and the environment. Among the variables are relative humidity, temperature, voltage across conductors, concentration of atmospheric reactive species, process residues, and finally leachable package contaminants extracted by diffusing water and transported to the corrosion site under a concentration gradient. Catalysts, such as chlorides, promote and accelerate the reactions and formation of corrosion products. Chloride ions form chlorides with most metals. Chlorides are generally water soluble and allow the metal ions (in this case Pb^{++}) to react with less aggressive anions, such as carbonate ions, to precipitate the insoluble and stable compound, lead carbonate. The chloride ions thus freed renew the process, finally leading to failure. Solder joints, pad metallizations, and aluminum circuitry are sensitive to traces of chloride ions. Metals, such as Cr, Ti, Ni, Co, Mo, and Pd, commonly found in thin-film and pad metallizations, are affected by chlorides to a degee that depends on

the difference in the oxidation potential between adjacent films in the structure. The kinetics of reaction depend on the activation energy, the relative humidity, and the electric bias. In the case of an encapsulant, the activation energy is that of diffusion regardless of whether the corrosives come from outside or from the encapsulating material, that is, hydrolyzable chloride in epoxy and SO_2 or CO_2 from the environment. If the metallization is in direct contact with the reactive gaseous species or electrolytes, the activation energy is that of forming activated complexes, which then decompose at a finite rate, giving rise to stable corrosion products. If two compounds are being formed, their rates of formation will differ (unless they have the same activation energy): at higher temperature, the compound with the higher activation energy will predominate, whereas at the lower temperature, the one with the lower activation energy will be favored. In such a case, the experimental results will yield an apparent activation energy approaching the larger of the two at the higher temperature, whereas it will approach the value of the lower activation energy at the lower temperature. Therefore, experimentally, one cannot obtain a straight line across the temperature range when the reciprocal of the time to failure t_{50} is plotted against the reciprocal of the absolute temperature. The curve is concave upward, characterized by a decreasing slope as $1/T$ increases. Figure 13 reflects the dual mechanism and associated activation energies. Relative humidity is by far the strongest factor in the corrosion process. Water vapor condenses on surfaces when its temperature drops below the dew point. The vapor is adsorbed on surfaces according to the BET adsorption isotherm which describes multilayer molecular adsorption as a function of relative humidity. The higher the relative humidity, the larger the number of water monolayers built on the surface, and the greater the probability for chemical or electrochemical reaction in the presence of reactive ionic species. The water film becomes an electrolyte capable of sustaining localized or more widespread corrosion depending on the availability and continuity of the water film. The BET isotherm, named after the discoverers, Brunauer, Emmett, and Teller (23), expresses the adsorbed water-film thickness δ, which is inversely proportional to the corrosion time to failure (experimental) and is a function of the relative humidity H:

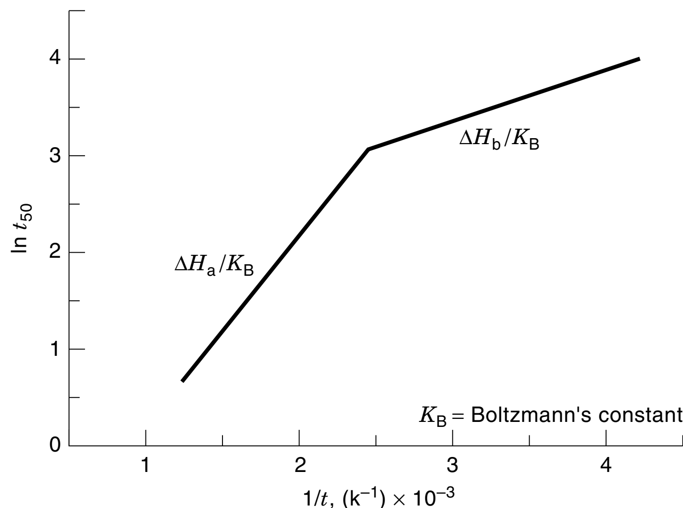


Figure 13. Illustration of two activation energies across an operating temperature range.

$$\delta \propto \frac{1}{t_f} = \frac{H}{K(1-H)[1+(\zeta-1)H]} \tag{38}$$

where

$\zeta \simeq \exp(\Delta H_v - \Delta H_A)/KT$, determined experimentally
ΔH_V = 44 KJ/mol, heat of water vaporization
ΔH_A = heat of adsorption which depends on the material and surface conditions
H = relative humidity
K = 8.3143 J/mol · K

which shows that as the relative humidity increases, the time to failure decreases. In general, corrosion time-to-failure data fit Eq. (38) for $\zeta \ll 1$, which dictates that ΔH_A be greater than ΔH_V. Therefore the usual form employed is

$$t_f = B\frac{(1-H)^2}{H} \tag{39}$$

The advantage of Eq. (39) is that it is based on adsorption and condensation phenomena and surface physics, which allows one to extrapolate the corrosion results with more confidence to lower relative humidities, for which no data exist. This is not the case when strictly empirically fitting a narrow range of data. This can lead to gross deviations, orders of magnitude apart, when projecting to field humidities of ≤10% at the corrosion site under operating conditions. It is much safer to adhere to a physical approach, such as the BET model, which has a more scientific basis. Several empirical forms used in chip reliability were developed by Lycoudes (24), Peck (25), Merrett, Bryant, and Studd (26), Sherley and Hong (27), Sim and Lawson (28), Polman and Fokkens (29), Anolick et al. (30), Guan et al. (31), Lawson (32), and Tam (33). Figure 14 is a plot of the relative humidity accelerating factor for some of the models versus H, demonstrating how drastically different the field projections are. The models agree within a factor of 3 down to 40% H. However, they differ by orders of magnitude around ≤10% H, which is what a device experiences in the field.

The voltage factor plays a fundamental role in corrosion and promotes the corrosion rate beyond the galvanic stage. For alloys, such as Pb–Sn solders and Al–Cu lands, there are always concentration gradients, precipitates, and intermetallics which promote galvanic corrosion. When an external electrical bias is applied, the corrosion rate is enhanced to an extent which depends on both the intensity of the local potentials and the applied voltage itself. The corrosion rate is normalized to the galvanic corrosion rate, so that when the applied voltage is zero, the model reverts to the galvanic component. The time to failure is given by (1)

$$t_f = A\frac{(1-H)^2}{H}\frac{e^{\Delta H/KT}}{1+\sinh \mu V} \tag{40}$$

where

ΔH = activation energy, J/mol
K = 8.3143 J/mol · K

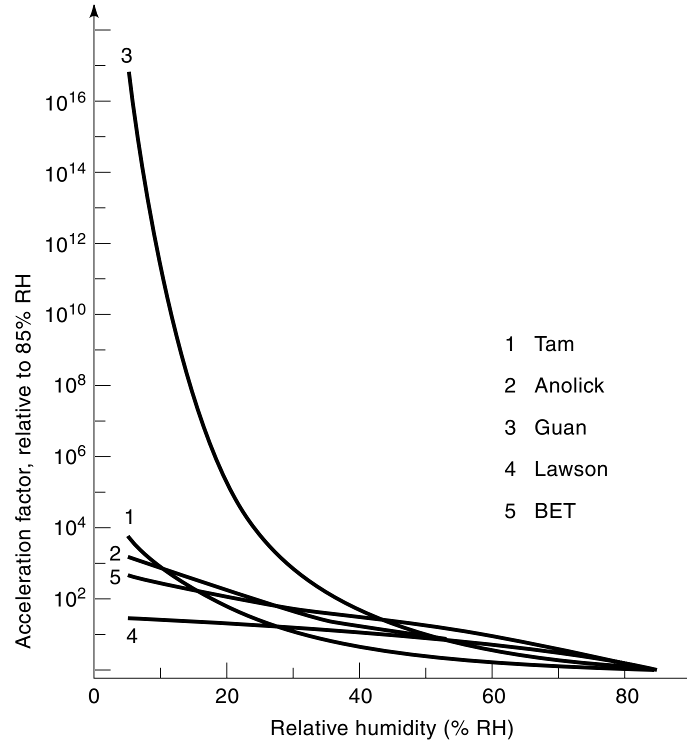


Figure 14. Graphical comparison of RH-factor empirical models.

H = relative humidity
T = absolute temperature, K
V = voltage, V
μ = voltage parameter, 0.05 to 0.10/V
A = constant of the materials, time

The voltage factor in Eq. (40), in particular, was applied to temperature–humidity corrosion under voltages varying from 0 to 25 V. The results shown in Fig. 15, fit Eq. (40) for μ = .075. The time to failure is related to the stress conditions via the proportionality constant A, which is a characteristic of the particular material system and process. In fact, the constant differs as the process varies because of different impurities

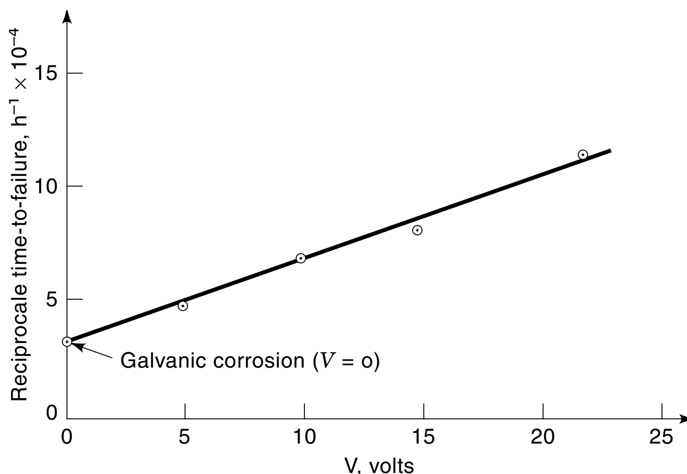


Figure 15. Normalization of the electric bias effect on corrosion.

and process residues. The proportionality constant and activation energy should be redetermined for a different process and materials system. The presence of chlorides in residues or in packaging materials causes major shifts in these two constants by altering the reaction kinetics through a low activation energy path. The effect of chlorides on corrosion is evaluated and quantified as a function of its concentration by Eq. (41) given here (1):

$$t_{\mathrm{f}} = B\frac{1+\beta C}{\beta C}\frac{(1-H)^2}{H}\frac{e^{\Delta H/KT}}{1+\sinh \mu V} \tag{41}$$

where β is a concentration parameter C determined at different levels of chloride. For values of $\beta C \ll 1$, the concentration factor in Eq. (41) varies directly with the reciprocal of the chloride concentration, whereas for $\beta C \gg 1$, the factor approaches 1, and the time to failure is independent of C. For intermediate values of βC, the dependence is sublinear. Based on corrosion studies, β is on the order of 10,000 m^2/kg when concentration is expressed in kg/m^2. For leachable chloride bulk concentrations in epoxy, the parameter $\beta \simeq 0.005$/ppm, which is empirically determined with a number of chlorides. The chloride factor has the form of the Langmuir adsorption isotherm which describes surface corrosion and confers a physical character on it. When dealing with gases instead of electrolytic solutions, the factor is expressed in terms of the partial pressure P instead of C: $\beta' P/(1 + \beta' P)$. This is an example of chemical corrosion unaccompanied by the generation of electric current. Notorious for corrosiveness is lead sulfide, PbS in S_8 vapors. Corrosion, however, is practically eliminated by hermetic modules for their lifetime. One can express a critical quantity Q_c of reactive species, such as Cl and S, that causes failure in time t_f, as a function of species leak or permeation rate at atmospheric concentration (α, kg/s), species reaction rate at atmospheric concentration (β, kg/s), and internal module volume V under accelerated test conditions of temperature, relative humidity, and bias:

$$Q_{\mathrm{C}} = \frac{\alpha\beta}{\alpha+\beta}\left[t_{\mathrm{f}} - \frac{V}{\alpha+\beta}\left(1-\exp -\frac{\alpha+\beta}{V}t_{\mathrm{f}}\right)\right] \tag{42}$$

When $\alpha \ll \beta$, the corrosion process is controlled by the permeation or leak rate. Whereas if $\beta \ll \alpha$, the process is controlled by the corrosion rate. The values of α and β are determined experimentally by measuring the air leak rate of the module (multiplied by the species concentration) (m^3/s)(kg/m^3) = kg/s and determining the corrosion rate in an open environment, using nonhermetic modules.

According to the transient term of Eq. (42), the partial pressure of the reactive species reaches a steady state when $\beta = \alpha$, which may be minutes, hours, or even days depending on the degree of hermeticity and the initial disparity between β and α. It must be emphasized that the values of β and α represent the corrosion rate in an open environment and the leak rate at start, respectively, assuming that the module has no pollutant inside at time zero. These rates, however, change within the module as the species partial pressure rises and approaches the steady-state level at which $\beta = \alpha$.

Galvanic Corrosion. This type of corrosion prevails in humid environments without applied voltages when the metal alloys have precipitates, intermetallics, or heterogeneous

compositions. The model for this type of corrosion is expressed by Eq. (41) for $V = 0$. In most cases the corrosion rate is diffusion-controlled when the reaction is limited by the transport rate of a reactant through a material, that is, a polymeric film, corrosion product, or an encapsulant. The corrosion rate increases or decreases depending on whether the permeability of the material is raised or lowered by a temperature change or by design. For a diffusion-controlled process, the thickness of an oxide or lead carbonate layer grows as the square root of the time(parabolic growth), whereas the rate diminishes as the reciprocal of the square root of the time. The behavior is observed in other metallizations. Aluminum orthophosphate forms at the interface of aluminum films and phosphosilicate glass (PSG) forms in the presence of moisture, when the phosphorous pentoxide (P_2O_5) is not properly diffused or driven into the glass structure through high temperature annealing. Some of the P_2O_5 remains free and picks up moisture to form orthophosphoric acid. It has been shown that the reaction is diffusion-controlled by the corrosion product (34). The time to fail t_f is given by

$$t_{\mathrm{f}} = \frac{\beta \Delta h_{\mathrm{c}}^2 e^{\beta\tau}}{2a^2 DC} \tag{43}$$

where

$DC = 1.86 \times 10^{-9}\, e^{-\Delta H/KT}$, kg/m · s
$\beta = 0.30$/min
τ = annealing time, minutes
$\Delta H = 53$ kJ/mol, activation energy of permeation
ρ = kg/m^3, density of phosphorous in the reacted layer
$\Delta h_c = m$, corrosion layer thickness

Polymer Film or Underfill Protection

Corrosion in the presence of a polymeric coating or an underfill is attenuated because of the polymeric protection. However, cleanliness of the interface and a strong bond are essential. If the interface is a continuous network of polymer bonds without gaps or contamination, the chance of corrosion is remote. In such an ideal situation, the polymer absorbs water until it saturates to the level dictated by the outside vapor pressure, acccording to Henry's law. The water remains in solution as long as there is no loss of adhesion or interfacial defect where corrosion can nucleate. Otherwise corrosion starts in those sites which become sinks to water and other corrosive species surrounding the area. As the corrosion product builds up, there is more water absorption so that the site becomes an electrochemical local cell (galvanic corrosion) which extends further along the interface by capillary action between the joint and polymer and degrades the interface chemically and by the pressure exerted by the corrosion product, leading to failure. Which species are rate-limiting in the process? Carbon dioxide (CO_2) in the atmosphere has a partial pressure of ~26 Pa whereas that of water vapor is two orders of magnitude higher. The water permeation rate through common encapsulants is also much larger than that of CO_2. Therefore, CO_2 controls the kinetics of lead carbonate formation, given that one molecule of each forms lead carbonate. What is important is the rate at which the carbon dioxide is supplied, which is dictated by its diffusivity and solubility in the polymer. If CO_2 were not rate-limiting, it would build up near the corrosion site, thus reducing its concentration gradient until a steady state is reached ($\alpha = \beta$).

The model that describes corrosion in the presence of a polymer has the same form as that for corrosion occurring without a protective film, except that the activation energy and the constant of proportionality are different. For a coating, the activation energy is that of permeation, whereas without the film, the activation energy is that characteristic of the reaction of the given species with the metal. When water is rate-limited by diffusion, the activation energy through polyimide is 40.5 kJ/mol, and through an epoxy the activation energy is 38.6 kJ/mol independent of the solder. Because permeability, which is the product of diffusivity and solubility, is different above the glass-transition temperature T_g, one must ascertain that the accelerated tests are performed below T_g, otherwise the test results are not relatable to the lower field temperature, which is assumed to be below T_g. Generally, the solubilities of light gases, such as N_2, CH_4, and CO_2, are low enough to be within the limits of Henry's law and thus increase linearly with their partial pressure in the atmosphere. Among the more notable reactants that promote corrosion and become part of the corrosion products are water vapor (H_2O), sulfur dioxide (SO_2), hydrogen sulphide (H_2S), nitrogen dioxide (NO_2), ammonia (NH_3), and chlorine (Cl_2). Concentrations vary from thousands of ppm for water vapors to a fraction of one ppm, which may be the case for Cl_2.

An important aspect of a polymer or elastomer film is the time delay it causes before corrosion can start. The delay results from the fact that pollutants from the atmosphere must diffuse through the film before they reach the metal interface, assuming that corrosion initiates when the first molecule arrives there. The delay time, calculated from Crank (35), gives the amount diffused as a function of time:

$$Q = \frac{C_0 D}{h}\left[t - \frac{h^2}{3D}\right] - \frac{2hC_0}{\pi^2}\sum_{1}^{\infty}\frac{1}{n^2}e^{-\pi^2 n^2 Dt/h^2} \tag{44}$$

For $h^2/3D \geq t$, $Q = 0$, which defines the delay time τ, when Q just becomes positive by receiving the first molecule:

$$t = \tau = \frac{h^2}{3D} \tag{45}$$

where τ is the extrapolated t value for $Q = 0$ when the exponent approaches zero. For SO_2, the diffusivity in epoxy is about 5×10^{-13} m^2/s at 45°C. If the encapsulant is 0.00002 m thick, the delay time is approximately 1500 h, which is considerable. Other examples can be worked out similarly.

Metal Migration

Metal migration through an electrolyte between thin-film conductors is another important failure mechanism caused by metallic dendritic growths across the gap under the driving force of an electric field. Figure 16 shows a Cu dendrite grown in a dilute solution of H_2SO_4 between Cu wires under a 5 V bias. For metal migration to occur, the current density flux must be above the critical value which varies from metal to metal. The mechanism requires a sufficient film of water to condense or adsorb in the gap, on the chip, or substrate surface to provide the migration medium for the ions, which migrate from the anode to the cathode, where they deposit and

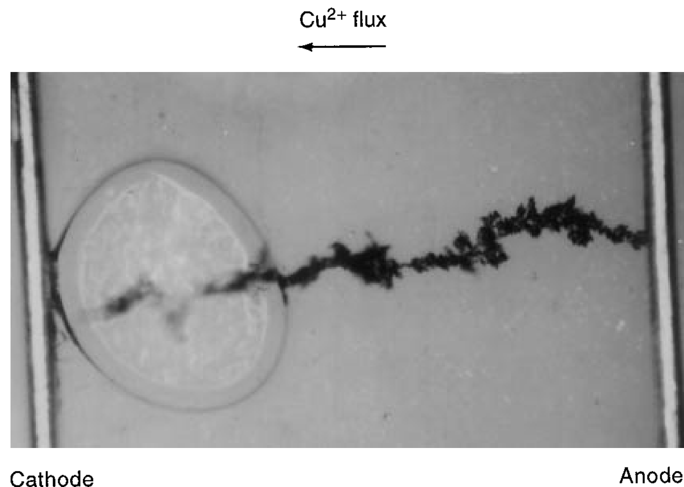


Figure 16. Cu dendrite grown from Cu wires in dilute H_2SO_4 solution. Courtesy A. Cammarano.

reduce on metallic protrusions to form dendrites. The electric field in the vicinity of the growing dendrite provides a focusing action for the ions, allowing a high deposition rate capable of sustaining dendritic growth without being poisoned by impurities. There is, however, a current density threshold associated with the electrolyte purity below which the dendrite cannot continue to form. This is the case when the crystallographic facets at the dendrite tip become poisoned, impeding orderly atomic deposition, because the energy to sweep the atomic step across the surface is too high. The closer the poisoning particles are, the harder it becomes for the crystallization process to proceed.

The migration mechanism requires chemical interaction between the surroundings and the metal to generate metal ions, a polar transport medium through which ionic migration occurs (i.e., water), and an electrical field. But for water to provide a continuous film of adequate thickness, the relative humidity has to be relatively high on a clean surface. If residues are present, aqueous condensation occurs at much lower levels of humidity, and therefore metal migration is more probable, assuming that the other conditions are met. Therefore, one can speak only of the probability of metal migration, which increases with the degree of aqueous condensation, which in turn depends on the relative humidity. On the other hand, the metal's susceptibility to migration depends on the oxidation potential, the chemistry of the electrolyte at equilibrium, the energy state of its surface, and the tenacity of the oxide layer, if any, through which ionic species are exchanged to keep the process active.

There is competition between the metal deposition rate and the adsorption rate of the capillary-active impurities. If the concentration of impurities is sufficiently low, they are easily buried by the growing lattice planes of the depositing metal. The critical current density corresponds to the steady state at which the diffusion/deposition rate of the impurities at the surface is equal to the rate at which they are buried by the metal deposition:

$$\dot{Q}_{\mathrm{TIP}} \geq \rho l / M t_{\mathrm{b}} \tag{46}$$

where

Q_{TIP} = metal deposition rate
ρ = Density of dendrite
l = Distance between biased circuit electrodes
M = Atomic weight
t_{b} = Maximum time to bridge without poisoning

The poisoning effect is explainable by the lattice-step, edge-energy per unit area E and the free energy decrease per unit volume ΔG in the deposition process (36,37).

The failure distribution that one obtains from the migration mechanism results from a long-term effect (incubation) which is a function of temperature, voltage, and relative humidity. One can think of metal migration as the result of a degradation mechanism which leads to the propitious conditions amenable to ionic migration, though the time for the dendrite to grow across the gap is of the order of seconds or minutes. Therefore the failure distribution reflects the time for each site to become active, which controls the kinetics and therefore the time to failure, which is described by the same variables as the corrosion model. In fact, the metal migration mechanism is an electrochemical process in which the oxidation-reduction involves no corrosion product because the metal ions are transported, reduced, and deposited at the cathode in the metallic state rather than combining with anions to form salts, hydroxides in situ. The model's relative humidity factor is the same as that for corrosion, based on the BET adsorption isotherm (38,39). This has been amply verified, whereas the thermal factor has an activation energy for hydrated-ion diffusion through the water film, which is independent of the ion type. The migration dependence on voltage (electric field), however, must meet the conditions that the process halts when the electric field goes to zero. This is not true for the corrosion mechanism where local reaction still occurs (galvanic) in the absence of an outside potential. The model is given by

$$t_{\mathrm{f}} = A' \frac{1+\beta C}{\beta C} \cdot \frac{(1-H)^2}{H} \frac{e^{\frac{\Delta H}{KT}}}{\sinh \frac{\mu (V - V_{\mathrm{T}})}{d}} \tag{47}$$

where V_{T} is the voltage threshold that can vary from 0.5 V to 2 V depending on contamination and local alloy composition. It is usually assumed zero for tests run at relatively high voltages. The parameter d is the gap between the anode and the cathode, and $\mu \simeq 0.0000075$ m/V, to fit migration's linear dependence on low voltages prevailing at use conditions. ΔH has been shown to be $\leq$0.3 eV, reflecting the diffusion-controlled process through water.

FORCED-DIFFUSION TRANSPORT MODELS

Electromigration

This mechanism affects solder interconnects and circuitry, especially aluminum lands and to a lesser extent Al–Cu, where Cu can be a maximum of 4% by weight. The electromigration failure mechanism becomes significant at 0.10 GA/m^2 for solder joints and an order of magnitude higher for Al–4%Cu. The failure mechanism of mass movement of the matrix metal ions in the direction of the electron flow causes metal deple-

tions, voids, and finally an open circuit. It is a cumulative effect which gradually produces shallow via-voids in solder bumps, thus reducing the cross section of the pad interface, which increases the current density and accelerates the degradation process until failure. Failure occurs precipitously under high test stress when the coalescing voids virtually extend across the entire cross section of the via, such that the last vestige of conducting metal melts due to Joule heating. As the void extends throughout the via, the thin-film metallization or ball limiting metallurgy (BLM) in the sidewall step must carry the total current. If the thickness of the thin-film structure is inadequate to carry the current load, failure occurs catastrophically as a result of thin-film melting. In field conditions, this mechanism is not of much concern for solder joints. Aluminum and Al–Cu stripes are similarly affected by electromigration. Opens occur where there is a mass flux divergence due to grain size differences along the stripe: large grains followed by small grains in the direction of the electron current, which provides more grain boundaries for the mass outflow (Fig. 17).

Electromigration results from two driving forces, the electric field and the "electron wind." The electric field exerts a force on the activated ions in a direction opposite to that of the electron flow. By colliding with activated metal ions, the electrons exert a force on them in the electron flow direction. Therefore, the activated ions are simultaneously subjected to these two competing forces. The metal ions (atoms with a net screening charge of *eZ**) drift in the direction of the predominant force, which for most common metals used in microelectronic circuits (including Pb–Sn solders) at moderate temperature is in the direction of the electron flow. The process is one in which the electrons transfer their momentum to the metal ions which will be able to jump across the energy barrier if the momentum added to the thermally activated ions is sufficient. In such cases, the drift of the metal ions is toward the anode or positive end, whereas the generated vacancies move toward the cathode. The two forces, electrostatic field and electron momentum transfer, however, need not have an opposite effect. A more complete statement is that if the drift occurs in the direction of the electron flow, then the electron momentum transfer is either dominant over the electrostatic force on a positive *eZ** or the electrostatic force is acting on a negative *eZ**, in which case it is in the same direction as the electron flow. On the other hand, if the drift is in the direction opposite to the electron flow, then *eZ** can be only positive and the electrostatic force predominates over the electron flow. The material properties which are important with respect to electromigration are the metal's grain boundary diffusion (self-diffusion or diffusion of a significant alloying element usually in solution), the activation energy, and the proximity of the operating temperature to the melting point of the metal or alloy. The activation energy for grain boundary diffusion is usually about half that of lattice diffusion. Therefore, the finer the grain size, the more predominant grain-boundary diffusion becomes at temperatures where both diffusion modes are significant. The minimum temperature at which bulk diffusion starts to dominate increases as the grain size diminishes. The question is often asked why electromigration, which invoves the transfer of a relatively small amount of energy to the drifting ions (compared to the activation energy of self-diffusion), has such a formidable effect on the failure mechanism. In general, such transfer energies amount to a few hundredths of one electron volt, hardly sufficient to dislodge the atom from the lattice. However, it may be adequate to cause ion drift by colliding with thermally activated ions which need only a small force or momentum transfer to make the transition across the energy barrier. The generic electromigration model is a function of the diffusivity of the diffusing element in an alloy (or self-diffusion in a metal), the current density, the stress differential between anode and cathode, the length of the conductor, the atomic volume, the electric resistivity, and the effective charge. It includes the backflow compressive stress. The derivation of the electromigration velocity based on the modified Nernst–Einstein expression (40,41), is written in terms of diffusivity D, electric field E, effective charge eZ^*, the constant K, and temperature T:

$$v = DEeZ^*/KT \tag{48}$$

Based on Ohm's law, Eq. (1) becomes:

$$v/J = \frac{\rho eZ^* D_0}{KT} e^{\frac{-\Delta H}{KT}} \tag{49}$$

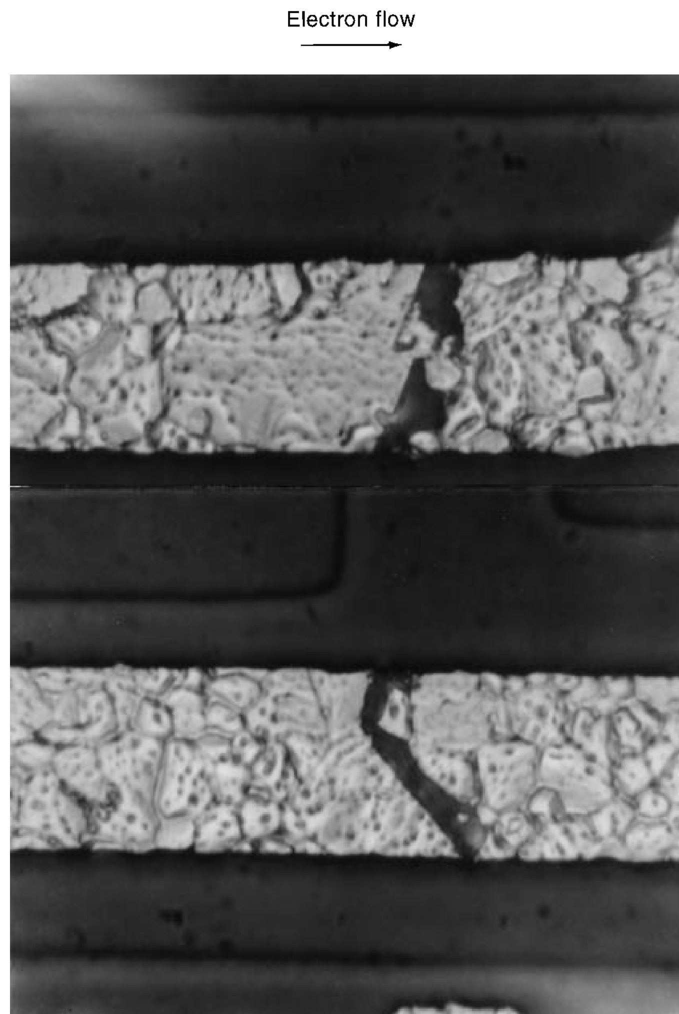


Figure 17. Electromigration circuit open where fine grains follow coarse grains in the electron flow direction (courtesy A. Cammarano).

and because (ρ/T) and eZ^* are reasonably independent of temperature

$$v = AJ \exp \frac{-\Delta H}{KT} \tag{50}$$

which is expressed in terms of transport rate per unit area normal to the flow as

$$\dot{Q} = \sigma v = A'J \exp \frac{-\Delta H}{KT} \tag{51}$$

where

$A' = \sigma(\rho/T)eZ^*D_0$
σ = Metal density

In addition, Eq. (48) states that v and D have the same temperature dependence. From Eq. (51), if a critical amount Q_c per unit cross section has to be depleted to cause a failure in the circuit, then

$$\frac{Q_c}{\dot{Q}} = t_f = \frac{B}{J} \exp \frac{\Delta H}{KT} \tag{52}$$

At field temperatures, high-lead Pb–Sn bumps that fail by electromigration do so through grain boundary diffusion. For example, for grain-sizes of a few micrometers in diameter, grain-boundary diffusion predominates up to a temperature around 60% of the melting point $T \le 0.6\ T_m$. However, if the grain size is increased by one or two orders of magnitude thus enormously reducing the total grain boundary's cross-sectional area, then electromigration is controlled by the lattice diffusion through the bulk of the grains. Grain boundary width is assumed to be a couple of angstroms. The effective diffusivity D_{eff} can be expressed as

$$D_{eff} = D_v + \left(\frac{\pi \delta}{d}\right) D_g \tag{53}$$

where

δ = Grain boundary width
d = Grain diameter

Increasing d in Eq. (53) emphasizes bulk diffusivity and accordingly curtails electromigration. One must also consider the effect of the backflow compressive stress that develops in the conductor (more compressive downstream) and is responsible for an electromigration threshold. In fact, the backstress can completely annihilate the mass transport rate due to electromigration. Such a condition is met when the product of the current density and the length of the conductor equals the product of the stress gradient and the atomic volume divided by the product of the effective charge and resistivity. This is an important aspect of the model because conductors are constrained by passivating films, whereas the chip solder-joints or C-4s are constrained by underfill materials. Under such conditions, compressive stress gradients can achieve sufficient magnitude to counter the effect of electromigration significantly. Blech and Herring (42) and Blech (43) showed the effect on TiN with aluminum films having different lengths, which developed stress gradients opposing the mass flow and were responsible for the current density threshold. The drift velocity is given by

$$V_d = (\rho e Z^* J - \Delta\sigma\Omega/L)(D/KT) \tag{54}$$

where

Ω = Atomic volume
L = Bump height
$\Delta\sigma$ = Backflow stress

For a sufficiently short conductor, the back stress will completely annihilate the electromigration rate when this condition is met:

$$(LJ)_c = \frac{\Delta\sigma\Omega}{eZ^*\rho} \tag{55}$$

where L_c is the critical length below which the drift velocity is zero for a current density satisfying Eq. (55). Based on this equation, one can improve electromigration resistance by using alloys consisting of metals with high $\Delta\sigma/eZ^*\rho$ values or high $E/eZ^*\rho$, where E is the elastic modulus. For this purpose, t_f is given by

$$t_f = \left[\frac{A}{kT}\left(1.6 \times 10^{-15} Z^* \rho J^n - \frac{10^6 \Delta\sigma\Omega}{L}\right) e^{-\Delta H/KT}\right]^{-1} \tag{56}$$

where

ΔH = activation energy of diffusion, J/mol
K = 8.3143 J/K · mol
n = experimental parameter
$\Delta\sigma$ = stress differential, Pa
A = proportionality constant, m/h · mol
Ω = atomic volume, m^3 (for Pb, 3.03×10^{-29} m^3)
Z^* = effective charge number

When one plots the reciprocal time to failure $1/t_f$ versus J, one obtains a straight line displaced to the right of the origin where there are no data for J values comparable to the field. It would require years of testing at very low current densities to determine the shape of the curve at such low stress values. As a result, one either fits the data with a straight line and introduces a J_{TH} intercept or fits the data by forcing the curve through the origin, in which case an exponent for J greater than 1 is obtained, as shown in Fig. 18. Both approaches are suitable and reasonable, though the latter is generally employed because it is more conservative. The empirical J exponent is in the range 1.6 to 1.8 for solder bumps, and ~2 for aluminum lands. An empirical model derived from Pb–Sn film depletion data obtained at different temperatures and current densities for a number of Pb–Sn alloys was developed by Di Giacomo (44):

$$t_f = \frac{C}{(J - J_{TH})} \exp \frac{\Delta H}{KT} \tag{57}$$

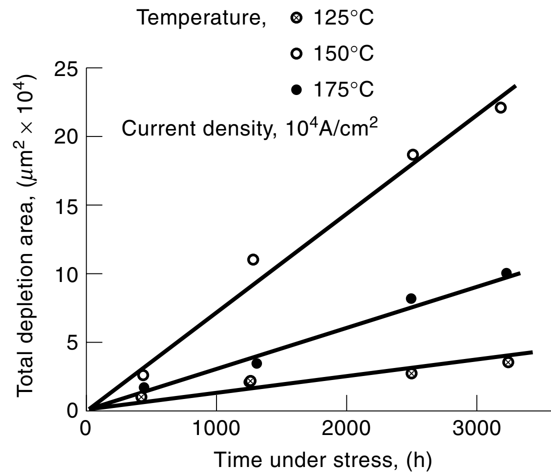


Figure 18. Solder depletion area vs. time at various temperatures. Source: Di Giacomo, G. 1979, "Electromigration Depletion in Pb–Sn Films." *17th Ann. Proc., Reliability Phys.* © 1979, IEEE.

where

C = 80,000 A·h/m^2
J_{TH} = 4.5 × 10^7 A/m^2
ΔH = 58 kJ/mol

Based on the model, at 85°C and 1.45 × 10^8 A/m^2, the time for 50% depletion t_{50} is ~10^6 h for a lognormal distribution sigma of 0.44. The Pb–5%Sn films were wide enough to have hundreds of depleted areas distributed over the film width. The results show that mass transport is linear with current density and that there is a current density threshold below which electromigration practically ceases. In fact the lowest current density stress was 5 × 10^7 A/m^2 (slightly above the threshold), at which the depletions were practically nonexistent. Note that the threshold here is based on direct observations of the damage in large and homogeneous films. A limited number of films do not reflect the statistics of solder bumps and therefore are more amenable to expression in terms of a threshold model. In fact, the results prove the existence of the threshold and its magnitude, though in practical statistics, the higher power of J is adopted because it is more representative of the process (affecting the threshold) and the statistics associated with it. Figure 18 shows the depletion rate versus time at various temperatures, from which the activation energy of diffusion was determined as 58 kJ/mol. Mass depletion occurs predominantly in those grains having the (100) and (111) planes parallel to the film surface and oriented in the ⟨110⟩ direction with the electron flow, most likely to occur because they form the most stable energy configurations on the basis of nearest neighbors (45). From the average mass transport calculated from the depletions, an effective electromigration velocity v and diffusivity D were also obtained at J = 10^{12} A/m^2, T = 423 K, and Z^* = 7 (46):

$$v = \frac{eZ^* \rho D(J - J_{\mathrm{TH}})}{KT} \simeq 10^{-14}\mathrm{m}^2/\mathrm{s} \tag{58}$$

where $D_{150°C}$ = 5 × 10^{-17} m^2/s.

Thermomigration

Thermomigration is a forced-diffusion mechanism analogous to electromigration except that the driving force is a thermal gradient ∇T rather than a potential gradient ∇V. The mechanism affects mainly solder joints which are susceptible to it when the thermal gradient is $\geq 10^5$ °C/cm at temperatures $\geq \frac{2}{3}$ the melting temperature. Failure occurs by depletion of the diffusing element in the direction of the thermal gradient, usually at an interface where such an element cannot by replenished because of flux divergence. According to Adda and Phillibert (47) and Wever (48), one can express the mass transport rate per unit area q by

$$q = A\frac{Q^* DC\nabla T}{RT^2} \tag{59}$$

where

Q^* = Heat of transport
$D = D_0 \exp(-\Delta H/KT)$, diffusivity of diffusing species
C = Concentration of diffusing species
A = Unitless constant of proportionality

From Eq. (59), the time to failure based on critical mass depletion criterion q_c, electric resistance delta, is written as

$$t_{\mathrm{f}} = A'T^2 \nabla T^{-n} \exp(\Delta H/KT) \tag{60}$$

where $A' = q_c k/Q^* D_0$, a constant of the material and n = 1 to 3. In practice, it is found by extrapolation that the reciprocal time to failure is a linear function of ∇T with an intercept or threshold ∇T_{TH}. Therefore, Eq. (60) is written as

$$t_{\mathrm{f}} = A'T^2(\nabla T - \nabla T_{\mathrm{TH}})\exp(\Delta H/KT) \tag{61}$$

Empirical failure data show that plots of $1/t_f$ versus ∇T extrapolate to a threshold of 8.5°C/0.0008 m or ~10,000°C/m. The data also show that when the curves are forced through the origin, $1/t_f$ depends on ΔT^3, which is the form used for failure projections because it is more conservative and represents the statistical nature of the data better. The curves are similar to those obtained for electromigration (Fig. 18) which are plots of $1/t_f$ versus J. The thermomigration data were generated in a study of Pb–In and Pb–Sn solders at IBM for different ∇T, varying from about 50,000°C/m to 320,000°C/m. This is an accelerated test compared to the field where the gradients are below 50,000°C/m. For high-lead Pb–Sn solders, a thermal gradient of 150,000°C/m is of concern at operating temperatures below 100°C. However, low-melting solders such as the Pb–Sn eutectic, the Bi–Sn eutectic, and high-tin Sn–Sb and Pb–In alloys are more vulnerable than high-lead Pb–Sn solders.

What defines the heat of transport Q^*? According to Huntington (49), the main contributors to Q^* in a diffusion vacancy mechanism are the migration energy ΔH_m and the vacancy energy of formation ΔH_f, written as $Q^* = \beta\Delta H_m - \Delta H_f$. Beta indicates the part of the energy residing at the side of the jumping atom, and β for an FCC lattice is 0.8. For indium, ΔH_f = 43.5 kJ/mol, and the activation energy for self-diffusion ΔH_D = 77.3 kJ/mol. This yields $\Delta H_m = \Delta H_D - \Delta H_f$ = 33.8 kJ/mol, and the vacancy contribution to Q^* therefore is about −14.5 kJ/mol. Evidently, Q^* is a quantity af-

fected by many variables and must be determined for each material. In practice, Q^* becomes part of the proportionality constant, and the experimental variables are ΔT across the joint and average T. Gurp, deWaard, and Duchatenier (50) studied the thermomigration of indium and determined the heat of transport Q^*. They employed wide stripes with underlying molybdenum as the heater, causing voids in the metal film above the heater stripe at 120°C, which produced depletions and hillocks in few hours at temperature gradients on the order of 2×10^6 K/m and compressive stress exceeding the yield point of indium. The activation energy was 0.8 eV (77.3 kJ/mol) equal to that of lattice diffusion. The heat of transport Q^* was estimated at 11.6 kJ/mol. For the thermomigration of an interstitial solute in a linear temperature gradient, the stationary state ∇T produces a stationary opposing concentration gradient ∇C, leading to $J_i = 0$ at steady state. In view of irreversible thermodynamics (51,52), at steady state

$$\frac{d \ln C_i}{d(1/T)} = \frac{Q_i^*}{K} \tag{62}$$

In a nonstationary state, the initial and final ΔC conditions must be known accurately, and the flux must be in accordance with

$$J = -D\left(\nabla C - C\frac{Q^*}{KT^2}\nabla T\right) \tag{63}$$

Despite the fact that Soret (53) discovered the effect over a century ago, it still is not quite understood.

OXIDATION

Solders oxidize readily to form oxides of the most chemically active metal within the alloy. For high-lead Pb-Sn, the predominant oxide formed is SnO, which covers the entire surface of the alloy despite the low Sn concentration. At temperatures above 100°C, some PbO also forms. Tin oxide performs a useful function for solder bumps, namely, that of protecting the joint from corrosion by lead carbonate formation. On the other hand, its presence at any stage of the process hinders solder wettability and prevents the bumps from achieving a spherical shape. Lead carbonate also increases contact resistance for pad-on-pad or edge connectors thus influencing the thermal properties and package thermal management. The solder oxidation kinetics are useful for process reasons and for coping with electric and thermal requirements. For Bi–Sn eutectic films exposed to 85°C and 81% RH, the oxide showed a parabolic growth for which the thickness can be expressed as $h_{ox}(t) = 1.2 \times 10^{-9} t^{1/2}$ m, and the oxidation rate $dh_{ox}/dt = (6 \times 10^{*} - 10)t^{-1/2}$ m/s$^{1/2}$. At the early stage, the oxide film is not continuous because nucleation takes place in discrete sites on the surface of the solder. As the oxide film grows, the texture and distribution are altered with time, which helps in understanding the degradation of electric and thermal resistance at interfaces. This behavior has been observed in Pb–Sn and Bi–Sn eutectic alloys (14,15). The Pb–Sn eutectic (0.05 mm thick) was electroplated on Cu films, reflowed at 220°C and aged at 130°C. The SnO thickness was determined by Auger analysis, which is expressed as a function of time and temperature: $h_{ox}(t, T) = 1.8 \times 10^{-11} t \exp -21{,}000/R(1/T - 1/T_0)$, where $T_0 = 403$ K. In addition to open modules, "hermetic" modules were also tested to determine the oxide growth rate under restricted oxygen conditions. The oxide film thickness in a hermetic environment, expressed as a function of oxygen leak rate L, the rate R at which oxygen is consumed during oxidation, temperature T, and time t is written as

$$h_{ox}(L, C, T, t) = \frac{\dot{h}_{ox}(T,t)L}{aC + L}\left\{t - \frac{V}{aC + L}\left(1 - \exp - \frac{aC + L}{V}t\right)\right\} \tag{64}$$

where

$\dot{h}_{ox}(T, t) = d/dt[h_{ox}(T, t)]$, measured oxidation per second
a = Area of solder, m^2
$C = (\rho_s/\rho_g)h_{ox}(T, t) \exp -21{,}000/R(1/T - 1/T_o)$, m^3(O$_2$)/s · m^2, O_2 consumption rate
ρ_s = density of oxygen in oxide, kg/m^3
ρ_g = density of oxygen in atmosphere, kg/m^3
L = Oxygen leak rate, m^3(atm)/s
T = 403 K or 130°C
t = time, s

At 130°C, $C = 7.9 \times 10^{-9}$m^3/s · m^2. If the area $a = 0.000025$ m^2, the leak rate is 10^{-15}m^3/s, the volume $V = 0.000002$ m^3, and the time is 80,000 h (2.88×10^8/s), then the oxide thickness is 10.7 nm instead of 710 nm under nonhermetic conditions. Depth profiles leading to the determination of the oxide film thickness have been investigated by DiGiacomo (14), Sen, Sen, and Bauer (54), and Okamoto, Carter, and Hercules (55) using Auger analysis. On the other hand, ESCA was employed to determine the oxide stoichiometry, whereby one can determine the binding energy of an electron in atoms of a given element, which exhibit small variations in differing chemical combinations. The depth analyzed in this process at the sample surface is on the order of 1 nm. The technique is very useful in compound identification and is frequently used in studying solders and other metals surface (55–57).

RH WITHIN SEMIHERMETIC MODULES

Inside semihermetic or hermetic modules, the relative humidity is much lower at any time than in an open environment at the component's operating temperature. In fact, the relative humidity increases with time and as a function of the module leak rate. On this basis, the corrosion model expressed by Eq. (41) is given by

$$t_{50} = A\frac{1 + \beta C}{\beta C} e^{\frac{\Delta H}{KT}} \frac{[1 - H_o(1 - e^{-at})]^2}{H_o(1 - e^{-at})(1 + \sinh \mu V)} \tag{65}$$

in which the relative humidity H has been replaced by

$$H = H_o[1 - \exp(-at)] \tag{66}$$

where $a = L/V$ and H_o is the outside relative humidity. When $at \gg 1$, Eq. (59) reduces to Eq. (35), the open, nonhermetic form. If the module has a polymer seal, then the permeation rate of water vapor into the module is a function of the seal perimeter length l, height h, width w, and water permeability

P, yielding the parameter $a = lhPp/wV = L/V$, from which the leak rate $L = hlPp/w$. Normalization of results from hermetic and nonhermetic modules requires this analysis, based on permeation or leak rate.

CONCLUSIONS

In this chapter, we have described fundamental and important failure mechanisms which are of most concern to the microelectronics packaging industry, and we have tried to learn how those failure mechanisms can be expressed in generic forms by characterizing material parameters and processes and by generating more quantitative fatigue data through better controlled experiments. The approach has provided the empirical and theoretical background necessary for widening the scope of reliability models and reducing the need for experimentation. We have tried to achieve flexibility and to extend the models to new applications and to cope with different materials and designs. This leads to greater understanding of the mechanisms and to more reliable failure projections in the field.

BIBLIOGRAPHY

1. G. Di Giacomo, *Reliability of Electronic Packages and Semiconductor Devices,* New York: McGraw-Hill, 1996.
2. P. A. Tobias and D. Trindade, *Applied Reliability,* New York: Van Nostrand-Reinhold, 1986.
3. G. H. Hahn and S. S. Shapiro, *Statistical Models in Engineering,* New York: Wiley, 1967.
4. R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics,* 4th ed., New York: Macmillan, 1978.
5. D. C. Montgomery, *Introduction to Statistical Quality Control,* 2nd ed., New York: Wiley, 1991.
6. P. J. E. Forsyth, *Acta Metall.,* **11**: 703, 1963.
7. K. C. Norris and A. H. Landzberg, *IBM J. Res. Devel.,* **13** (3): 266–271, May 1969.
8. G. Di Giacomo and U. M. Ahmad, Thermal cycle frequency-threshold below which solder joints fatigue damage saturates as stress relax, *Proc. IMAPS,* 1997, San Diego, CA.
9. G. DiGiacomo, Thermal diffusion model describing transient strain in chip solder joints under thermal cycling, *Proc. Int. Symp. Microelectron.,* Reston, VA: SPIE **1847**, 1992, pp. 737–742.
10. R. Satoh et al., *IEEE Trans. Compon. Hybrids Manuf. Technol.,* **14**: 264–272, 1991.
11. K. U. Snowden, The effect of atmosphere on the fatigue of lead, *Acta Metall.,* **12**: 295–303, 1964.
12. R. Berriche et al., The effect of environment on fatigue of low tin lead-base alloy, *Electron. Packag. Corros. Microelectron.,* Metals Park, OH, 1987, pp. 169–174.
13. G. Di Giacomo, Modeling the effect of oxygen on the fatigue lifetime of solder joints, *Int. Reliability Phys. Proc.,* Las Vegas, 1995, pp. 93–96.
14. G. DiGiacomo, Oxidation kinetics of Pb-Sn eutectic solder and degradation of contact thermal resistance, *Proc. Int. Symp. Microelectronics,* Atlanta, 1986, pp. 322–327.
15. G. Di Giacomo et al., Oxidation kinetics of Sn-Bi eutectic and thermal contact resistance, *Proc. ISHM Symp.,* Minneapolis, 1987, pp. 587–593.
16. G. Di Giacomo, Effect of minicycles on the solder joint fatigue as a result of their superimposition on main cycles, *Proc. ISHM '96,* Minneapolis, 1996.
17. C. F. Felter, *Acta Mettall.,* **21**: 351, 1972.
18. T. K. Hong et al., Threshold stress for cyclic creep accelerated copper, *J. Mater. Sci.,* **23**: 1569–1572, 1972.
19. I. Finnie and W. R. Heller, *Creep of Engineering Materials,* New York: McGraw-Hill, 1959.
20. M. F. Ashby and R. A. Werrall, Diffusion-accommodated flow and superelasticity, *Acta Metall.,* **21**: 149–163, 1993.
21. J. Weertman, *Trans. Metal Soc. AIME,* **227**: 1475–1476, 1963.
22. G. Di Giacomo, Modeling solder joints cyclic creep, *Proc. Int. Electron. Packag. Conf.,* Atlanta, 1994, pp. 748–758.
23. S. Brunauer, P. H. Emmett, and E. Teller, *J. Amer. Chem. Soc.,* **105**: 309, 1983.
24. N. Lycoudes, The reliability of plastic microcircuits in moist environments, *Solid State Technol.,* **21** (10): 53–62, 1978.
25. D. S. Peck, Comprehensive model for humidity testing correlation, *IEEE Int. Reliability Phys. Symp.,* 1986, pp. 44–50.
26. R. P. Merrett, J. P. Bryant, and R. Studd, An appraisal of high temperature humidity stress tests for accelerating plastic encapsulated semiconductor components, *IEEE Int. Reliability Phys. Symp.,* 1983, pp. 73–82.
27. C. G. Sherley and C. E. C. Hong, Optimal acceleration of cyclic C.H.B. Tests for plastic packaged devices, *IEEE Int. Reliability Phys. Symp.,* 1991, pp. 12–22.
28. S. P. Sim and R. H. Lawson, The influence of plastic encapsulant and passivation layers on the corrosion of thin aluminum films subjected to humidity stress, *17th Proc. Reliability Phys. Symp.,* San Francisco, 1979, pp. 103–112.
29. H. L. A. Polman and H. Fokkens, Autoclave tests of 64K and 256K SRAMS, *Quality Reliability Eng. Int.,* **9**: 425–430, 1993.
30. E. S. Anolick, A. S. Amendola, and P. Levitz, Acceleration factors for environmental testing of integrated circuits, *9th Annu. Proc. Reliability Phys. Symp.,* New York, IEEE, 1971, pp. 228–232.
31. D. Y. Guan et al., New failure mechanism in temperature and humidity stress, *Proc. 36th Annu. Electron. Compon. Conf.,* pp. 107–110, 1986.
32. R. W. Lawson, The accelerated testing of plastic encapsulated semiconductor components, *IEEE Int. Reliability Phys. Symp.,* pp. 243–249, 1974.
33. S. M. Tam, Demonstrated plastic encapsulated microcircuit reliability for missile application, *Proc. 1994 Adv. Microelectron. Qualification/Reliability Workshop,* Aug. 16–18, 1994, pp. 116–130.
34. G. Di Giacomo, Reaction kinetics of Al films with phosphosilicate glass (PSG) in semiconductors, *19th Annu. Proc. Reliability Phys.,* Orlando, FL, 1981, IEEE Catalog No. 81 CH 1619-6.
35. J. Crank, *The Mathematics of Diffusion,* London: Oxford Univ. Press, 1956.
36. J. L. Barton and J. O. Brockis, The electrolytic growth of dendrites from ionic solutions, *J. Electrochem. Soc.,* **268A**: 485–505, 1962.
37. P. B. Price, D. A. Vermilyea, and M. B. Webb, On the growth and properties of electrolytic whiskers, *Acta Metall.,* **6**: 524–521, 1968.
38. A. W. Adamson, *Physical Chemistry of Surfaces,* New York: Interscience, 1970.
39. J. T. G. Overbeck, *Colloid and Surface Chemistry,* Cambridge, MA: Center for Advanced Engineering Studies, MIT, 1971.
40. H. B. Huntington and P. S. Ho, Electromigration of metals, *Proc. Int. Conf. Crystal Lattice Defects, Conf. J. Phys. Soc. Jpn.,* **18**, Supp. II:202–208, 1963.
41. H. B. Huntington and A. R. Grone, *J. Phys. Chem. Solids,* **20**: 76, 1961.

42. I. A. Blech and C. Herring, Stress generation by electromigration, *Appl. Phys. Lett.,* **29** (3): 1976.
43. I. A. Blech, *J. Appl. Phys.,* **47**: 1303, 1976.
44. G. Di Giacomo, Electromigration depletions in Pb-Sn films, *Proc. Reliability Phys. Symp.,* San Francisco, 1979, pp. 72–76.
45. R. Vanselow, R. Masters, and R. Wehness, Crystal forms of hillocks and voids formed by electromigration on ultrapure gold and silver wires, *J. Appl. Phys.,* **12**: 341–345, 1967.
46. H. B. Huntington, Effect of driving force on atom motion, *Thin Solid Films,* **25**: 265–280, 1975.
47. Y. Adda and J. Phillibert, *La Diffusion Dans Le Solider,* Paris: Presses Univ. de France, 1966, chap. 15.
48. H. Wever, *Electro-und Thermotransport in Metallen,* Leipzig: Barth.
49. H. B. Huntington, in H. Aaronson, ed., *Diffusion,* Metal Park, Ohio: ASM, p. 155.
50. G. J. Van Gurp, P. J. deWaard, and F. J. Duchatenier, Thermomigration in indium and indium alloy films, *J. Appl. Phys.,* **58** (2): 728–735, 1985.
51. H. I. Yoo and B. J. Wuensch, *J. Phys. Chem. Solids,* **54** (5): 613–620, 1993.
52. J. H. Matlock and J. P. Start, *Acta Metall.,* **19**: 923, 1971.
53. C. Soret, *Arch. Sci. Phys. Nat.,* 348, 1879.
54. K. S. Sen, S. Sen, and C. L. Bauer, Determination of the oxidation states of tin by auger electron spectroscopy, *Thin Solid Films,* **82**: 157–164, 1981.
55. Y. Okamoto, J. W. Carter, and D. M. Hercules, A study of the interaction of Pb-Sn solder with O_2, H_2O, NO_2 by ESCA and AES, *Appl. Spectrosc.,* **33** (3): 1970.
56. R. J. Bird, Corrosion-resistant lead-indium alloys: Surface studies by photoelectron spectroscopy (ESCA), *Metal Sci. J.,* **7**: 109–113, 1973.
57. C. L. Lau and G. K. Wertheim, Oxidation of tin: An ESCA study, *J. Vac. Sci. Technol.,* **15** (2): 1978.

GIULIO DI GIACOMO
DONALD W. SCHEIDER
JASVAR S. JASPAL
International Business Machines

DESIGN, HUMAN CENTERED. See HUMAN CENTERED DESIGN.