Lifetime Modeling and Prediction of Power Devices

Mauro Ciappa,
Integrated Systems Laboratory, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland

Abstract

Accurate lifetime prediction based on realistic mission profiles represents a challenge for the design of complex devices and systems. This paper proposes several examples where this problem has been solved by proper analysis, by dedicated physical modeling, and by efficient calculation tools.

1 Introduction

Thanks to their ability in handling large currents at high voltage and at high switching frequency, Insulated Gate Bipolar Transistors (IGBT) have almost completely replaced bipolar power transistors and they are challenging the position of Gate Turn-Off Thyristors (GTO) in their traditional fields of application. In the last years, the need to increase the reliability of high-power IGBT multichip modules has been one of the most powerful drivers that forced engineers to design new products, especially intended for traction (railway and automotive), for power transmission, and for power distribution applications. In the case of railway and automotive traction, the main requirements imposed by the increasing complexity of the systems are weight and space reduction, as well devices operating at high temperature and at increasing voltages. This also applies to other application fields, where this trend is leading more and more to the deployment of advanced integrated technologies. The daily experience in dealing with such complex devices has clearly shown the traditional reliability assurance programs based on the a posteriori failure rate assessment cannot be used. In fact, even a simple measurement of the failure rate as function of time would require many millions of cumulated component-hours. Instead of this, modern power devices are designed according to built-in reliability programs taken from the manufacturing of ULSI devices. At present, new approaches are in use, which are based on the knowledge of the root cause of the failure mechanisms and on the relationship among the product specifications, its constituting elements, the variations in the manufacturing process, as well as the interactions of product materials with the loads and their impact on the product reliability. Recently, efficient practices have been established, which require that the development of lifetime models is done in parallel (or even prior) to the development phase, and that these model are calibrated/validated by targeted experimental procedures. The final target of these activities is to confirm that a given product is reliable and that the process is suitable for mass production. The best practices that are presently in use are a systematic combination of physical models for the dominating failure mechanisms with the proper statistical and stochastic procedures. This paper presents a concise review of the necessary tools, including, the main definitions, the usual lifetime metrology tools, quantitative models for the most common failure mechanisms observed in the field, and the methodology for the calculation of the lifetime of power devices based on mission profiles.

2 Definitions

2.1 Lifetime vs. Failure Rate

The lifetime t of a non-repairable device is a random variable, which is defined as the time span between the initial operation and failure, where the failure condition is clearly identified by a failure criterion. Estimates either of the median (t_{50}), or the empirical mean ($E(\tau)$) of the mechanism-specific distribution are measured during lifetime tests (very often under accelerated conditions). For engineering purposes even a more important role is played by the time-dependent failure rate $\lambda(t)$, which is extracted from the previous distribution, and that is defined as the probability that a device will fail in the next operation hour, provided that it did not fail before. The correct measurement unit for $\lambda(t)$ is the FIT (1 failure in 10^9 operation hours) and not the part per million, as usually assumed [1]. This value is very important for the designer, since it univocally expresses the number of failures, which can be expected after a given operation time of the devices, and given a specified availability of a repairable system, it is used to design efficient preventive maintenance strategies.

2.2 Complexity

System topologies have to be taken into consideration, since they can produce either a gain, or a loss in the measured lifetime, due to elements that are connected either in parallel (redundancy), or in series in the sense of the reliability block diagram. Typical examples at device level are the parallel-connected emitter and the series-connected gate bond wires in power modules [2].

2.3 Mission profiles: some examples

The mission profile of a device (or system) is the specific task, which must be fulfilled during a stated time and under specified conditions.
A first example of particularly challenging mission profiles are power converters used in airborne applications to drive electrical actuators in the close vicinity of the jet engine. In this case, the power devices are required to survive at the same time static environment temperatures of about 200°C during the flight and extreme thermal cycles between –55°C and 200°C with a rate of 10°C/min during landing and take off. The requested lifetime of the equipment is 50'000 hours corresponding to about 500 landing/take off cycles [3]. Power devices used in electrical hybrid vehicles are normally specified for a typical lifetime of 10'000 hours, which corresponds to several millions of power cycles with different duration and amplitude. In this case, the environment temperature strongly depends on the design of the cooling circuitry of the vehicle. If the power converter shares the same cooling circuitry with the thermal engine, a baseline temperature of 90°C up to 120°C can be expected. This leads to junction temperatures of the switching devices in the 200°C range. With a typical baseline temperature of 60°C, the working conditions of the devices are less severe in vehicles, which use a dedicated cooling circuitry. Up to now no standardized mission profile is available in the automotive field, which has been expressly developed for reliability prediction purposes. Car manufacturers are assessing the reliability performance of their products by using standardized driving cycles, which are originally intended for the measurement of exhaust emissions and of the fuel consumption (e.g. NMVEG, Artemis). Finally, basing on the past experience with GTO and on logistic considerations (no preventive maintenance for semiconductors devices) several railway operators usually express two main requirements. The first refers to the useful life of a device that is specified in at least 30 years. The second is the failure rate of a single module, which is specified not to exceed 100 FIT over the whole useful life of the device.

3 Failure mechanisms

Like in the case of integrated circuits (IC), the failure mechanisms, which affect power device modules can be subdivided in two categories. The first includes the extrinsic mechanisms, which result from poorly controlled or poorly designed manufacturing processes. The second category includes the intrinsic failure mechanisms, which lead to a time-dependent degradation of the performances of the device during its useful lifetime. It is important to notice that in opposite to IC, the lifetime of power modules is usually limited by intrinsic (wear out) mechanisms, since devices and materials are often operated close to their physical limits. Therefore, one among the main tasks during a prototyping phase is to classify the observed failure mechanisms, in order to define appropriate corrective actions for the extrinsic mechanisms, and to develop quantitative models for the intrinsic mechanisms, with the scope with the scope to design intrinsically reliable devices.

3.1 Package-related mechanisms

Multichip modules for high power IGBT devices are complex multilayered structures consisting of different materials, which have to provide a good mechanical stability, good electrical insulation properties, and good thermal conduction properties. The failure mechanisms that are most frequently observed in power modules affect these capabilities and are due to the thermo-mechanical low fatigue stress of the package materials as a consequence of the thermal cycles experienced during the operation. The main driving forces of these mechanisms are the mismatch in the coefficient of thermal expansion (CTE) of the different materials, the characteristic length of the layers, and the local temperature swing they are subjected.
Bond wire fatigue. Multichip IGBT modules for high-power applications typically include up to 800 wedge bonds, which are connected by ultrasonic wedge bonding either onto the aluminum metalization (with a thickness ranging from 3 to 5 μm), or onto a strain buffer. Since about half of them are bonded onto the active area of semiconductor devices (IGBT and freewheeling diodes), they are exposed to almost the full temperature swing imposed both by the power dissipation in the silicon and by the ohmic self-heating of the wire itself. Emitter bond wires are usually 300 up to 500 micrometers in diameter. The chemical composition of the wire can be different from manufacturer to manufacturer, however in all cases, the pure aluminum is hardened by adding some few thousand parts per million of alloying elements, such as silicon and magnesium, or nickel for corrosion control. Under normal operating conditions, the current within a single aluminum bond wire does not exceed 10 A, such that the maximum ohmic power dissipation is between 100 and 400 mW, depending on the wire diameter. Failure of a wire bond occurs pre-

dominantly as a result of fatigue caused either by shear stresses generated between the bond pad and the wire, or by repeated flexure of the wire. The fracture mechanics at bonded interfaces and the modeling of the crack propagation within the welded joint with time is a quite complex issue. There is experimental evidence that the crack leading to the failure is initiated at the tail of the bond wire, and propagates within the wire material until the bond wire completely lifts off. The failure of a single or of multiple bond wires causes a change either in the contact resistance or in the internal distribution of the current, such that it can be traced by monitoring V_{CEsat} [4]. The observed failure mode can be different depending on the stress the devices are submitted. If the test is not interrupted after exceeding a predefined threshold, the end of life failure mode observed during power cycles is melting of the survivors bond wires. On the contrary, during high-voltage test or field operation, a frequently observed secondary failure mechanism is the triggering of parasitics.

Bond wire heel cracking. Bond wire heel cracking rarely occurs in advanced power modules. However, it can be observed mainly after long endurance tests and especially in cases where the ultrasonic bonding process is not optimized. The failure mechanism is due again to a thermo mechanical effect. In fact, when the wire is subjected to temperature cycles it expands and it contracts undergoing flexure fatigue. In the case of a typical bond wire length of 1 cm and of a temperature swing of 50°C, the displacement at the top of the loop can be in the 10 µm range producing a change in the bending angle at the heel of about 0.05°. An additional stress is introduced by the fast displacement of the bond wire (e.g. at the turn on) within the silicone gel, which can be considered as a very viscous fluid. In those cases, where the temperature change within the bond wire is dominated by the ohmic self-heating, heel cracking can also be observed at the wire terminations welded on the copper lines of both IGBT chips and freewheeling diodes.

Aluminum reconstruction. Although reconstruction of the aluminum metalization is an effect, which has been encountered since the early times of microelectronics, the occurrence of this degradation mechanism in IGBT multichip modules has been firstly reported in [5]. During thermal cycling of IGBT devices and of freewheeling diodes, periodical compressive and tensile stresses are introduced in the thin metalization film by the different CTEs of the aluminum and of the silicon chip. Due to the large thermo mechanical mismatch between both materials and due to the stiff ness of the silicon substrate, the stresses, which arise within the aluminum thin film during pulsed operation of the device can be far beyond the elastic limit. Under these circumstances, the stress relaxation can occur by diffusion creep, grain boundary sliding, or by plastic deformation through dislocation glide, depending on temperature and stress conditions. In the case of IGBT devices, the strain rate of the metalization is controlled by the rate of temperature change. Because the typical time constants for thermal transients in IGBT are in the range of the hundreds of milliseconds, if the devices are operated cyclically at maximum junction temperatures above 110°C, the stress relaxation occurs mainly by plastic deformation at the grain boundaries. Depending on the texture of the metalization, this leads either to the extrusion of the aluminum grains or to cavitation effects at the grain boundaries. In any case, aluminum reconstruction reduces the effective cross-section of the metalization and results into an increase of the sheet resistance of the aluminum layer with time. This effect contributes to the observed linear increase of V_{CE} as function of the number of cycles during power cycling tests. Aluminum reconstruction may become a reliability hazard in presence of pre-existing step coverage problems at the emitter contact vias. In this case, thermo mechanical and electromigration effects can coalesce resulting into a complete depletion of the metalization from the wall of the via.

Solder fatigue and solder voids. A main failure mechanism of power modules is associated with the thermo mechanical fatigue of the solder alloy layers. The most critical interface is represented by the solder between the ceramic substrate and the base plate, especially in the case of copper base plates [6]. In fact, at this location one finds the worst mismatch in the CTEs, the maximum temperature swing combined with the largest lateral dimensions. Nevertheless, fatigue phenomena occurring in the solder between the silicon chip and ceramic substrate cannot be neglected. This is also the case of process-induced voids, which can both interact with the thermal flow and with the crack initiation within the solder layer. Both gross voids and extended fatigue-induced cracks can have detrimental effects on dissipating devices. In fact, they can significantly increase the peak junction temperature of an IGBT or of a diode and therefore accelerate the evolution of several failure mechanisms including bond wire lift off and solder fatigue.

3.2 Burn out failures

Device burnout is a failure mode, which is very frequently observed either as the final act of wear out, or as consequence of a robustness issue. Burnout is often associated with a short circuit condition, where a large current flows through the device (or through a portion of it), while it is supporting the full line voltage. Sustaining a short circuit over a long time interval inevitably leads to thermal runaway and finally to a fast destruction of the device. In fact, since IGBTs do not require any *dI/dt* snubbering, the device itself limits the current increase rate. Therefore, after the failure the current may increase at a rate up to 10kA/µs, leading to a current maximum in the 100 kA range and to decay within 100µs. In this case, the main part of the stored capacitive energy is released in few hundreds of nanoseconds reaching a peak power up to 100 MW. The capacitive energy is dissipated by the ohmic components of the circuit, i.e. mainly by the bond wires and by the silicon chip. As consequence of

the adiabatic heating process, the bond wires evaporate, by producing a preferential conductive path for arching through the module. The resulting shock wave rapidly propagates through the silicon gel by leading to the catastrophic destruction of the device. Advanced IGBT multichip modules have been expressly designed for minimizing the consequences of such an explosion in order to match the tight requirements in terms of safety imposed by traction applications. They are many systems, environmental and wear out related causes, which may turn into a short circuit condition. Among these there are operation of the device outside the safe operating area, gate unit malfunction, inhomogeneous current sharing [7], overheating due to the degradation of the thermal resistance, dielectric breakdown, and cosmic ray irradiation. Since the investigation of the root causes associated with system design and device application related problems are outside the scope of this paper, just the case of the latch up is briefly discussed here since it is an inherent failure mechanism to IGBTs, which plays an important role in determining the availability of a power system. Nevertheless, it has to be noted that latch up is mainly a problem related to the ability of a certain device design to survive stresses out-of-specification. Thus, strictly speaking, it is a robustness issue rather than a reliability concern. The latch up mechanism (static and dynamic) manifests itself through a sudden collapse of the collector to emitter voltage, and once this failure mechanism is activated the device cannot be longer controlled through the gate. The failure mode associated with latch up is always a generalized low-ohmic short circuit of collector, emitter, and base. Catastrophic burnout of IGBT devices can also be initiated through local self-sustaining filamentary discharges produced in the silicon by recoil nuclei, which result either from neutron scattering, or from the decay of neutron-activated isotopes within the semiconductor. At normal operating conditions, high-energy neutrons are usually associated to terrestrial cosmic rays.

4 Lifetime Metrology

In order to estimate the failure rate from experimental data (either from field data, or from accelerated lifetime tests), the empiric distribution has to be computed on the base of the observed failure-free-times t_i. Assumed that n failures with the related failure-free-times $t_1, \ldots, t_n$ have been observed, the first step is to rank these values such that $t(1) \leq t(2) \leq \ldots \leq t(n)$. In the next phase the empiric distribution is built as a step function, i.e. $F_{emp}(t)$: $t(i) - F_{emp}(t(i)) = i/n$. In the following, the empiric distribution has to be associated to an analytical distribution in order to extract the distribution parameters. This can be done by fitting the empiric distribution by the analytical distribution to be tested according to the maximum likelihood criterion. The goodness of the fit can be tested in different ways, i.e. either by using the Kolmogorov-Sminorv, or the Chi-square (χ^2) test [8]. Once the more accurate approximating distribution is identified, the related parameters can be extracted to compute the failure rate $\lambda(t)$. Usual parameters, which can also be determined from such a plot is the time required to reach a given percentile of the distribution, as for example the median (t_{50}) that represents the time for getting 50% of the device population failed. Nowadays, this task is performed automatically by dedicated software tools, or by graphic techniques, where the empiric data appear as a straight line in dedicated probability charts, associated to each distribution.

In this kind of analysis different distributions are used, depending on the nature of the degradation mechanism, which is observed. Therefore, it is mandatory to perform failure analysis of the failed devices to define the physical mechanism that produced the failure, in order to sort the different failures mechanisms to be modeled.

4.1 Exponential distribution

The expontial distribution $F(t) = 1 - exp(-\lambda t)$, is typically used if a constant failure rate λ is expected to occur over the whole operating time of the device, i.e. when the stochastic process is memory-less. This means that the failure rate of a device does not depend on how long it has been operated in the past. From a modeling point of view it represents reasonably the failure regime in the flat region of the bathtub, which is characterized by random failures. This is a realistic model, in the case of mature devices, which have been properly screened to eliminate infant failures, and that did not yet reach the end-of-life region. In power modules, the exponential distribution can be used to model semiconductor-related failure mechanisms, only. Under circumstances it can also be used to model robustness-related issues like cosmic rays.

While wear out failure mechanisms can be attacked by adequate design rules, random failures are not necessarily related to a given failure mechanism. In fact, they express the random character of both the occurrence of physical processes (failure mechanisms), and of the variation of manufacturing processes. Random failures play a relevant role in defining the survival probability of (non-reparable) mature systems with a very long operating life, as it can be the case for inverters in railway traction applications. For instance, the survival probability after 30 years operation of a unit consisting of 6 modules, each having a constant failure rate of 100 FIT is close to 0,85. In converse, for a constant failure rate of each module of 400 FIT, the survival probability is in the 0,5 range. This simple, but realistic, example evidences the requirement of an accurate estimate of λ with sufficient statistical significance. For this scope, it is recommended to use a symmetric interval estimate for the constant failure rate, which is defined by a lower (λ_l)

and upper (λ_u) limit, in conjunction with a confidence level β (e.g. $\beta = 0.1$, *conf. level [%] = 100 (1 - 2β) = 80%*). It can be shown by the use of the Poisson statistics [8], that in the case of k failures observed during the cumulative operating time T, the interval limits are given by

$$\lambda_l = \frac{\chi^2(2k,\beta)}{2T} \quad , \quad \lambda_u = \frac{\chi^2\left(2(k+1),1-\beta\right)}{2T}$$

where $\chi^2(x,y)$ is the Chi-square distribution with the arguments x and y that can be easily obtained from tabulated values. If no failure is observed during the cumulative operating time T, the interval limits are given by

$$\lambda_l = 0 \quad and \quad \lambda_u = \frac{\ln\left(1/\beta\right)}{T}$$

When the devices are subjected to wear out, the exponential model assuming a constant failure rate can be a momentary approximation of a time-dependent failure rate, only. Therefore, the use of the mean time to failure $MTTF = 1/\lambda$ for the calculation of the lifetime of an item can lead to wrong results. For example, a momentary failure rate of 10 FIT does not necessarily correspond to a lifetime of 11415 years, since in the meanwhile wear out could occur, which terminates the device. Finally, it should be noted that in the case of the exponential distribution, about 63% of the original devices are already failed at $t = MTTF$.

4.2 Weibull distribution

There is enough experimental evidence that the failure-free times measured associated to wear out mechanisms can be modeled (at least piecewise) by a Weibull distribution of the form

$$F(t) = 1 - \exp\left(-(\lambda t)^{\beta}\right)$$

where λ is the scale and β is the shape factor, respectively. Therefore, a Weibull distribution exhibiting a failure rate that increases in time, is the most suited distribution to model the wear out (package-related) failure mechanisms observed in power modules. On the contrary, semiconductor-related failure mechanisms are better described by the constant failure rate of the exponential distribution. Furthermore, when repeating a power cycling experiment at a constant ΔT and average temperature, the observed number of cycles to the failure due to thermo mechanically activated failure mechanisms (e.g. bond wire lift off, substrate delamination) is also Weibull distributed.
The related time-dependent failure rate associated to a Weibull distribution is given by

$$\lambda(t) = \beta\lambda(\lambda t)^{\beta-1}$$

and it is monotonically increasing if $\beta > 1$.
In the case of non-reparable items affected by a failure rate increasing with time, if the use of preventive maintenance is not an option, the task of the reliability engineer is to design the device with a failure rate during the operating lifetime that does not exceed a pre-defined threshold, which can be technically tolerated. This requires a number of dedicated accelerated tests to ascertain the dependency of the failure rate on the different technology parameters.
In many cases, the experimental data cannot be described just by a single distribution, but a combination of two, or more distributions is required. This can be the case of the co-existence of two competing failure mechanisms (e.g. bond wire lift off and substrate delamination during power cycles), which can lead to the termination of the device. If the related Weibull parameters are λ_1, β_1 and λ_2, β_2, respectively, it can be demonstrated [1] that the resulting failure rate is given by

$$\lambda(t) = \beta_1\lambda_1(\lambda_1 t)^{\beta_1-1} + \beta_2\lambda_2(\lambda_2 t)^{\beta_2-1}$$

When investigating the field reliability of systems using non-repairable devices exhibiting a lifetime, which is much shorter than the operating time of the systems (MTBF), it is usual to compute the mean time between failures by the means of a running average algorithm. In many cases, the MTBF value delivered in time by such a procedure converges towards a constant value. It has to be stressed, that even if the obtained MTBF is constant in time, it does not mean necessarily that the failure rate of the devices is also constant in time, i.e. the distribution of the failure-free time of the devices is not necessarily exponential, and the failure mechanisms is not necessarily random. If failure analysis confirms that the devices are failing due a wear out mechanism, whose failure-free time is Weibull distributed, it can be easily demonstrated, that the asymptotic limit obtained trough the running average algorithm is

$$E(\tau) = \frac{\Gamma\left(1+1/\beta\right)}{\lambda}$$

where Γ is the gamma function, and $E(\tau)$ is the mean of the associated Weibull distribution.

5. Lifetime Modeling and Design

The end-of-life period of complex multi-chip modules is often defined by thermo-mechanics related failure mechanisms. The lifetime for these wear-out mechanisms is normally estimated on the base of deterministic models, which are calibrated with data extracted from accelerated power cycling experiments. Furthermore, these estimates are referred to a given application profile, which is specific for the technical system under consideration. Simple lifetime predic-

tion models, based on the bimetallic approximation for the thermo-mechanical stresses, and based on principle of the linear accumulation of the damage related to low cyclic fatigue are in use since years [9]. The main issue about these models is the procedure for defining and extracting the number, the amplitude and the duration of the thermal cycles, which are experienced by the system when it is submitted to a given mission profile. Alternative prediction procedures base on some fundamental equations of the thermo-mechanics and takes into consideration the creep suffered by compliant materials when they are submitted to thermal cycles. Due to the complexity of the systems to be considered, strongly simplified models are used. Nevertheless, the experience shows that predictions based on more complex models (e.g. realistic finite element modeling) are not necessarily more accurate than those obtained by simple physics-based models. The reason often resides in the lack of experimental data to extract the large amount of parameters required by complex models. In particular, finite element simulation low fatigue phenomena are strongly limited by the fact that relevant parameters as the initial stress distribution cannot be conveniently measured, and the long computational time required by three-dimensional finite element modeling just make possible to simulate the damage evolution during some few cycles. On the contrary, realistic finite element modeling can be very helpful to estimate the stress (and strain) levels, to be used either in simple analytical expressions, or to design the system such that possibly the stress levels are kept below the threshold of fatigue limit.

5.1 Low-cycle fatigue mechanisms

During thermal cycling of power devices the thermo mechanical stress is often high enough to produce plastic deformation of the materials. This is the case of almost all package-related failure mechanisms proposed in Section 3.1. Under these circumstances the number of cycles to the failure N for a constant ΔT, can be conveniently described by the maximum strain experienced in conjunction with the Coffin-Manson relation. By using for sake of illustration the simple example of a bimetallic system consisting of joint between copper and alumina (coefficients of thermal expansion α_{Cu} and α_{Alu}, resp.), with a typical length L, and submitted to a temperature swing ΔT, we obtain an estimate of the total strain

$$\varepsilon_{tot} \approx L\left(\alpha_{Cu} - \alpha_{Alu}\right)\Delta T$$

Considering the large size of the copper and ceramic plates, and in particular taking into account that the strain is of the same order of magnitude as the thickness of the solder layer joining both plates yields

$$\varepsilon_{tot} = \varepsilon_{elastic} + \varepsilon_{plastic} \approx \varepsilon_{plastic}$$

that inserted in the Coffin-Manson delivers

$$N_f \approx \varepsilon_{plastic}^{-n}$$

where N_f is the number of thermal cycles to the failures associated with the quantile f of the distribution, and n is the exponent depends on the materials. Combining both previous expression yields

$$N_f \approx a\left(\Delta T\right)^{-n}$$

where a is a proportionality constant, which has to be determined experimentally in conjunction with the exponent n.

This kind of model is used to predict N_f by assuming different (electrical or mechanical) failure criteria for the different thermo mechanical failure mechanisms. Among these, the bond wire lift off in modules, the delamination, and the bond wire heel cracking.

For the bond wire lift off, it is common to use the model in its easiest form [6], i.e.

$$N_f = a\left(\Delta T_j\right)^{-n}$$

where ΔT_j is the temperature swing of the junction temperature, and the parameters a and n, are usually provided by the module manufacturer.

Similarly, a Coffin-Manson-like model can be used to predict the number of cycles to failure, due to the delamination at critical interfaces (e.g. DCB-base plate, chip-DCB), according to

$$N_f = 0{,}5\left(\frac{L\,\Delta\alpha\,\Delta T_{sub}}{\gamma\, x}\right)^{1/C}$$

where L is the typical size of the system to be considered, $\Delta\alpha$ the thermo mechanical mismatch of both plates, ΔT_{sub} the swing of the temperature cycle at the interface (e.g. DCB-base plate, chip-DCB), x and γ are the thickness and the ductility factor of the solder layer, respectively, C the exponent to by fitted on experimental data.
Finally, the Schafft model [6] predicts the number of cycles to failure, due to heel cracking of the wire bonds, based on

$$N_f = A\left(\frac{r}{\rho_0}\left(\frac{ar\cos\left(\cos\psi_0\left(1-\Delta\alpha\,\Delta T\right)\right)}{\psi_0} - 1\right)\right)^n$$

where ΔT is the swing of the temperature cycle, A and n fit parameters, $\Delta\alpha$ the thermo mechanical mismatch between aluminum and DCB material, r the radius of

the bond wire, ρ_0 the bending radius at the heel, and ψ_0 the angle between chip plane and bond wire.

These simple models do not take into consideration the fact that N_f has been shown to depend on the temperature swing ΔT of the thermal cycle, but also on the maximum temperature T_{max}, which is reached during the thermal cycle. In particular, a correction is necessary, if T_{max} approaches the melting temperature (T_{MP}) of the materials, as it is often the case of low-temperature solder alloys. Usually, this dependency is represented heuristically as a correction factor, multiplying the traditional Coffin-Manson expression (which depends on ΔT, only). An example [6] of such a correction is

$$N_f\left(\Delta T, T_{\max}\right) = N_f\left(\Delta T\right) \quad \frac{1}{\exp\left(m\left(T_{\max} - T_{MP}\right)\right)+1}$$

where m is a fitting parameter to be defined experimentally. Alternative expressions are based on the average cycle temperature (T_{aver}) and on an Arrhenius-like term

$$N_f\left(\Delta T, T_{aver}\right) = N_f\left(\Delta T\right) \quad \exp\left(\frac{C}{k\,T_{aver}}\right)$$

where C is again a fitting factor.
Additional relevant parameters for which theoretical models exist are the cycle frequency, cycle dwell-time, and rise- fall-time of the pulse. However, calibration data have never been published for power modules.

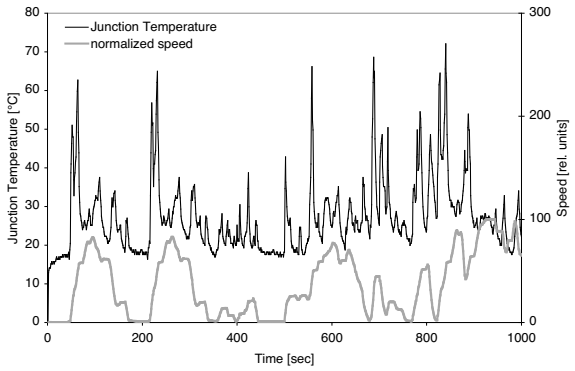

Fig. 1 Time evolution of the junction temperature of an IGBT of a hybrid car operated according to the driving cycle Artemis [6].

5.2 Linear Accumulation of the damage

Simple lifetime prediction models, based on the bimetallic approximation for the thermo-mechanical stresses, and based on the principle of a linear accumulation of the damage related to cyclic fatigue have been proposed and investigated in the past. Under the simple assumption of linear accumulation of the cyclic fatigue damage, a fatigue function can be defined as

$$Q\left(\Delta T\right) = \frac{N\left(\Delta T\right)}{N_f\left(\Delta T\right)}$$

where is the number of cycles performed at a given temperature excursion ΔT and N_f is the number of cycles to failure at the same ΔT, respectively. Differentiation and integration of Q [2] over a complete mission temperature profile yields

$$Q\left(after\ 1\ mission\ profile\right) = \frac{1}{a}\int_{\Delta T\,\min}^{\Delta T\,\max} \frac{g\left(\Delta T\right)}{\Delta T^{-q}}\, d\left(\Delta T\right)$$

where $g(\Delta T)$ represents the frequency distribution of the thermal excursions inside one mission profile. Equation (1) already includes the assumed dependency of N_f on ΔT according to a Coffin-Manson law with coefficient a and exponent q, to be determined experimentally by accelerated tests. The ratio $1/Q$ returns the lifetime of the system under investigation expressed in number of missions profile to failure.

5.3 Analysis of the mission profile

The main issue about these models is the procedure for defining and extracting $g(\Delta T)$, i.e. the number, the amplitude and the duration of the thermal cycles, which are experienced by the system when it is submitted to a given mission profile. The usual cycle counting procedure used in thermo mechanics is the rain flow method. However, alternative techniques have been introduced [2], which better fit the requirements imposed by the power traction applications. The choice of the correct cycle counting algorithm is very important. In fact, as been shown in [2] the estimated lifetime may spread by more than one decade, depending on the assumed criterion. Particular attention has to be paid to filter out those spurious cycles, which are introduced either by the limited resolution of the equipment, or by the sampling procedure. The extracted distribution can be either analytical, or numerical. The analytical form can be obtained by fitting the empirical distribution with the usual statistical distributions, i.e., Normal, Lognormal, or Weibull.
As an example, Fig. 1 represents the evolution of the junction temperature of an insulated gate bipolar transistor (IGBT) device in a power module, operated within a hybrid car, which is driven according to the standard mission profile Artemis [9] in an urban environment. Fig. 2 shows the associated distribution, as a function of the cycle amplitude and of the cycle duration. It can be observed that the highest bin corresponds to thermal cycles having a temperature excursion lower than *5 K* and duration shorter than *5 s*. About 80% of the temperature cycles are lower than *25 K*, while their duration does not exceed *15 s*.

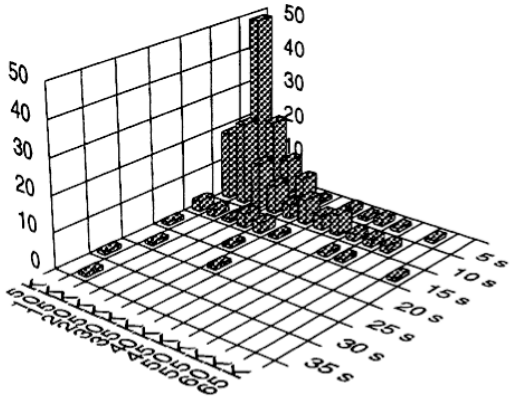

Fig. 2 Distribution of the temperature cycles from Fig. 1 as a function of the cycle excursion and cycle duration (cycle definition see [2]).

The observed maximum temperature swing is *65 K*, the maximum cycle duration is *35 s*, and the total number of cycles for a driving period of *1000 s* is *206*. This means that with a specified lifetime of *8'000 hours* the vehicle experiences up to *5 millions* of cycles with different amplitude and duration.

5.4 Thermo mechanical model

In order to overcome the problems involved with the arbitrary definition of the thermal cycle, prediction models have been developed, which make use of the fundamental thermo mechanical equations describing the creep mechanics in materials submitted to cyclic loads. In particular, following simplified model assumes that stresses above the yield point cause a plastic irreversible deformation, which turns into a hysteresis loop in the stress-strain representation. The uniaxial strain is computed from the thermal expansion of the bimetallic system, while the corresponding shear stress is derived from the creep relation. Once represented in the strain-stress plane, these values provide the deformation energy dissipated during a thermal cycle of arbitrary shape. In fact, it can be easily demonstrated that the deformation energy is proportional to the area included within the corresponding hysteresis loop. The lifetime of the system can be computed under the assumption that the system reaches its end-of-life as soon as a total amount of deformation work (W_{tot}) has been accumulated. Thus, the lifetime of the system (expressed in number of mission profiles) can be computed by calculating the ratio between W_{tot} and the deformation work associated to a given mission profile. Starting from these principles, two different models for the creep behavior of solder materials can be derived. The first model includes only the dependency of the shear stress on the temperature swing and can be summarized by the following equations

$$\gamma(i) = \left(T(i) - T_0\right) \cdot \Delta\alpha$$

$$\tau(i) = \mathrm{sgn}(\gamma(i) - \gamma(i-1)) \frac{G}{A^{1/n}} \cdot \left(\frac{abs(\gamma(i) - \gamma(i-1))}{t_{slot}} \right)^{1/n}$$

where γ is the shear strain, τ is the shear stress, $\Delta\alpha$ is the mismatch in the thermal expansion coefficients of the layers, which are joined by the solder and t_{slot} is the time discretizing unit. T is the instantaneous temperature and T_0 is the temperature, at which the strain is zero. Finally, G, A, and n are parameters that depend on the materials and on the geometry of the system.

The second model takes into account the dependency of the creep behavior both on the temperature swing and on the instantaneous temperature. This dependency is represented by an Arrhenius term, yielding

$$\tau(i) = \mathrm{sgn}(\gamma(i) - \gamma(i-1)) \frac{G}{A^{1/n}} \cdot \left(\frac{abs(\gamma(i) - \gamma(i-1))}{t_{slot}} \exp\left(\frac{E_a}{k\,T(i)} \right) \right)^{1/n}$$

where E_a is the activation energy and k is the Boltzmann constant. All equations are already expressed in their discretized form to allow the straightforward numerical computation. Special attention must be paid in both models to the temperature vector *T(i)*, as it represents the solder (not the junction) temperature evolution. Therefore, before evaluating *T(i),* the junction temperature mission profile (Fig. 1) should be somehow filtered to estimate the thermal effect of both the package and the heat sink. Accurate procedures have been developed [10] to calculate efficiently the temperature evolution at a specific interface, by taking also in account the local variation of the rise- and fall-time of the thermal cycles.

Once calibrated [2], the model applied to the mission profile of Fig. 1 returns the instantaneous strain-stress values, which are plotted with the time as a parameter in Fig. 3. The area enclosed within every single loop represents the total plastic deformation work associated with the mission profile.

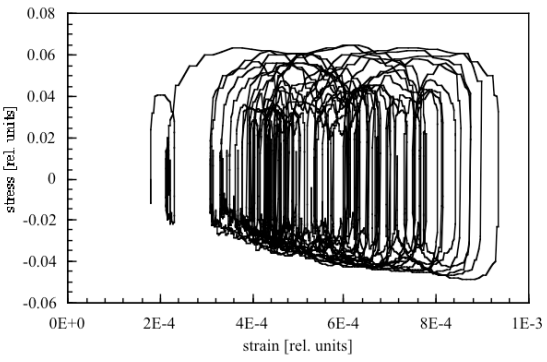

Fig. 3 Stress-strain plot for the mission profile of Fig. 1.

In spite of its apparent complexity, the model based on the creep behavior is robust and returns lifetimes that are not strongly dependent on the uncertainty of the assumed parameters. Furthermore, the lifetime can be easily extrapolated by standard numerical procedures. The main limitation of the new model consists of the fact that, similar to the Coffin-Manson approach, it fails in predicting the decrease of the module lifetime when operated at higher average temperatures. This is due to the assumption that creep is the only plastic deformation mechanism, which leads to the thermal cycling damage. More complicated models can be developed, which take into account additional mechanisms as for instance stress relaxation, anisotropic effects, and micro structural changes. Unfortunately, the number of free parameters associated with these approaches would make the calibration of those models virtually impossible.

6 Conclusions

Accurate and robust procedures for the reliability metrology are an important prerequisite for a consistent modelization and prediction of the lifetime of an item. Models have been discussed for the cases of a constant failure rate (exponential) and for a constant rate increasing in time (Weibull). The exponential distribution is more adequate to model semiconductor-related random failures, whereas package-related wear out mechanisms are more properly described by a Weibull distribution.

The physics behind the low-cycle fatigue wear out processes, which lead first to the degradation of the device parameters and then to the functional failure of a device are manifold and rather complex. However, there is a general consensus that the related number of cycles to the failure can be at least empirically described by a Coffin-Manson-like relationship to be calibrated by experimental data.

The experience has shown that complex physical models taking into account all possible driving forces are not necessarily more accurate in predicting the lifetime of a device as one-dimensional models, which reproduces the basic evolution of the degradation as a function of the time. This is mainly due to the fact that, in general, only few experimental data are available to calibrate prediction models. Therefore, simple and numerically robust models including few calibration parameters can be handled in a more transparent way and deliver more realistic results.

7 Literature

[1] M. Ciappa
Some Reliability Aspects of IGBT Modules for High-Power Applications
Hartung-Gorre Editor, 2001, Konstanz

[2] M. Ciappa, F. Carbognani, P. Cova, W. Fichtner
A Novel thermo-mechanics-based lifetime pre diction model for cycle fatigue failure mechanisms in power semiconductors
IEEE Trans. on Dev. and Mat. Reliability 3(2003)191-196

[3] M. Ciappa
Lifetime Prediction on the Base of Mission Pro files
Microelectronic Reliability 45(2005)1293-1298

[4] M. Ciappa, W. Fichtner
Lifetime prediction of IGBT modules for traction applications
IEEE Int. Reliab. Physics Symp. 38(2000)210-216.

[5] M. Ciappa, P. Malberti
Plastic Strain of Aluminum Bond Wires in IGBT Multichip Modules under Thermal Cycling
Quality and Reliability in Engineering International 12(1996)297-303

[6] M. Ciappa
Selected Failure Mechanisms of Modern Power Devices
Microelectronics Reliability 42(2002)653-667

[7] A. Castellazzi, M. Ciappa, M. Mermet-Guyennet, G. Lourdel, W. Fichtner
Compact Modeling and Analysis of Power-Sharing Unbalances in IGBT-Modules Used in Traction Applications
Microelectronics Reliability 46(2006)1754-1759

[8] A. Birolini
Reliability Engineering
Springer (2007)

[9] M. Ciappa
Lifetime Prediction on the Base of Mission Pro files
Microelectronics Reliability 45(2005)1293–1298

[10] M. Ciappa, W. Fichtner, T. Kojima, Y. Yamada, Y. Nishibe
Extraction of Accurate Thermal Compact Models for Fast Electro-Thermal Simulation of IGBT Modules in Hybrid Electric Vehicles
Microelectronics Reliability 45(2005)1694–1699