Diss. ETH Nr. 13790

Some Reliability Aspects of IGBT Modules for High-Power Applications

A DISSERTATION

submitted to

SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

For the degree of

DOCTOR OF TECHNICAL SCIENCES

Presented by

Mauro P. M. Ciappa

Dipl. Phys. University of Zurich

born January 13, 1961
citizen of Claro (TI), Switzerland

accepted on the recommendation of

Prof. Dr. W. Fichtner, examiner
Prof. Ing. F. Fantini, co-examiner

2000

Seite Leer /
Blank leaf

Dies kundzutun, steht uns nicht an:
Sei standhaft, duldsam und verschwiegen !
Bedenke dies; kurz, sei ein Mann,
Dann, Juengling, wirst du maennlich siegen.

W.A. Mozart, E. Schikaneder
Die Zauberflöte, Act 1 Scene 3

Acknowledgements

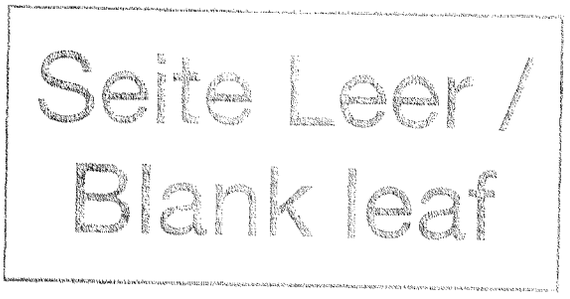
I would first like to thank my advisor prof. Wolfgang Fichtner for encouraging me to undertake this thesis and for his continuous support and advice. I thank prof. Fausto Fantini, for serving as co-examiner, but in particular for being a constant presence since the very beginning of my professional career. I am also very indebted to prof. Alessandro Birolini for all the professional opportunities he provided.

I wish to express a special thank to Paolo Malberti for the long co-operation.

I thank all the partners of the European RAPSDRA project, but in particular prof. E. Wolfgang (Siemens), Dr. H.R. Zeller (ABB Semiconductors), Dr. P. Zani (Ansaldo Semiconductors), Mr. D. Newcombe (Mitel Semiconductors), Dr. H. Berg (Eupec), Dr. R. Zehringer (ABB Research), Dr. E. Herr (ABB Semiconductors), Dr. L. Fratelli (Ansaldo-Breda), Mr. G. Coquery (INRETS), Dr. P. Cova (University of Parma), and Mr. W. Nerozzi (Ferrovie dello Stato) for the very illuminating technical discussions and the good co-operation.

I want to thank my collegues of the Integrated Systems Laboratory, in particular, Dr. Dölf Aemmer, Dr. Norbert Felber, and Lorenzo Ciampolini for providing a pleasant and creative working environment.

I thank Dr. F. Bonzanigo for his continuous advise, and the other former collegues of the Reliability Laboratory, in particular G. Nicoletti and P. Scacco for preparing some of the samples presented in this work.


Seite Leer /
Blank leaf

Content

Abstract

The assurance that a technical system will perform its intended function for the required duration and within a given environment requires a variety of engineering activities to be performed, which start from the project definition and continue during the whole life cycle. In the case of non-repairable systems this task is accomplished essentially by concurrent interdisciplinary efforts with the scope to contribute to the system architecture design, to select materials, processes, and components, as well to validate the selections made by means of tests, modeling, and analysis.

Present work deals with some aspects related to the reliability physics of Insulated Gate Bipolar Transistors (*IGBT*) for high power applications and it is organized in three main thematic sections.

In the first thematic section (*Chapter 2*) a compendium of failure mechanisms is presented, we observed to arise either during accelerated tests, or in field applications. The list mainly includes thermomechanical failure mechanisms, but it also refers to failure mechanisms, which result into burnout events. For every failure mechanism, we provided the failure modes, the physical or chemical process that leads to the failure, possible countermeasures, and where it applied also quantitative prediction models.

In the second thematic section (*Chapters 3* and *4*) physical and chemical techniques are presented, which have been specially adapted to the failure analysis and the to the thermal characterization of *IGBT* devices. Failure analysis methods include both non-destructive (e.g. electrical characterization) and destructive procedures (e.g. selective deprocessing). All techniques are illustrated in very detail and they are demonstrated

basing on real failure analysis case histories. Special attention has been paid to the preparation and to the selective delayering of the *IGBT* chip. The recipes and the working conditions of the chemical solutions we applied successfully for these processes are specially documented for enabling reproducibility. Due to the relevance of the junction temperature and of the junction temperature evolution in activating the most important failure mechanisms, two experimental techniques are presented for the quantitative characterization of this parameter. The first method is an application of the infrared microradiometry for temperature mapping at the steady state. In this case the sensitivity and the reproducibility of the technique have been improved by the use of a dedicated surface coating layer. The second method relates to an experimental set up we developed for acquiring the transient cooling curve of a single *IGBT* chip device. Because of the favorable experimental conditions the junction temperature measured in the steady state by both techniques has been demonstrated to agree within *10%*.

In the third thematic section (*Chapters 5* and *6*) two models have been developed for predicting the lifetime due to two specific failure mechanisms: time dependent dielectric breakdown (*TDDB*) and bond wire lift off. The model for *TDDB* assumes that extrinsic and intrinsic breakdown occur according to the formalism for two competing *Weibull* distributions. Basing on this assumption all relevant reliability parameters are derived analytically. Unlike the common *TDDB* models present approach is non-deterministic and enables to predict the failure rate even for the lowest quantiles of the distribution function. The dependence of the distribution parameters on the applied gate voltage has been derived from the invariant transforms of the *Weibull* distribution. The acceleration law has been validated with experimental data and it has been used for the optimization of screening procedures, in order to achieve the minimum yield loss. Finally, a model has been developed for the bond wire lift off mechanism, which predicts quantitatively the lifetime due to bond wire lift off in devices submitted to cyclic loads as they are encountered in current converters of railway systems. It assumes linear accumulation of the thermal-cycle fatigue damage, it is calibrated basing on data from accelerated tests, and it takes into account the redundancy of the bond wires within a complex multichip module.

Riassunto

Per assicurare che un sistema esegua la funzione richiesta durante un periodo stabilito e nelle condizioni ambientali fissate è necessario attuare tutta una serie di misure a partire dalla fase di progettazione e durante tutto il periodo di vita utile. Nel caso particolare di sistemi non riparabili, la realizzazione pratica di questa strategia passa essenzialmente per la definizione di un'architettura di sistema robusta, per la scelta corretta dei materiali, dei processi e dei componenti, come pure per la valutazione dei prototipi mediante adeguate prove di collaudo, analisi di guasto e modellizzazioni. Tutto questo richiede il concorso di numerose discipline tecniche e scientifiche.

In questo lavoro vengono trattati alcuni aspetti della fisica dell'affidabilità relativi ai componenti noti come Integrated Gate Bipolar Transistor (*IGBT*) utilizzati per applicazioni di alta potenza. La trattazione seguente è suddivisa in tre ambiti tematici.

Nella prima parte (*Capitolo 2*) si presenta un compendio di meccanismi di guasto osservati a seguito di prove accelerate e di collaudi in campo. In aggiunta ai meccanismi in prevalenza di origine termomeccanica, vengono trattati anche alcuni meccanismi di guasto intrinseci. Per ogni meccanismo di guasto sono stati specificati i sintomi caratteristici, i processi fisici o chimici che ne stanno alla base, eventuali contromisure tecnologiche e ove la cosa fosse possibile anche modelli per la predizione del tempo di vita.

Nella seconda parte (*Capitoli 3 e 4*) vengono trattate quelle tecniche di caratterizzazione chimico-fisica che son state specificamente adattate alle esigenze dell'analisi di guasto e della termometria negli *IGBT*. Esse includono metodiche non-distruttive (*p.e.* la caratterizzazione elettrica) e metodiche distruttive (*p.e.* deprocessamento del chip). Le tecniche sono discusse in dettaglio e vengono illustrate sulla base di casi reali di analisi di guasto. E' stata prestata particolare attenzione alle tecniche di preparazione del semiconduttore e agli attacchi selettivi per i diversi strati del chip. La composizione e le condizioni di utilizzazione delle diverse

soluzioni chimiche sono state documentate dettagliatamente in modo da garantire la riproducibilità dei risultati. In considerazione dell'importantanza della temperatura di giunzione nell'attivazione dei principali meccanismi di guasto, vengono presentate due metodiche specifiche per la determinazione quantitativa di questo parametro. La prima tecnica è un'applicazione della microradiometria infrarossa per l'acquisizione di mappe di temperatura di *IGBT* all'equilibrio termico. In questo caso la sensibilità e la riproducibilità della metodica sono state di molto migliorate rispetto alle prestazioni standard, grazie allo sviluppo di una particolare vernice per l'equalizzazione dell'emissività dei diversi materiali. La seconda tecnica utilizza un banco di prova appositamente progettato per l'acquisizione del transitorio di raffreddamento negli *IGBT*. Si è potuto dimostrare che, grazie alle condizioni sperimentali favorevoli, le temperature di giunzione misurate con entrambe i metodi non si discostano di più del *10%*.

Nella terza parte (*Capitoli 5* e *6*) sono stati elaborati due modelli per predire il tempo di vita di un dispositivo a seguito di due meccanismi di guasto specifici: la rottura del dielettrico ad alti campi (*TDDB*) e il distacco dei fili di alluminio (*bond wire lift off*). Nel modello proposto per il *TDDB* si presuppone che entrambe i meccanismi di breakdown intrinseco ed estrinseco siano descritti dal formalismo relativo alla competizione di due distribuzioni di *Weibull*. Questo assunto consente di derivare analiticamente i principali parametri affidabilistici. Contrariamente ai modelli tradizionali, quello proposto non è deterministico e consente di calcolare il tasso di guasto anche per i primi quantili della distribuzione cumulativa. Inoltre, la dipendenza dalla tensione di gate dei parametri della distribuzione cumulativa è ricavata a partire dalle trasformazioni invarianti della distribuzione di *Weibull*. Le leggi di accelerazione ottenute in questo modo sono state validate mediante dati sperimentali e sono state inseguito applicate per ottimizzare i processi di setacciatura in modo da rendere massima la resa, riducendo al minimo gli scarti. Infine è stato messo a punto un modello per la predizione di vita per il meccanismo del distacco dei fili (*bond wire lift off*) in *IGBT* sottoposti a cicli termici rappresentativi di quelli incontrati in convertitori per sistemi di trazione ferroviaria. Il modello assume che il danneggiamento prodotto dalla fatica termomeccanica sia accumulato in modo lineare e tiene in debita considerazione la struttura ridondante dei fili all'interno di un modulo complesso.

Chapter 1

Introduction

1.1 Working background

Thanks to the recent improvements made in handling large currents at high voltage and at high switching frequency, *Insulated Gate Bipolar Transistors* (IGBT) have almost completely replaced bipolar *power transistors* (BPT) and they are challenging the position of *Gate Turn-Off thyristors* (GTO) in their traditional fields of application.

In the last five years, the need of 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, for power transmission, and for power distribution applications. In order to cope this demand, various wide-band research projects have been started which involved both industry and academia.

The European Project Reliability of Advanced *High Power Semiconductors for Railway Traction Applications* (RAPSDRA, 1995-1999) has been based on requirements defined by railway operators, who were apart of a wide consortium including also European device manufacturers and research institutes.

Basing on the experience with *GTOs* and on logistic considerations (no preventive maintenance for semiconductors devices) the consortium has summarized the project reliability goals in 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* (1 failure in 10 millions cumulated operating hours) over the whole useful life of the device. In addition, field measurements on locomotives have lead to the definition of equivalent environmental conditions describing typical railway applications (e.g. *Table 1.1*).

Table 1.1 Equivalent junction temperature swings and number of cycles over 30 years operation as defined in [1] (A: stop to stop, B: station to station, C: day to day, D: year to year)

	A		B		C		D	
	ΔT	Cycl	ΔT	Cycl	ΔT	Cycl	ΔT	Cycl
Metro, Tram, Bus	40K	10^7	50K	10^6	60K	10^4	170K	30
Suburban	50K	$2\ 10^6$	60K	10^5	80K	10^4	170K	30
IC and high speed	60K	$2\ 10^5$	80K	$4\ 10^4$	100K	10^4	170K	30

Starting from existing *IGBT* products in bond wire technology, the efforts of several working groups within *RAPSDRA* have been focused on two objectives: postpone the occurrence of wearout mechanisms beyond the useful life of the device and keep the failure rate due to random failures within the specified limits.

While wearout failure mechanisms can be attacked by adequate design rules (and represent essentially a cost optimization problem), 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 quality of manufacturing processes. Nevertheless, as shown in *Figure 1.1*, random failures play a

very relevant role in determining the survival probability of a mature system.

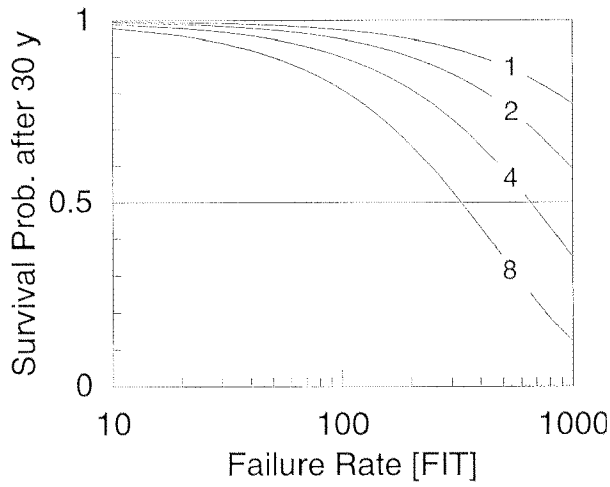


Figure 1.1 Survival probability after 30 years operation of a unit consisting of a different number of modules (parameter) without redundancy as a function of the constant failure rate of every single module.

For instance, the survival probability after *30 years* operation of a unit with *6* modules, each having a constant failure rate of *100 FIT* is close to *0.85*. In converse, for a constant failure rate of *400 FIT*, it is in the unacceptable range of *0.5*. This simple but realistic case stresses the importance of developing efficient techniques for controlling the failure rate during all phases of the useful life of a complex device.

Under previous assumptions, the traditional reliability metrology approach based on the *a posteriori* failure rate assessment cannot be used here. In fact, even a simple measurement of the failure rate as function of time would require many millions of cumulated component-hours (even at accelerated conditions). In addition it would not provide relevant information about the corrective actions for reliability improvement.

Instead of this a more sophisticated reliability engineering approach has been used in *RAPSDRA*, being based on the concept of built-in reliability. The built-in reliability strategy is currently used in *ULSI* manufacturing for improving the reliability of already mature technologies [3]. The main idea behind this approach is the continuous control of those process parameters, which may affect the reliability of the final product. This initial phase is followed by dedicated experiments and characterizations, which are intended to investigate the system response over the variation

of a given parameter. Finally, the obtained information is returned into a *feedback loop* for finely tuning the process conditions. The engineering discipline dealing with the theoretical and experimental aspects of the back end of the built-in reliability process is called *reliability physics.*

A successful example of reliability growth of high power *IGBT* devices is presented in *Figure 1.2.*

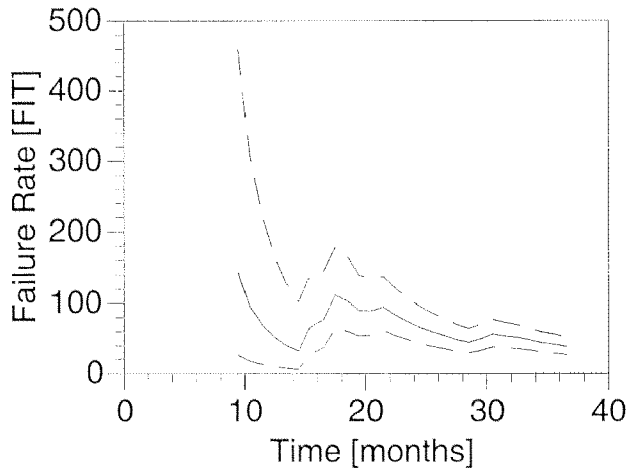


Figure 1.2 Reliability growth of high power IGBT devices. Dashed lines represent the upper and the lower limits of the failure rate interval estimated with a confidence level of 0.9. The solid line is the point estimate of the failure rate.

Figure 1.2 refers to the investigation of non-package-related field failures observed in a population of *150'000* high-power *IGBT* devices delivered between *1995* and *1998* and corresponding to about *700 millions* of cumulated *device-hours* [4]. The rapid decrease of the failure with time and the consistent saturation to a level of about *60 FIT* is a practical demonstration of the efficiency of the built-in reliability approach, as well of the capability of the *IGBT* technology to reach the required reliability specifications.

1.2 Scope, main results, and content of the thesis

The work behind this thesis was intended to address three different aspects of the reliability physics of high power *IGBT* multichip modules, namely modeling of intrinsic and package related failure mechanisms,

development of efficient failure analysis techniques, and modeling of the lifetime for dominant failure mechanisms.

Modeling of intrinsic and package related failure mechanisms

In *Chapter 2*, the failure mechanisms issued either from accelerated tests or from field failures (except partial discharge) have been identified, documented and characterized. The associated physical (or chemical) mechanisms have been described with the observed failure modes, the probable root causes, and possible technological countermeasures. Finally, where it was required by the relevance of the failure mechanism, quantitative lifetime model has been provided as function either of the time, or of the number of thermal cycles.

Development of efficient failure analysis techniques

In *Chapter 3*, efficient failure analysis techniques and procedures for IGBT modules have been especially trimmed for *IGBT* devices and successfully tested. After describing the fundamentals, the advantages and the drawback of each technique, they have been illustrated basing on original case histories related to *IGBT* devices. The use of several microscopy techniques is demonstrated, in particular of optical microscopy, scanning electron microscopy, emission microscopy, and scanning probe microscopy. Special attention has been paid to sample preparation techniques. They include procedures for non-destructive encapsulation of *IGBT* modules, for selective delayering and for structure delineation of *IGBT* chips. The related process conditions and recipes have been clearly documented.

In *Chapter 4*, an experimental set up has been realized, which enable to measure both the thermal resistance and the thermal impedance of *IGBT* subassemblies. The equipment combines static temperature measurements by infrared thermography with the acquisition of cooling transients. The sensitivity of the infrared thermograph has been improved by the use of an innovative surface coating, which highly enhances the emissivity of the sample without perturbing the electrical behavior of the device. The impedance is measured by the acquisition of the V_{CE} transient at a constant low level injection of I_C when the power is switched-off. Temperature drifts with time of the heat sink can be eliminated by a differential measurement technique controlled either manually or by a

simple finite state machine. The static temperature values provided by both systems have been shown to deviate by less than *0.5 °C*.

Modeling of the lifetime for dominant failure mechanisms

Since gate oxide breakdown events occur as random failures due to the presence of process-related defects in the dielectric, a quantitative model has been developed in *Chapter 5*, which enables to predict the failure rate of *IGBT* devices, basing on the results of breakdown experiments on representative *MOS capacitors*. The model assumes that the robust and the defective sub-populations are both statistically described by two different *Weibull* distributions. Starting from this statistical model, the most relevant reliability parameters have been derived analytically from a probabilistic model, which bases on the formalism for competing risks. The model has been strictly derived from clear assumptions and some inconsistencies found in the literature have been clarified. A new approach assuming the invariance of the *Weibull* statistical model has been pursued by leading to definition of time transforms, which simplify the calculations involved in the design of screening strategies and the evaluation of the related yield.

Finally, in *Chapter 6* a model has been developed for the bond wire lift off mechanisms, which predicts quantitatively the lifetime of devices submitted to cyclic loads as they are encountered in current converters of railway systems. It takes into account the redundancy of the bond wires within an *IGBT* module, the thermomechanical stress due to realistic application profiles, and it assumes linear accumulation of the thermal-cycle fatigue damage. Model calibration has been carried out with experimental data obtained from power cycling experiments. Bond wire redundancy has been shown to play a minor role. On the contrary, the model has been shown to be very sensitive both against the *Coffin-Manson* parameters and on the frequency distribution of the temperature swings suffered by the device.

1.3 Investigated devices

The devices that have been investigated are *NPT IGBT* multichip modules from different manufacturers rated for *300 A* up to *1200 A* and with a

blocking voltage ranging from *1.2 kV* up to *3.5 kV*. All the investigated devices were mounted with anti-parallel freewheeling diodes.

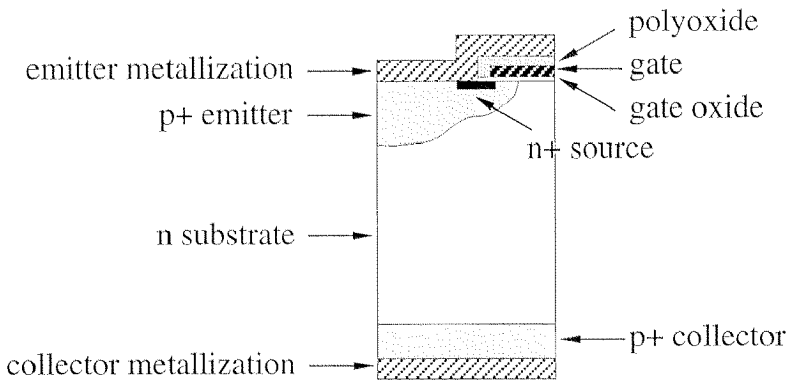


Figure 1.3 Schematic cross-section of a non-punch-trough (NPT) IGBT

The simplified cross-section of an *IGBT* device is presented in *Figure 1.3*. More technological details can be found in *Chapter 3*, and especially in the micro-sections of *Figure 3.21* and *3.22*.

The *IGBT* is basically a four-layer structure, which does not exhibit regenerative turn-on. The device is brought in on state by applying a constant gate bias. The conducting *n-channel MOS* structure enables the injection of electrons into the *n-substrate* and causes the p^+*-collector* to inject holes into the *n-substrate*. The resulting carrier density in the *n-substrate* exceeds by more than three orders of magnitude the background doping concentration. The output characteristics of an *IGBT* are similar to those of a *bipolar transistor*, except in the saturation regime, where they are dominated by the *MOS* characteristics. The *IGBT* is fully controlled by the *gate electrode*, since, when the *gate bias* is removed, the electron injection is instantly interrupted, by stopping also the hole current from the anode. The state of the art and the future developments of the *IGBT* technology are exposed in [2].

The structure of high-power *IGBT* multichip modules in bond wire technology for is discussed in detail in *Chapter 2*, while recent packages in *press-packaging technology* are presented in [5].

Chapter 2

Failure mechanisms of IGBT modules

2.1 Introduction

Failure mechanisms are physical, chemical, or other processes resulting into a failure. For practical purposes they can be divided in two categories. The first includes mechanisms, which result from poorly controlled or poorly designed manufacturing processes. The second category includes those failures, which occur during the normal operation of the device. In the case of a mature product the first category of failures is normally suppressed by a suitable reliability assurance program, which can include inspections and other screening procedures. 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 first category, and to develop quantitative models for the second with the scope to realize the concept of built-in reliability.

In the following, we will shortly review the most frequent failure mechanisms we observed either during reliability tests, or in field applications. Almost all failure mechanisms listed below are package-related and refer to thermomechanical stresses. This is because in our case the vast majority of the failures originate from accelerated tests, and in particular from thermal cycling experiments. At present, there is no statistical data accounting for the occurrence probability of each mechanism in field applications. More process related failure mechanisms, like those associated with crystal and oxide defects and ionic contamination are discussed in *Chapter 3* in conjunction with some targeted failure analysis techniques.

Where applicable, we provide some simple *lifetime models*, mainly based on *power laws*. The presented models are semi-quantitative and have not been especially validated for *IGBT* devices. Where dedicated material or model parameters were not available, engineering estimates have been used, which have been extrapolated from similar microelectronic applications. Additional models and analytical procedure for estimating the mechanical stresses, which arise in multilayered structures, are presented in [6].

2.2 Package-related failure 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 capabilities. The schematic cross-section through a module of *type A* (*e.g.* a standardized *E2* package) is represented in *Figure 2.1a*, and the related physical parameters are listed in *Table 2.1*. Starting from the bottom one can recognize the *base plate*, the *direct copper bonded ceramic substrate*, the *silicon chip*, the *aluminum metalization* (not shown), and the *bond wire*.

A *type B* module is schematically sketched in *Figure 2.1b*. The most important structural difference with the *type A* module is the use of a *strain buffer layer* soldered on the top of the *IGBT* chip (layers *a'* and *b'* in *Figure 2.1b*).

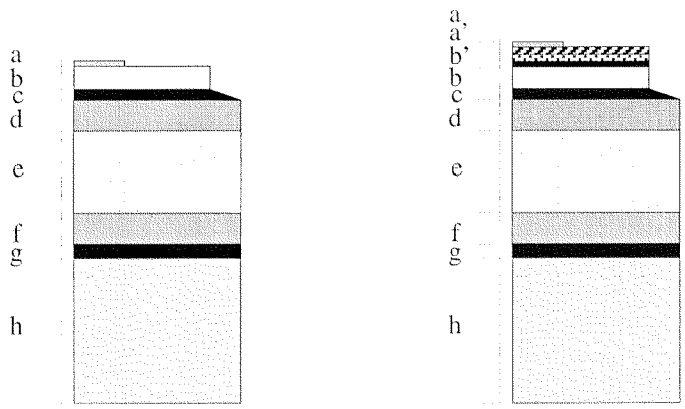


Figure 2.1 Schematic representation of the multilayer in a multichip module of type A (a) and (b) of type B

In both module types they are additional compliant layers, which are placed at the interface of materials with large differences in thermal expansion. This is the case of the *solder layers*.

Table 2.1 Thickness (t), coefficients of thermal expansion, typical length (L)

	Material	t [μm]	*CTE [ppm/°C]*	*L [mm]*
a	Al	300	22	1
b	Si	250	3	12
c	solder	100	compliant	
d	Cu	280	not relevant	
e	Al_2O_3 or AlN	1000	7 or 4	30 - 55
f	Cu	280	not relevant	
g	solder	180	compliant	
h	Cu or AlSiC	4000	17 or 8	

When considering thermal cycling of these multilayered structures and the consequent thermomechanical fatigue induced failure mechanisms, it is important to take into account all the factors, which play a role in determining thermomechanical stresses.

In first approximation, they are the mismatch in the *coefficient of thermal expansion* (CTE), the *characteristic length* of the layer, and the *local temperature swing* (*see* also *Chapters 3* and *6*). The first two parameters for the relevant materials and layers in a multichip module are listed in *Table 2.1*. In *Figure 2.2a* we represent the computed (one-dimensional approximation) relative temperature swing at different interfaces within two modules having the same geometry (as in *Table 2.1*), but with layers

made of different materials. In particular, the first module includes a ceramic substrate of Al_2O_3 and a base plate of copper, while the ceramic substrate of the second module is *AlN* and the base plate *AlSiC*.

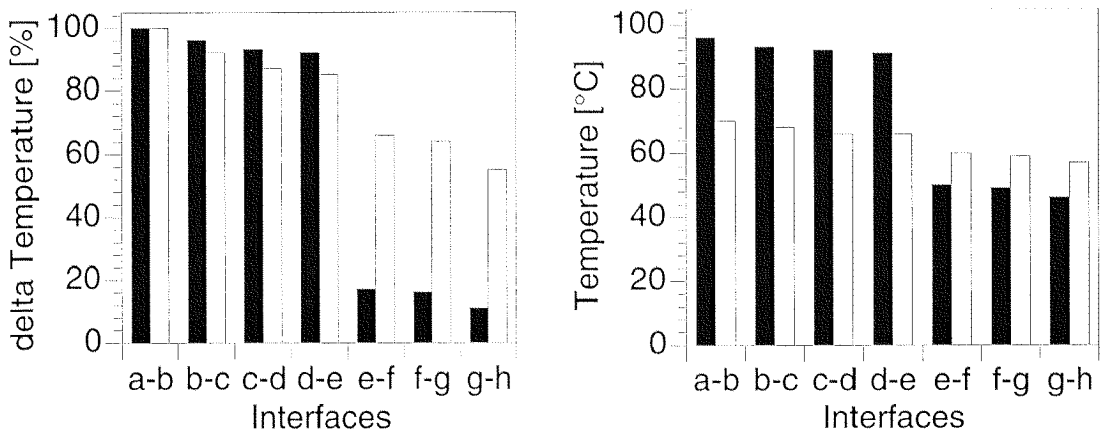


Figure 2.2 (a) Temperature swing distribution at the interfaces of both stacks in Figure 2.1 (black: type A, white: type B; a-b: Si, b-c: Si-solder, c-d: solder-Cu, d-e: Cu-ceramic, e-f: ceramic-Cu, f-g: Cu-solder, g-h: solder-base plate. (b) Temperature at the interfaces for a dissipated power of 100 W and a heat sink temperature of 40°C (black: type A, white: type B).

From *Figure 2.2b*, it can be seen that, due to the good conductivity of *AlN*, the maximum temperature swing in the multilayer of *type B* is about *50%* lower than in the stack of *type A*. In converse, it can be seen that the most relevant temperature drop (about *80%*) within the first multilayer occurs across the Al_2O_3 ceramic substrate. In the second multilayer, the largest temperature drop virtually occurs across the *AlSiC* base plate and the isothermal heat sink, such that all interfaces experience almost the full temperature swing.

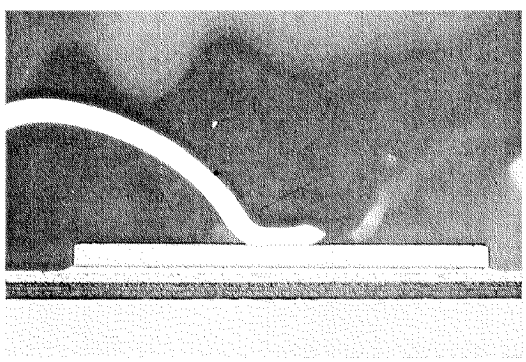


Figure 2.3 Cross-section through the gate bond wire of an IGBT module of type A (Cross-section, optical image 3x)

From *Table 2.2*, it can be seen that the largest difference in *CTE* affects the aluminum (bond wires, metalization) and the silicon chip. In the first multilayer, the mismatch is worsened by the fact that both materials are in intimate contact. On the contrary, the strain buffer used in the second stack, which consists of an aluminum layer bonded onto a molybdenum plate (*CTE 2.5 ppm/°C*), dramatically reduces the thermomechanical stresses experienced by aluminum bond wires. On the second and third place in terms of *CTE* mismatch, one can mention the ceramic substrate and the base plate (especially Al_2O_3 and *copper*), and the silicon and the ceramic base plate (especially *silicon* and Al_2O_3), respectively. Compliant solder layers separate the last two couples of materials.

Since the smaller is the lateral size of a layer, the smaller is the thermomechanical stress, several small-sized structures are preferred instead of single large plate (especially for brittle materials). Unfortunately, the size of almost all components is determined by physical constraints (*e.g.* size of the *IGBT* chip). The only degree of freedom concerns the ceramic substrate; such that in advanced packages it is partitioned in squares with a side length, which ranges from *30* to *55 mm*. Small-sized ceramic substrates are normally used when combining *AlN* with *copper* base plates.

Table 2.2 Differential elongation at the conditions of Figure 2.2

Type A		Type B	
Si – bond wire	2 on1000 μm	Si – bond wire	0 on 1000 μm
Si - Al_2O_3	4 on12000 μm	Si - AlN	1 on 12000 μm
Al_2O_3 - copper	28 on 55000 μm	AlN - AlSiC	7 on 30000 μm

Table 2.2 summarizes the combined effect of the *CTE* mismatch, temperature swing, and size in the case of the stacks represented in *Figure 2.1a* and *2.1b*. The differential elongations, which have been computed according to the one dimensional approximation (*see Chapter 6*) for a heat sink temperature of *40°C* and a dissipated power of *100 W*, clearly identify the critical interfaces of both types of multichip module.

In the following, we consider those package-related failure mechanisms, which are activated directly or indirectly by thermomechanical stresses. For sake of clarity, they are classified into *bond wires fatigue*, *metalization fatigue*, *brittle cracking* and *fatigue*, *solder fatigue*, and *stress corrosion*.

2.3 Bond wire fatigue

Multichip *IGBT* modules for high-power applications typically include up to *800* wedge bonds. 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. The current capability of a bond wire decreases as well-known over-proportionally with the length and just slightly depends on the substrate temperature. The maximum *DC current* capability of a bond wire is limited by melting due to ohmic self-heating. In a *1 cm* long wire loop in air it is of *25 A* for *300 µm* (*35 kA/cm²*) and of *60 A* for *500 µm* (*30 kA/cm²*) aluminum wires. 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. During switching operation the current density distribution across the section of a bond wire is strongly inhomogeneous due to the *skin effect*. The wires are connected by ultrasonic wedge bonding either onto the aluminum metalization (with a thickness ranging from *3* to *5 µm*), or onto the strain buffer.

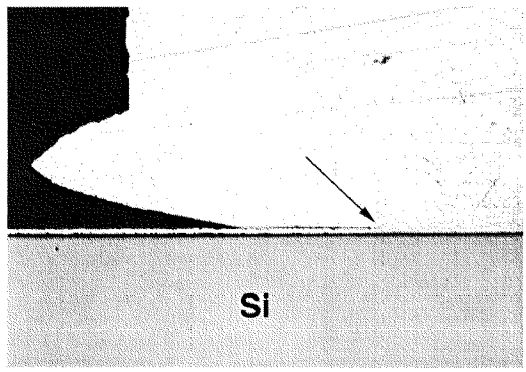


Figure 2.4 Cross-section of a virgin wedge bond (tail side) on aluminum metalization, showing the transition to the interdiffused region (Optical image, 120x)

Ultrasonic wire bonding involves heat and pressure (solid state welding). The ultrasonic vibrational energy provided by the bonding tool renders

the wire material temporarily soft and plastic and causes it to flow under pressure. During bonding, the temperature rise at the wire-metalization interface can approach *30* to *50 percent* of the melting point of aluminum. In *Figure 2.4* is represented a cross-section through a wedge bond, before thermomechanical stress. The arrow indicates the transition from the non-bonded to the welded region, where the bond wire material cannot be distinguished from the aluminum metalization.

Failure of a wire bond occurs predominantly 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 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} [7]. 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.

2.4 Bond wire lift off

Although the bond wire lift off is treated in very detail and quantitatively modeled in *Chapter 6*, some additional remarks are reported below, due to the relevance of this failure mechanism. Bond wire lift off has been observed to affect both *IGBT* and freewheeling diodes. However, since power cycling experiments are usually performed with unipolar current sources, these last are often ignored. No bond wire lift off occurs at the wire terminations bonded onto copper lines. This is mainly due to the fact that copper lines do not experience large temperature swings. Additionally, the *CTE* mismatch between aluminum and copper is less severe than with silicon.

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 (*Figure 2.4*), and propagates within the wire material until the bond wire completely lifts off.

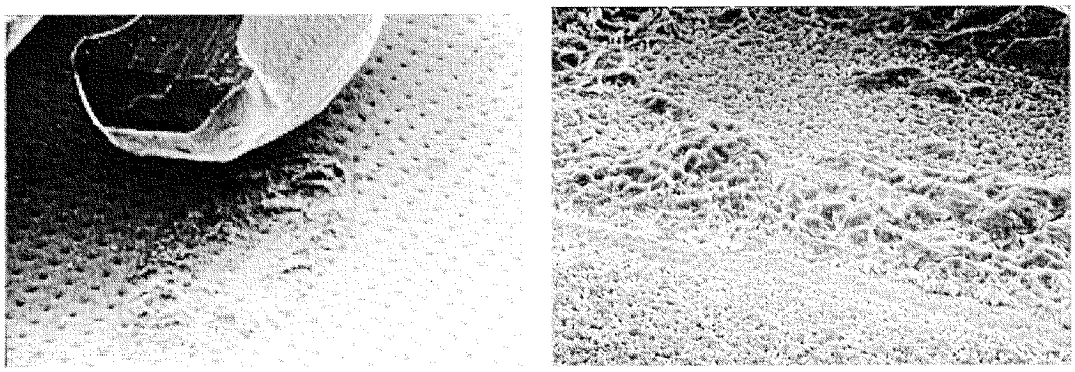


Figure 2.5 (a) Bond wire lift off (SEM image, 40x). (b) Close view of the footprint of an aluminum bond wire after lift off (SEM image, 100x).

Polycrystalline metals exhibit yield strength. If the stresses exceed this value flow is very rapid. If they are below, the compliant behavior of the material depends on the amplitude of the applied stress and on the time. The kinetics of the flow process is controlled by effects at atomic scale, like the glide motion of dislocations and the diffusive flow of individual atoms [10]. Several attempts have been made to estimate the operating lifetime of bond wires by numerical simulation basing either on *continuous mechanics* models [8] or on *quasi-atomistic models* including grain boundaries [9]. Generally, the quantitative use of simulations is limited by the complexity of the three-dimensional structure of the bond, and due to the uncertainty in evaluating the initial stresses induced by the strong deformation of the wire through the bonding tool. This is the reason why the reliability of different types of solid-state welded contacts is still investigated experimentally.

Figure 2.5a shows a bond wire after lift off. Due to the spring action exerted by the aluminum wire loop, the wire looses the electrical continuity with the *IGBT* chip.

The close up into a footprint of a lifted bond wire in *Figure 2.5b*, clearly indicates that the crack propagates within the wire material and not at the interface as it would be the case either of poor welding, or of delamination of the metalization layer. Furthermore, it can be seen, that welding just occurs at the periphery of the joint, while in the central region, the wire is not in contact with the metalization, as it can be deduced from the occurrence of reconstruction. The continuous rim around the footprint is due to the pressure exerted on the metalization layer during thermal expansion. Additional images on this failure mechanism can be found in *Chapters 3* and *6*.

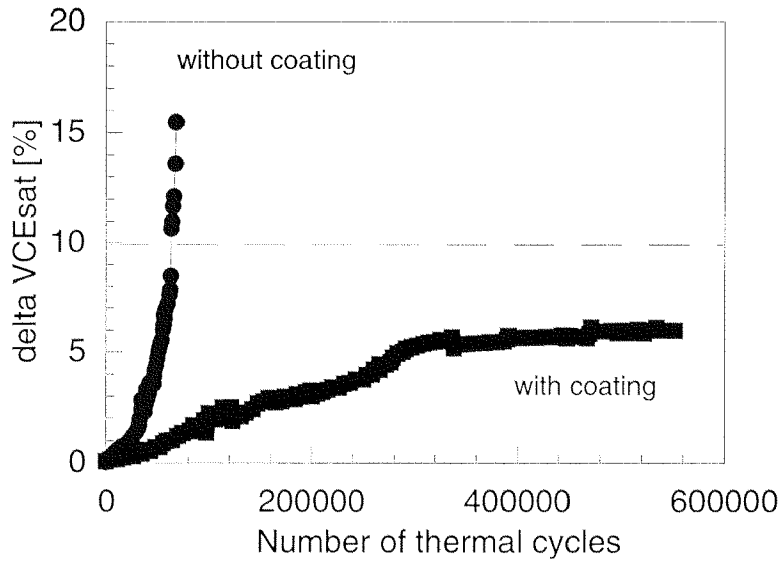


Figure 2.6 Degradation of V_{CEsat} in an IGBT module without and with a polymeric bond wire coating layer; $T_l = 65°C$, $T_h = 125°C$, $t_{on} = 0.8$ s, duty cycle 0.5.

At present, two main technological countermeasures are common for facing the bond wire lift off failure mechanism. The first one makes use *molybdenum-aluminum strain buffers* [12], which are mounted on the top of the *IGBT* and of the diode chips, with the scope to eliminate thermomechanical fatigue by distributing the *CTE* mismatch of aluminum and silicon across a thick layer (*Figure 2.1b*). The second solution is a symptomatic countermeasure, which aims to avoid the physical separation of the wire from the bond pad, once the welding joint fails. This scope is achieved by gluing the bond wires with a coating layer. The coating consists of one or of multiple polymeric layers with graded hardness, which are painted onto the wires immediately after ultrasonic bonding. *Figure 2.6* reports the results of a very early experiment [11], where the efficiency of polymeric coatings in slowing down the consequences of the bond wire lift off is clearly shown. Additional solutions like direct chip cooling for quenching large temperature swings at the chip surface are envisaged for the future.

2.5 Bond wire heel cracking

Bond wire heel cracking rarely occurs in advanced *IGBT* multichip 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 thermomechanical 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.

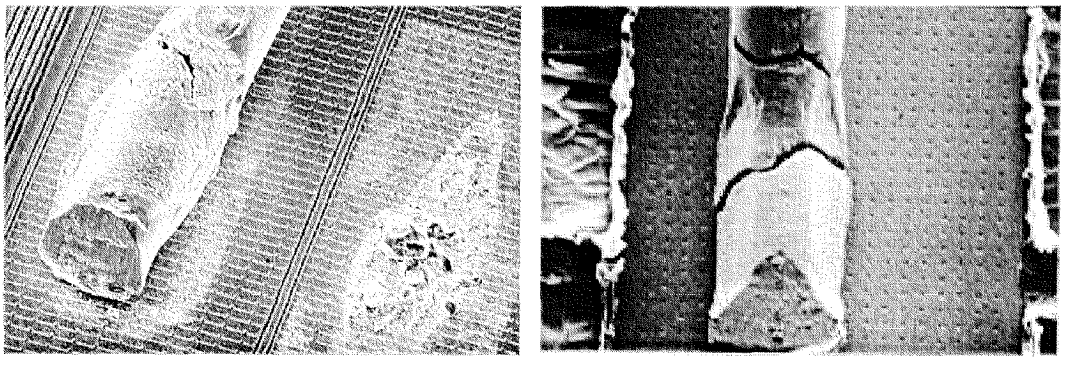


Figure 2.7 (a) Bond wire heel cracking due to low-cycle fatigue stressing (SEM image, 25x). (b) Bond wire cracking due to improper bond wire coating (SEM image, 25x).

Figures 2.7a and *2.8a* show two examples of heel cracking at a single and at a double bond, respectively. In the first case heel cracking and bond wire lift off occur at the same time. However, while the adjacent bond has been completely removed by lift off, the cracked wire still presents some electrical continuity with the chip. This is a clear indication of fact that even if the bonding parameters are not too close to the optimum, heel cracking is slower than the lift off mechanism.

The couple of wires in *Figure 2.8a* indicates that heel cracking preferably occurs at those locations where the aluminum wire has been previously damaged by the bonding tool. In fact, the bond wire at the left in *Figure 2.8a* presents a thin crack at the same location where the crack fully developed in the wire at the right side. Additionally, it has to be mentioned that the temperature distribution in double bonds due ohmic

self-heating and indirect heating through the chip is much more asymmetric than for single bonds.

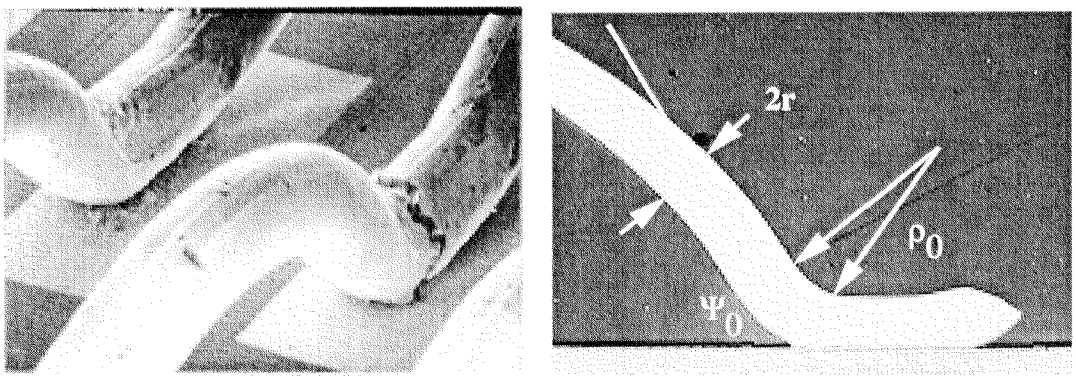


Figure 2.8 (a) Heel cracking in a double wire bond. Crack initiation can also be observed in the double bond in the back (SEM image, 25x). (b) Parameter definition for the lifetime model.

The failure imaged in *Figure 2.8b* could lead to a wrong identification of the failure mechanism. In fact, in this case, the wire rupture has not been caused by heel cracking, but by the shear stress arising due to the use of a rigid bond wire coating (selectively removed in *Figure 2.8b*).

The model of *Schafft* [13] enables to predict analytically the number N_f of thermal cycles (ΔT) to heel cracking due to bending stress. It bases on the power law

$$N_f = A\,\varepsilon_f^n \tag{2.1}$$

where A and n are constants for a particular material and the wire strain ε_f is computed according to

$$\varepsilon_f = \frac{r}{\rho_0}\left(\frac{ar\cos\left(\left(\cos\psi_0\right)\left(1-\Delta\alpha\,\Delta T\right)\right)}{\psi_0}-1\right) \tag{2.2}$$

$\Delta\alpha$ is the mismatch in the *CTE* of aluminum and silicon, while ψ_0, ρ_0, and r are geometrical parameters defined in *Figure 2.8b*. The values $A = 3.9\ 10^{-10}$ and $n = -5.13$ are engineering estimates [6,14], which are usually encountered in microelectronic applications of aluminum bond wires with a diameter below $100\ \mu m$.

2.6 Aluminum reconstruction

Although reconstruction of the aluminum metalization is an effect, which has been encountered since the early times of microelectronics [15,16], the occurrence of this degradation mechanism in *IGBT* multichip modules has been firstly reported in [17,18].

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 thermomechanical mismatch between both materials and due to the stiffness 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.

Figure 2.9a and *2.9b* show how the metalization of an *IGBT* and of a freewheeling diode appears after reconstruction. In optical images reconstructed regions look dark, because of the light scattering due to the surface roughness. Reconstruction is more evident at the center of the chip, where the junction temperature reaches its maximum. It has been shown by infrared thermography [17] that surface reconstruction is negligible in those peripheral regions of the chip, where the maximum junction temperature does not exceed *110°C*.

Figure 2.9b shows that surface reconstruction sometimes occurs as a secondary mechanism in conjunction with bond wire lift off. In fact, after release of the bond wires on the left side of the diode, the (pulsed) current has been carried by the bond wire on the right side only, by leading to an increase of the local temperature with consequent reconstruction of the metalization.

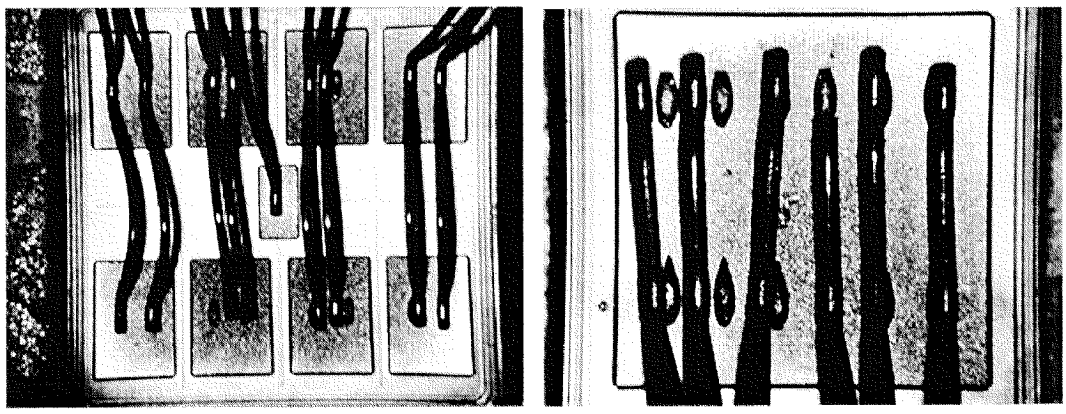


Figure 2.9 (a) Reconstructed emitter and gate metalization of an IGBT (Optical image, 4x). (b) Reconstructed metalization of a freewheeling diode (Optical image, 5x).

In *Figure 2.10a* and *2.10b* the emitter metalization of a virgin *IGBT* chip is compared with that of a similar device, which survived *3.2 million* of cycles between *85°C* and *125°C*. After stress, it can be seen that non-columnar aluminum grains are extruded from the thin film surface, while voids are present at the boundaries of larger grains.

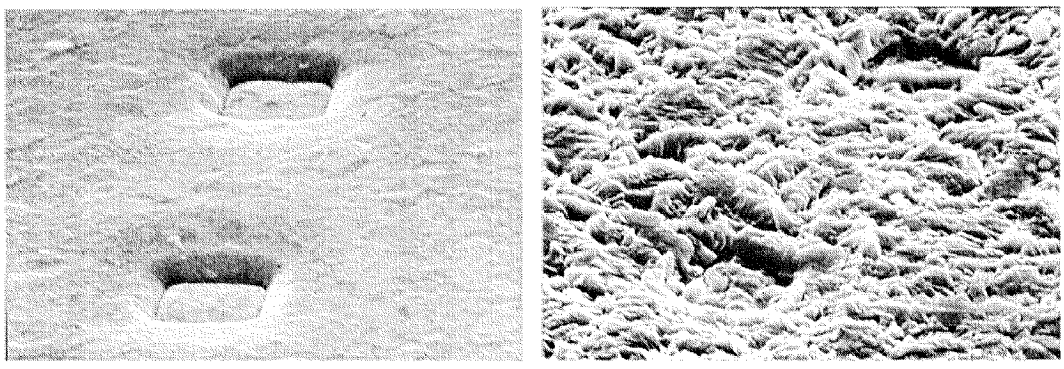


Figure 2.10 (a) Emitter metalization of an IGBT chip before power cycling (SEM image, 1000x). (b) Reconstructed emitter metalization after 3.2 millions of power cycles between 85°C and 125°C (SEM image, 1000x).

In field failures turning into a destructive burn out of the device, aluminum reconstruction may be less evident, due to remelting of the metalization as consequence of the high temperature levels that can be reached. 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, thermomechanical and electromigration effects can coalesce resulting into a complete depletion of the metalization from the wall of the via.

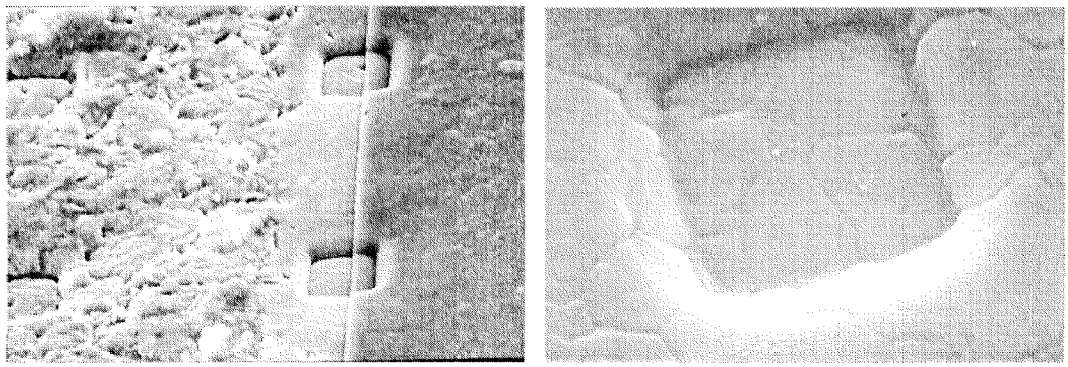


Figure 2.11 (a) Reconstructed emitter metalization after removal of the polyimide passivation (SEM image, 800x). (b) Grain boundary depletion in a passivated emitter contact after power cycling (SEM image, 1500x).

Figure 2.11a shows the role of a compressive layer in suppressing reconstruction phenomena in aluminum layers. In fact, after selective removal of the polyimide passivation, it can be clearly seen that this overlayer has almost inhibited the extrusion of metal grains in the center of the image. Therefore, the use of compressive overlayers can be considered an effective countermeasure for controlling the increase of the sheet resistance of metalization layers submitted to large temperature swings. *Figure 2.11b* shows a close view of the aluminum metalization at an emitter contact that has been coated with a compressive layer and then power cycled with a maximum temperature of *125°C*. As expected no reconstruction occurs. However, one can clearly see that the grain boundaries have been depleted as a consequence of cavitation effects. Voiding of the grain boundaries has been also observed in non-passivated metalization layers submitted to long power cycle testing with a maximum junction temperature below *100°C*.

2.7 Brittle cracking

The brittle materials used in advanced *IGBT* multichip modules are the single crystal silicon, the thin insulating layers on it, and the ceramic substrate. One among the main assumptions in fracture mechanics of brittle materials is that the sharp stress concentration at pre-existing damages leads to the rupture under the influence of external mechanical stresses. *Ultimate brittle fracture* can occur suddenly without any plastic deformation, when an initial crack is present, whose length exceeds a critical size, which is a characteristic of every brittle material [19]. Failures due to brittle cracking are usually observed immediately after mounting or powering the device. However, even if the initial crack does not reach the critical length, it can develop by fatigue crack propagation under the influence of the applied stresses, until the threshold for brittle fracture is exceeded. This results usually into early field failures, as in the case of the short circuit presented in *Chapter 3* (*Figures 3.5a* and *3.5b*), where a crack propagated through the polyoxide as a consequence of a pre-damage introduced during wire bonding.

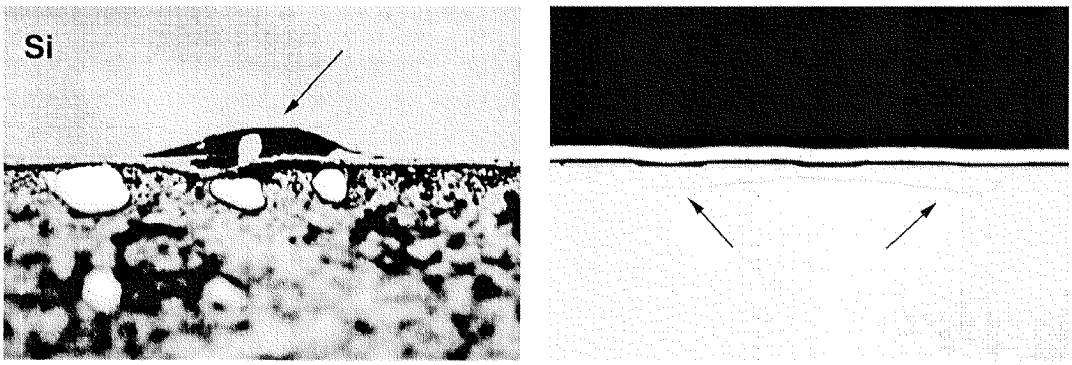


Figure 2.12 (a) Notch in the silicon chip (Micro-section, optical image, 250x). (b) Crack in the silicon chip due to bending stresses in the base plate (Micro-section, optical image, 300x).

Pre-existing defects can be originated for example by processing problems (*e.g.* during dicing), by assembly problems (*e.g.* hard wire bonding), or by soldering (*e.g.* voids in solder alloys). *Figure 2.12a* shows a notch in the bottom side of an *IGBT* chip, which has been caused during diamond sawing of the silicon wafer.

There are different sources of stress, which can lead to brittle failures. One among these is the bending stress, which arises while mounting modules with a bowed base plate onto a flat heat sink. This failure cause

is less frequent in advanced modules since the uncontrolled bimetallic warpage of the base plate is reduced by using partitioned ceramic substrates and by using bow-shaped base plates. *Figure 2.12b* shows an unusual horizontal crack in the sub-surface region of an *IGBT* chip, which developed very likely as consequence of the peeling stress arising when mounting a module with an excessively convex base plate.

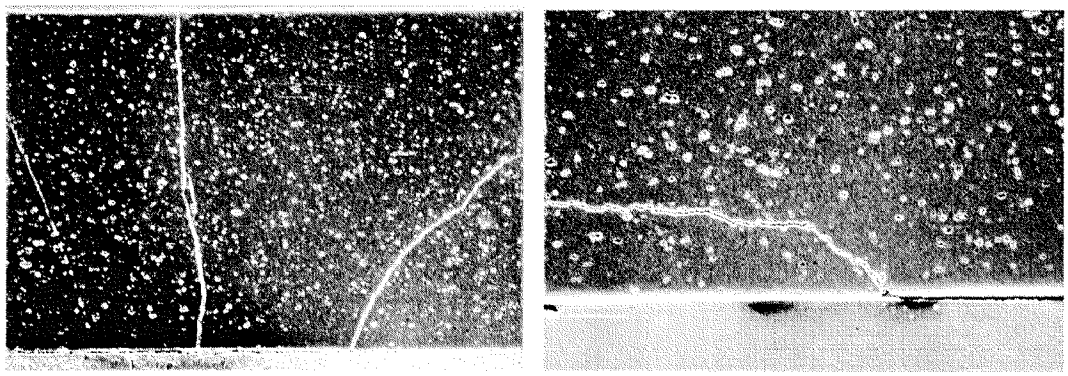


Figure 2.13 (a) Vertical crack within an Al_2O_3 ceramic substrate, due bending stresses (Micro-section, SEM image, 400x). (b) Crack within an Al_2O_3 ceramic substrate initiated from an inhomogeneity in the solder layer (Micro-section, SEM image, 600x).

Figure 2.13a shows the vertical cracks caused across an Al_2O_3 ceramic substrate by the horizontal tensile stress produced by the same failure cause as in previous case. In *Figure 2.13b* a similar crack is represented but propagating from the border of a large void within the solder layer between the ceramic substrate and the base plate. Cracks across the ceramic substrate are particularly insidious, because they can transform with time into insulation failures, which can dramatically impair the partial discharge properties of a multichip module.

Because of the conservative design of the compliant layers, brittle cracking of the silicon chip and of the ceramic substrate due to thermomechanical mismatch only is unusual in advanced *IGBT* multichip modules. However, thermomechanics may concur with some pre-existing stresses in initiating and propagating the fracture. This is the case of *extreme thermal shocks*, where the thermal transient can be as fast, that it cannot be followed by the stress relaxation through the plastic deformation of the compliant layers.

2.8 Corrosion

Corrosion of aluminum is a well-known failure mechanism since the early times of microelectronics. When pure aluminum (*e.g.* bond wires) is exposed to an oxygen containing atmosphere, a thin native Al_2O_3 surface layer is grown that passivates the metal. Aluminum is self-passivating also in pure water, where the native aluminum oxide is converted into a hardly soluble layer of aluminum hydroxide $Al(OH)_3$. When exposed to other solutions, *aluminum hydroxide* is amphotheric, *i.e.* it is dissolved both by strong acids (*e.g.* phosphoric, hydrofluoric and hydrochloric acid) and by strong bases (*e.g.* potassium hydroxide). This step is followed by the exposition of the bare aluminum surface to further chemical or electrochemical attacks. In converse, strong oxidizing agents (*e.g.* nitric acid) leave $Al(OH)_3$ unaffected. In presence of an electrolytic or of a galvanic cell aluminum is corroded according to the corresponding *redox reaction* [21]. The corrosion immunity of the aluminum as a function of the pH of the electrolyte and as function of the voltage applied in an electrolytic cell is described by the different regions of the related *Pourbaix* diagram [22].

In failure analysis both anodic and cathodic aluminum corrosion are found. *Anodic corrosion* occurs in electrolytic and in galvanic cells in presence of halides (*e.g.* chloride and bromide) with a two-steps reaction. The *aluminum hydroxide* passivation is firstly made soluble in the electrolyte by the reaction

$$Al(OH)_3 + Cl^- \rightarrow Al(OH)_2 Cl + OH^-$$

Aluminum chloride is formed after exposing the bare aluminum

$$Al + 3Cl^- \rightarrow AlCl_3^- + 3e^-$$

Finally, the *chloride* dissociates and enters again in solution

$$AlCl_3^- + 3H_2O \rightarrow Al(OH)_3 + 3Cl^- + 3H^+$$

Usually, the main source of chlorine contaminants are process residuals, and in particular residuals of halides activated fluxes, used for improving the wettability before of the surfaces to be soldered.

Cathodic corrosion mainly affect devices, which make use of phosphosilicated glasses as passivation or as insulating layers. If the

phosphor doping exceeds 5%, it can be hydrolized and can form *phosphoric acid*, which corrodes the metallization. However, this last failure mechanism is not expected to play a dominant role in *IGBT* devices.

However, following galvanic corrosion mechanisms have been observed to attack in the different ways the metallic components of a module. The *bimetallic corrosion* is caused by the difference in the electrochemical potential associated with two dissimilar materials. This results into the preferential attack of the material with the higher standard potential (less noble, anodic). The *thermogalvanic corrosion* results from a galvanic cell caused by a thermal gradient. Also in this case anodic and cathodic areas are formed. Galvanic corrosion can also occur when a *concentration cell* forms on the surface of a metal exposed to an electrolyte varying in composition or concentration. Typical concentration cells are *oxygen cells,* in which the corrosion is faster at those locations with the lower oxygen concentration. *Pitting corrosion* occurs when the passivation layer breaks down locally. The surface that is exposed acts as an anode, while the passivated metal plays the role of the cathode. The corrosion due to the electrolyte within the unpassivated area causes a localized attack resulting into a pit. Pitting corrosion is commonly produced by halides (especially chloride). *Stress corrosion* cracking is the cracking of a material produced by the combined action of corrosion and tensile stress. This stress can be either due to an external load, or due to the residual stresses in the metal (*e.g.* by wire welding). The resulting microcracks can be both intergranular or transgranular. *Dealloying* is the selective removal of one element from a solid alloy by corrosion. It can be observed in aluminum metalizations, which include copper precipitates.

The identification of which driving force is promoting corrosion in *IGBT* multichip modules is a quite complex issue, since several causes may concur to the failure. In fact, in *IGBT* packages one is faced with multiple contamination sources, with different metals and alloys, with temperature gradients, as well with static and periodic mechanical stresses. Furthermore, the active devices and the bond wires are embedded in silicon gel, whose influence on the corrosion is not completely understood.

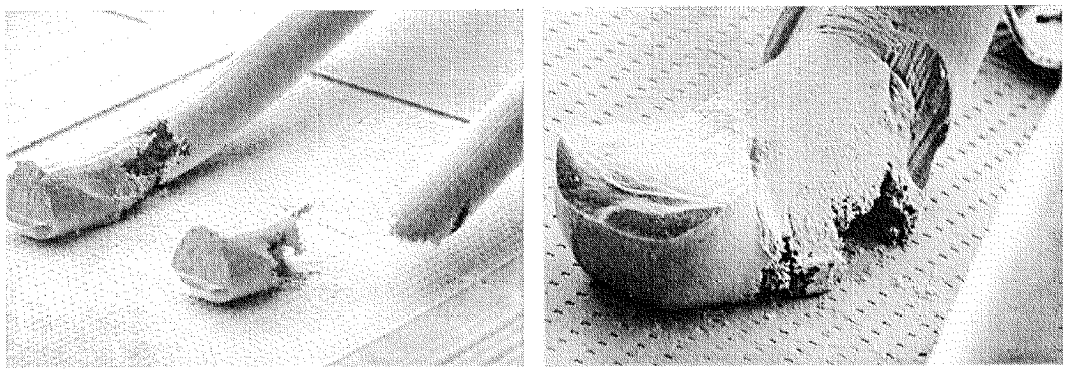


Figure 2.14 (a) Rupture of emitter bond wires due to stress corrosion (SEM image, 30x). (b) Detail of a corroded emitter bond wire (SEM image, 80x).

Figure 2.14a and *2.14b* show aluminum bond wires with no strain buffer that have been corroded at different grades during power cycle tests, which lasted over one million of cycles. This kind of corrosion has been encountered during power cycles performed at low voltage (typically *8V*), as well during lifetime tests at high voltage. The corroded areas were mainly located at those sites of the bond where the wire suffered the most severe deformation, at the heel of the bond wire, and at the top of the wire loop.

These corrosion events have been observed to occur in conjunction with the local formation of gaseous inclusions within the silicone gel (*Figure 2.15b*) that can be sometimes noticed during high-temperature operation of the devices. After package opening by wet chemistry, no corrosion by-products are left at the attacked locations.

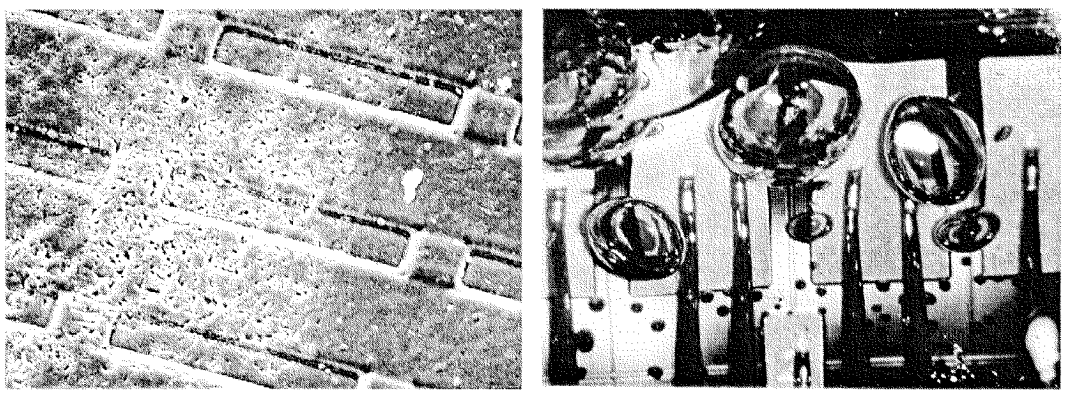


Figure 2.15 (a) Corroded emitter bond pad close to an emitter bond wire (SEM image, 160x). (b) Formation of gaseous inclusions into the silicone gel during power cycling (Optical image, 8x).

This combination of symptoms leads to the conclusion that the observed bond wire corrosion is strongly correlated with the mechanical stresses, which arise either due the thermomechanical cycling, or due to residual deformation stresses in the bond. The absence of reaction by-products and the corrugated surface of the corroded bond wires, may indicate that the corrosion occurs at the grain boundaries of the aluminum. Once completely separated the grains get loose and are removed during package opening. In summary, these indications are compatible with the stress corrosion failure mechanism. Intergranular corrosion is observed also to occur in the adjacent aluminum metalization, but with a less destructive effect than in the much thicker bond wires. This is probably due either to the lower mechanical stress, or to the beneficial effect of the additional alloying elements (silicon and copper). The nature and the source of the contamination are not completely understood. The corrosion could be due to the presence of chlorides originated either from process residuals, or by thermal segregation of the silicone gel.

The incidence of this failure mechanism has been mitigated by more effective cleaning processes after assembly, by the control of the water content in the silicone gel, and by the use of corrosion-hardened bond wires (including nickel).

2.9 Solder fatigue and solder voids

A main failure mechanism of *IGBT* multichip modules is associated with the thermomechanical 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* [26]. In fact, at this location one finds the worst mismatch in the *CTEs*, the maximum temperature swing combined with the largest lateral dimensions (*see Table 2.1* and *Figure 2.3a*). 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.

Voids – 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. Furthermore, since the heat flow within an *IGBT* module is almost one-dimensional, when a relatively large void is present in a solder layer, the heat must flow around it by creating a large local temperature gradient such that the heat dissipation performances of the assembly are degraded. On the contrary, if the large void is broken up into many smaller voids, the perturbation to the heat flow is less evident and has a much smaller impact on the overall thermal resistance of the multilayer. Critical sizes and the most critical sites of contiguous voids in power devices have been investigated experimentally and by numerical simulation in [20].

Since *IGBT* are vertical devices the die attach has to provide at the same time an efficient thermal and electrical conduction path. Therefore, the most insidious voids within the die attach are those, which hinder the thermal flux to the heat sink without inducing any noticeable reduction of the current distribution within the semiconductor. They are for instance edge cracks or shallow voids and delaminations at the interface with the ceramic substrate.

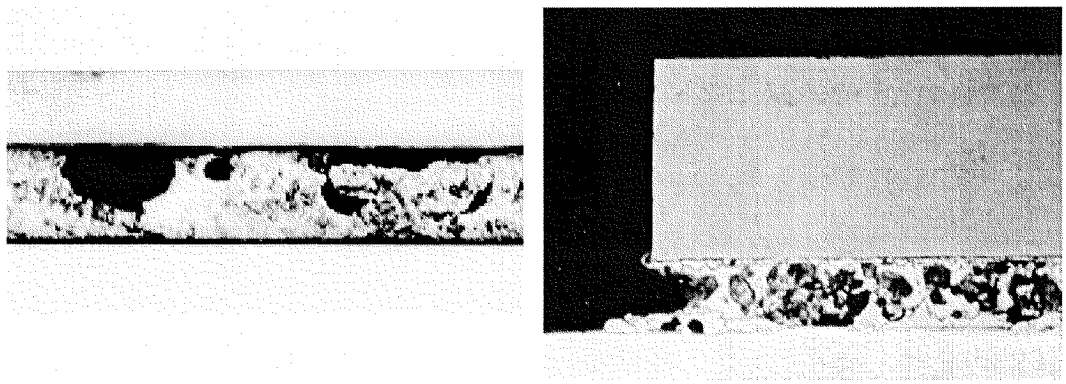


Figure 2.16 (a) Voids in the solder between ceramic substrate and base plate (Microsection, SEM image, 100x). (b) Solder grain coarsening in a die attach of an IGBT mounted on Al_2O_3 after thermal cycles from –20°C to 125°C (Microsection, optical image, 20x).

In advanced assembly processes, special care is taken to avoid the formation of gaseous inclusions within solder layers, by using *e.g.* vacuum ovens and clean processes. During the packaging phase the control of the temperatures profiles during soldering and during the successive annealing steps is essential for avoiding an excessive growth of brittle intermetallic layers. Nevertheless, the quality of solder joints between large size plates is still considered a critical issue. Examples of

large voids in solders are presented in *Figure 2.16a* and in *Figure 3.10* (*Chapter 3*).

Fatigue – The most frequent solders used in advanced *IGBT* multichip modules are based on *tin-silver*, *indium*, or *tin-lead alloys*. They have excellent electrical properties and as soft solders they exhibit good flow characteristics. For sake of simplicity, solders are often modeled as a single homogeneous phase. However, when *copper* is soldered for example with a standard *lead-tin* alloy, the bond is mainly provided through the formation of a Cu_5Sn_6 intermetallic phase located close to the *copper plate* [23]. Two additional distinct phases, one tin-rich and one lead-rich, are formed in the central part of the solder layer upon solidification. During power cycling, these phases coarsen rapidly due to the high homologous temperature at which the alloy is operated. An example, of severe solder coarsening of a tin-lead alloy in a die-attach is shown *Figure 2.16b*. Since the copper phase is much more brittle than the tin-lead phases, thermomechanic fatigue cracks often propagate within the copper rich intermetallic. Due to the larger *CTE* mismatch and to the higher temperature, fatigue cracks are found preferably in the vicinity of the intermetallic layer immediately below the ceramic substrate. This situation is clearly shown in the scanning acoustic microscopy images in *Figure 3.11a* and *3.11b* (*Chapter 3*). Metallographic preparations have shown [24] that cracks initiate as expected at the border of the solder joint, where the shear stress reaches its maximum. Additionally, crack formation is highly promoted by the presence of sharp angles at the edges of the ceramic substrate. This problem requires a dedicated engineering of the *solder fillets*. Thermal cycle tests with ΔT up to $100°C$ have shown that the number of cycles to the failure of the solder between ceramic substrate and base plate just weakly depends on the temperature swing. Because of the very severe conditions imposed by this kind of accelerated test, the results can hardly be extrapolated to real operating conditions. In fact, with a junction temperature swing of $100°C$ and the typical material constants listed in *Table 2.1*, the expected plastic strain can be estimated in about $50\ \mu m$. Since this value has the same order of magnitude of the thickness of the solder layer, it can be expected that the failure mechanism leading to the degradation of the solder during the accelerated test is not representative for the lower temperatures encountered in field applications. In fact, during field operation, the most critical system (Al_2O_3 on *copper*) rarely experiences temperature swings over $30°C$. Finally, in recent experiments, severe thermal cycles have shown some adhesion failures of *AlN* ceramic substrates due to the peeling of the copper metalization [25]. However, this unusual failure mechanism is not expected to impact the field reliability of *IGBT* devices.

The number of cycles to the failure of large solder joints due to thermomechanical fatigue can be simply modeled by a *Coffin-Manson*-like power law of the form

$$N_f = 0.5\left(\frac{L\,\Delta\alpha\,\Delta T}{\gamma\, x}\right)^{\frac{1}{c}} \tag{2.3}$$

In *Equation 2.3* L represents the typical lateral size of the solder joint, $\Delta\alpha$ the CTE mismatch between the upper and the lower plate, ΔT the temperature swing, c is the fatigue exponent. x and γ are the thickness and the ductility factor of the solder, respectively. The values $\gamma = 1.1$, and $c = -0.49$ are conservative engineering estimates usually encountered for the *In-70% Pb-30%, Sn-40% Pb-60%,* and *Sn-10% Pb-90%* solder alloys [6,14]. From *Equation 2.3* one can also derive some simple design rules for minimizing the fatigue of solder joints. In fact, it can be easily seen that the lifetime is improved by reducing the size of the solder joint, by matching the CTE of the materials, by reducing the edge voids, and by increasing the thickness of the solder (compatibly with the requirements imposed by the thermal resistance).

2.10 Burnout 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 failure cause occurring randomly. 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 $10kA/\mu s$, leading to a current maximum in the *100 kA* range and to a decay within $100\mu s$ [27]. 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 module [28] 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 [29].

They are many system, 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* [30], overheating due to the degradation of the thermal resistance, dielectric breakdown, and *cosmic ray* irradiation.

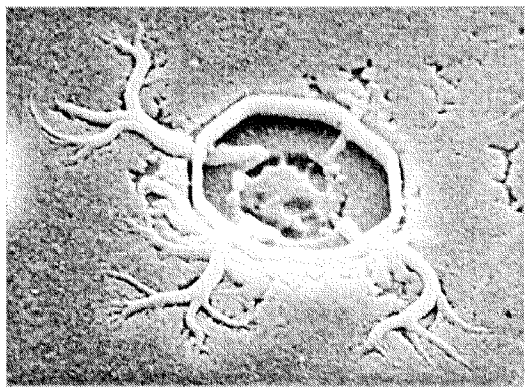


Figure 2.17 Polysilicon filamentation in an IGBT as a consequence of a short circuit between gate and emitter, due to pre-damaged insulation during wire bonding (SEM image, 2000x).

Figure 2.17 shows a characteristic *polysilicon filamentation*, which occurred during high voltage testing of an *IGBT* module, because of the short circuit between gate and emitter. The root cause of this failure is a pre-damage introduced in the polyoxide by the bonding tool. The relatively small damage produced is due both to the very localized fracture in the dielectric, as well by the short circuit detection and consequent current limitation in the tester.

Since the investigation of the root causes associated with system design and device application related problems are outside the scope of this work, they will be not considered in more detail. On the contrary, we will briefly discuss a failure mechanism, which is inherent with *IGBT* devices, *i.e. latch up*. This phenomenon is of special relevance, because most of

the root causes mentioned above activate this mechanism, such that it 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) manifest 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.

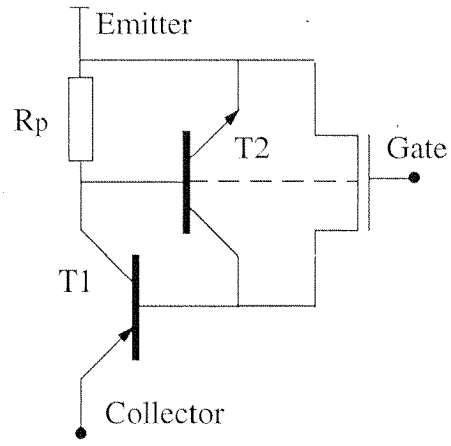


Figure 2.18 Simplified equivalent circuit of an IGBT

Figure 2.18 represents the simplified equivalent circuit of an *IGBT*, which takes into account just quasi-static effects. Under normal forward operating conditions the voltage drop caused by the collector current over the *p emitter* diffusion (represented by the resistor R_p) is almost negligible. Therefore, the parasitic bipolar transistor *T2* is in the non-conducting state and the *IGBT* device is controlled by the electron flow injected into the base of *T1* through the *MOS* transistor. On the contrary, if the collector current reaches the critical value for which the voltage across R_p exceeds *0.6 V*, *T2* enters in conduction and provides the base current to *T1*. Since the additional base injection turns into an increase of the collector current, this effect is regenerative and leads to the thermal destruction of the device, which is not controlled by the gate voltage anymore. This simple quasi-static model illustrates how latch up may arise in *n-channel IGBTs*, while forcing the collector current to increase. This situation can occur in an *IGBT* module, if the number of operating cells within a module is reduced with time, due to a degradation

mechanism, as for example bond wire lift off. Nevertheless, more complex physical and numerical models are required for taking into account all dynamic effects, which concur in triggering this failure mechanism [31,32].

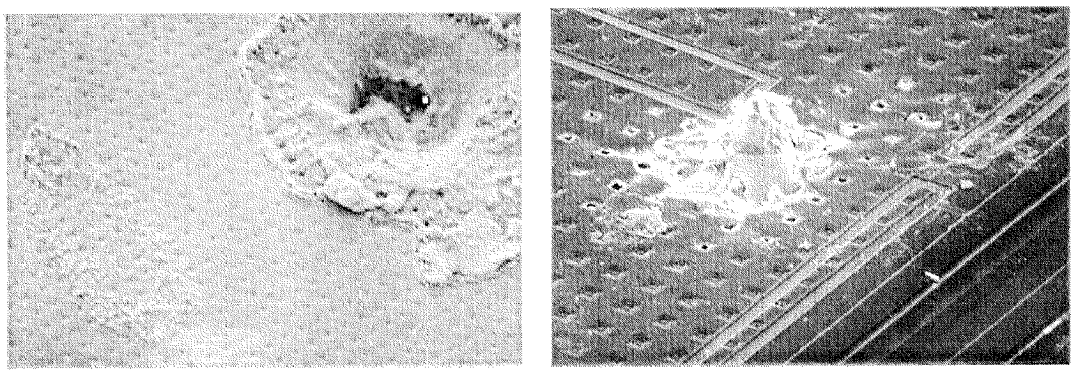


Figure 2.19 (a) Melted pit in an IGBT due to a latch up event, which occurred in conjunction with bond wire lift off (SEM image, 50x). (b) Same effects than in the previous image but localized to some few cells (SEM image, 90x).

Figure 2.19a represents a melted pit on an *IGBT* device, which resulted from the latch up event during a *lifetime test* at high voltage. During failure analysis the module showed clear evidence of distributed bond wire lift off. The melted path usually crosses the *IGBT* chip down to the die attach and a silicon-solder alloy is formed. Craters have been often observed in immediate vicinity of melted emitter bond wires, indicating that those bond wires were still attached shortly before the latch up event. This fact also suggests that the local current density is increased also in consideration of the sheet resistance degradation due to the reconstruction of the metalization. *Figure 2.19b*, shows the effect of a latch up event, which occurred in an *IGBT* device during a *long-term frequency test.* The latch up interested just some few cells and the melted area has been kept under control by external limitation of the collector current.

Cases have been reported [33], where catastrophic burnout can be caused by *second breakdown.* However, since second breakdown is basically related to avalanche carrier multiplication at high electron current regimes, it mainly affects *p-channel IGBT* devices.

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 radiations* [34]. A universal curve has been derived from the *Zeller* model [34], which predicts the failure rate of bipolar devices (thyristors, *GTO*, diodes) as function of the electric field parameter

$$S = \sqrt{\frac{V}{\rho}} \tag{2.4}$$

where V is the applied voltage in *Volts* and ρ is the *n base* resistivity in *ohm-cm*. *IGBT* devices show an increased sensitivity to cosmic ray in respect to thyristors, *GTO*, and diodes. In fact, the measured failure rate exceeds by at least an order of magnitude the value predicted by the universal curve. Furthermore, a design-dependent threshold S_{crit} is observed, such that for $S < S_{crit}$ the failure rate abruptly decreases to zero and that decreases with increasing voltage. This indicates that *1200 V* devices are inherently robust with respect to the failure mechanism, while the failure rate of *3500 V IGBTs* (or higher) is more prudently predicted by neglecting the effect of S_{crit} [35].

Chapter 3

Failure analysis techniques and procedures for IGBT devices

3.1 Introduction

The scope of the failure analysis is to investigate the physical and/or the chemical causes of a failure, in order to design those technological countermeasures, which enable to avoid such a failure in future product generations. A failure analysis has to include four relevant aspects: identification of the *failure mode(s)*, identification *the failure mechanism(s)*, identification the *root cause* of the failure, and the technological *corrective actions* for avoiding the failure. Operatively, it consists of five sequential phases, which are intended to provide the required information without affecting or preempting a subsequent step. They are failure detection and description, non-destructive failure analysis, semi-destructive failure analysis, destructive failure analysis, and failure mechanism analysis. At the end of the whole procedure, this information is summarized in a failure analysis report.

The analytical techniques mentioned in the following are specially focused on *IGBT* devices. They have been extensively used in our laboratory, and they have shown to be successful in most cases.

3.2 Parametric and functional tests

Curve tracer analysis is the most important technique for assessing the electrical condition and integrity of a device. This is the first electrical evaluation used for establishing the present condition of the device under investigation. For example, a low current/voltage can provide information about either the integrity of the gate oxide, or about the leakage current flowing between the different terminals of the *IGBT*. This is a non-destructive technique, which can also be referenced to a known good device for comparison. In any case, it has to be always considered that in a multichip package, *IGBTs* are either mounted in parallel or in series with or without an anti-parallel freewheeling diode. In the simplest case of a single device without freewheeling diode, the characteristic between gate and emitter as well between gate and collector is an open circuit. Further, if the gate is grounded and an *AC voltage* is applied between emitter and collector, one can inspect the characteristic of the reverse-biased *p^+-emitter n-substrate junction* and of the reverse-biased *n-substrate p^+-collector junction*, alternatively. Finally, if the gate is set beyond the threshold voltage, the *AC characteristic* of the *IGBT* is that of a *pn-junction* with a series resistance. The usual failure modes, which can be detected in forward bias, are high series resistance, soft diode knee, and shunt resistor. The possible failure modes in reverse bias are reduced breakdown voltage, soft breakdown, channeling, breakdown walkout, and jittering [40]. An additional parameter to be monitored is the *threshold voltage* of the whole module and of each *IGBT* chip. An interesting case of parameter variation after thermal cycles is V_{CE}. An increase of V_{CE} measured at constant temperature and at low I_{CE} can be an indicator of the occurrence of bond wire lift off. On the contrary, an increase of V_{CE}, measured at high values of I_{CE} and at a constant ambient temperature can be the symptom of the degradation of the thermal resistance.

Additional failure modes, which can be encountered, are instabilities and intermittencies. Instabilities can be either functional or parametric failures. They result into time-dependent characteristics, which can slowly fade away and, which are not necessarily reproducible under the same operating conditions. Instabilities can be due to surface and oxide contamination, humidity, parasitics, interface traps, *etc.*. Usually, the

cause of instability cannot be recognized directly. This is the case for example of an ionic contamination due to alkali ions, which can be very insidious even at concentrations far below the resolution limits of the most common analytical techniques. The usual procedure followed in this case consists into a sequence of actions intended either to remove or to dilute the causes of the instability. These operations are: *external package cleaning* with different solvents for organic or ionic contaminations, *high-temperature storage* (typically *16 hours* at *150°C*, under bias or not), package opening and *surface cleaning*, selective *removal of the passivation* and of insulating layers, and mechanical *partitioning of the circuit*. The removal of the instability cause has to be checked after each one of previous operations by electrical test of the device. Intermittent failures can occur in conjunction either with an applied external mechanical stress or just after a prolonged operating time, and usually disappear if the stress is removed or if the temperature is lowered (or increased). There can be several causes associated with intermittent failures, like thermomechanical mismatch or fatigue, microcracks either in the bond wires, connections or in the silicon die, temperature-dependent triggering of parasitics, *etc.*. Although these mechanisms can be almost detected by visual inspection, special attention has to be paid when verifying the occurrence of the failure. In this case, an electrical test while cooling, heating, tipping, or vibrating the device can be very helpful..

3.3 Encapsulation

The overall scope of encapsulation is to expose the semiconductor chips and the interconnections without affecting the functionality of the device. As for usual microelectronic circuits, the package has to provide to *IGBT* devices a barrier against mechanical and chemical attacks, the electrical interface with the outside world, as well a proper thermal contact for the thermal management. For high-voltage devices, the package has to provide in addition the necessary dielectric strength for avoiding partial discharge. For this reason, the removal of critical components of the module (*e.g.* the silicone gel) may completely impair the high-voltage capabilities of the device under test. Nevertheless, almost all electrical characterizations during the failure analysis (excepted the problems involving the blocking voltage of the device) are performed at voltages lower than *50 V*. Before opening, the package should be submitted to external inspection. Special attention must be paid to the presence of

cracks in the outer shell, and in particular to the status of the *thermal grease* layer on the bottom side of the base plate. In fact, an uneven distribution of the grease residuals can be an indicator that the failure is due to improper thermal dissipation. On the other side, cracks within the case may suggest that the malfunction of the device resulted into a catastrophic explosion of the module. Such cracks are sometimes due to excessive mechanical stress caused by vibration. If it is the case, they occur in conjunction with the complete or partial release of the terminals as a consequence of *low-cycle fatigue* of the solder joints.

There are a number of module designs, which are especially used for applications other than traction. For this reason, before package opening a thorough knowledge of the internal structure is necessary either by *x-ray* microscopy, or by using a sacrificial device. The most common packages encountered in traction applications are represented in *Figure 3.1.*

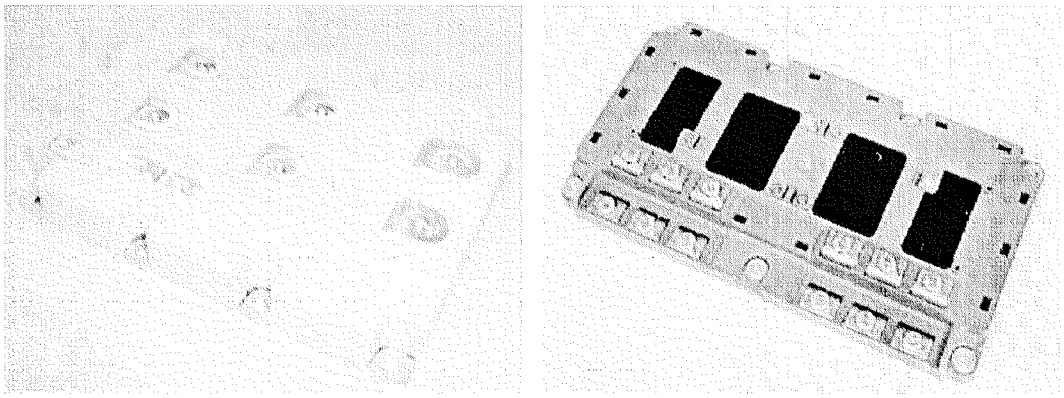


Figure 3.1 (a) Industrial standard E2 module including 24 IGBT chips, 12 diodes, and with a footprint of 140 on 190 mm. (b) Flip module [28] with a footprint 140 on 260 mm including 24 IGBT chips and 8 diodes.

The *E2* module (*Figure 3.1a*) is a relatively traditional design. *IGBT* chips and diodes are soldered onto multiple *direct-bonded copper* ceramic substrates, whose metallization is patterned in order to provide through aluminum bond wires the emitter, the gate and the collector contacts to each chip (*see Figure 3.2b*). The chips are paralleled by bridges and the outside contacts are realized through wide and thick copper strips, which are soldered, on the metallization of the ceramic substrates. The ceramic substrates are soldered onto the base plate. Chips, bond wires, ceramic substrates, and terminals are embedded in silicone gel. The mechanical stability of the terminals is provided by an epoxy mold on the top of the silicon gel (*Figure 3.2a*) and by an external case of

thermoplastic material. The *Flip* module [28] has a different three-dimensional design. Also in this case, the *IGBTs* and diodes are soldered onto metallized ceramic substrates and then soldered onto the base plate. However, in *Flip* modules the bond wires are not bonded onto the metallization of the ceramic substrate. In fact, emitter, gate, and collector contacts are provided by a stack of insulated metal plates, which also realize the outer terminals. The cavities within the insulated stacks, which host the *IGBTs* and the diodes are filled with silicon gel. As it can be seen from *Figure 3.1b*, due to safety reasons, the outer case presents large windows in correspondence with these cavities. For this reason *IGBT* and diode chips can be directly accessed without the need of removing hard molds.

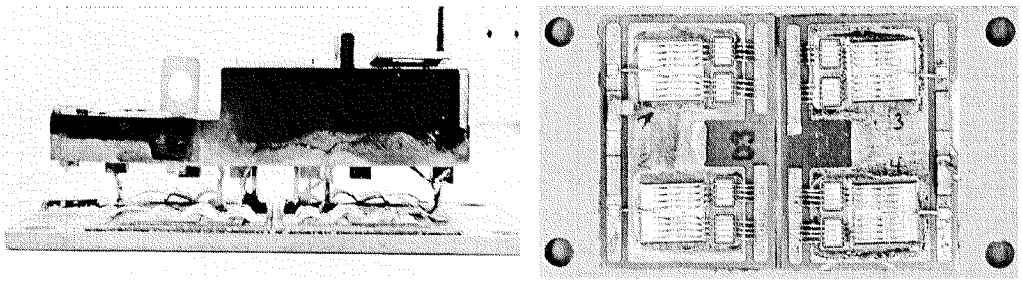


Figure 3.2 (a) Side view of an IGBT module rated for 300 A after partial mechanical removal of the outer case and complete dissolution of the silicone gel (0.2x). (b) Top view of the same IGBT module after sawing of the terminals and removal of the epoxy mold (0.3x)

The first phase of the encapsulation procedure consists into the partial removal by mechanical means of the outer shell, such that the silicone gel and the epoxy mold (in the case of *E2-like* modules) are exposed at the sides of the module. The next step is the selective removal of the silicone layer by *wet chemistry*. In order to accelerate the process, some holes can be drilled from the tops side through the thermoplastic case and the epoxy mold. Once, the silicon gel is reasonably exposed the full module is immersed into an organic solvent. Very good results have been achieved with the commercial product *Panasolve 215* (containing *alkyle glycol* [57]) at a temperature of *65°C*. The dissolution of the silicone gel is usually very slow and may take many hours. However, if the solvent is not contaminated with water, *Panasolve 215* is very selective over the aluminum metallization and just slightly etches the copper films directly bonded on ceramic substrates. The final result of this process is represented for a *300 A* rated module in *Figure 3.2a*. This operation can be highly accelerated by cutting the copper terminal posts with a manual

or a diamond saw, as soon they are visible, in order to remove the epoxy mold. An alternative solvent to *Panasolve 215* is the *Losolin IV* at *80°C*. This product is normally used as a resist stripper and contains *alkyaryl sulphonic acid* and a high boiling point organic solvent. It removes most of the silicon gel within *30 minutes* without etching the metals. After catastrophic failures, it happens often that carbonized residuals are still left on the surface of the device. In general they are very difficult to remove. However, local dropping of *red fuming nitric* acid at *60°C* can be helpful. If the silicon chips are concerned, one could tray to remove small carbonized particles by underetching the passivation layers. After complete removal of the gel and eventually of the bond wire coating, the device is neutralized in running water and finally dried by vacuum storage. The result of the encapsulated procedure is shown in *Figure 3.2b*. For further electrical characterization suitable contacts must be soldered on the stub of the electrical terminals. It is finally noted, that if *IGBTs* and diodes are provided with *molybdenum strain buffer plates*, the visual inspection of the chip surface by non destructive techniques is almost impossible.

3.4 Microscopy Techniques

3.4.1 Optical microscopy

Optical microscopy in the visible range is the most common technique for failure analysis. It can be used either as a non-destructive imaging tool or as a real analytical method after proper sample preparation. The main limitations of the traditional (*far field*) optical microscopy are the lateral resolution and the depth of field, since they cannot be achieved at the same time. Recently, such limitations have been partially removed by advanced techniques like *confocal microscopy* or the *near field* optical microscopy. Nevertheless, the equipment required is still not largely used for failure analysis; additionally it not very versatile in the case of very large multichip modules. *Infrared microscopy* is normally used for inspecting the active area of *IGBT* chips from the backside after selective etch of the backside metallization and eventually after thinning of the silicon chip down to *50-80 μm*. Of course, the lateral resolution of this technique is limited to several micrometers, due to the wavelength of the incident light.

Stereoscopic microscopy is normally used either for inspecting reliefs or for simple micro-surgical operations that is when a large depth of field and a three-dimensional view of the sample are required. Stereoscopic microscopes can reach a maximum magnification in the *100x* range, only. However, magnifications up to *1000x* can be attained by *inspection microscopes*. A critical issue of inspection microscopes for failure analysis of multichip modules is the working distance of the objectives. In fact, the objects to be imaged are often placed within cavities, and they cannot be focused with usual objectives.

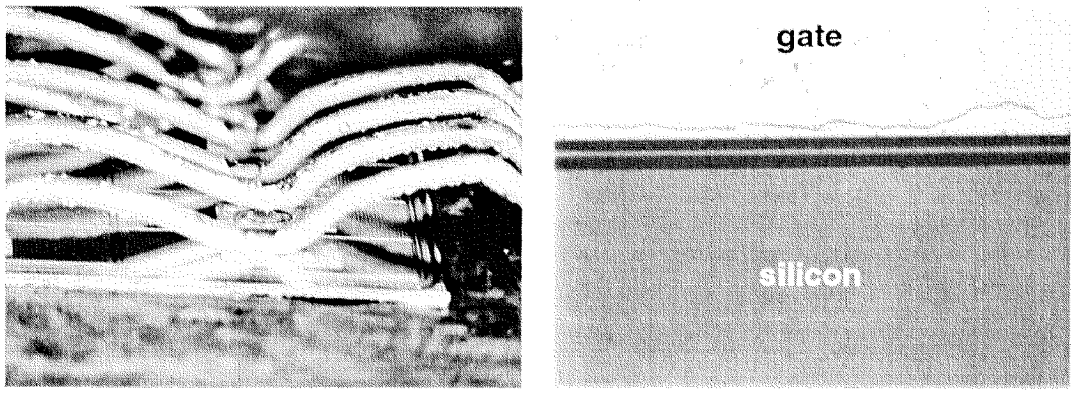


Figure 3.3 (a) Optical stereoscopic image of lifted off emitter bond wires (30x). (b) Bright field microscopy image of a microsection through the gate bond wire of a device submitted to power cycles (300x). The preparation shows the propagation of the crack within the aluminum bond wire.

High-magnification objectives with working distances up to several millimeters are available, however such a working distance can only be obtained on expenses of the numerical aperture, that is on expenses of the maximum lateral resolution. Three main operating modes are known for the inspection of semiconductor samples: the *bright field*, the *dark field*, and the *Nomarsky differential contrast*. In the bright field microscopy, which is the most common technique, the illumination path is coaxial with the objective. The image is formed through the difference in absorption, reflectivity, refraction index, and thickness of the probe. In the dark field technique the sample is illuminated under a glazing angle. Perfectly flat and reflecting objects are imaged as dark surfaces. On the contrary, objects, which scatter the impinging light, produces bright images. Thus, edges, particles, and rough surfaces produce brilliant features onto a black background. For performing the *Nomarsky* differential contrast technique the microscope must be equipped with a

polarizing filter and with a *Wollaston* prism, which splits the illuminating light into two rays. Some of the rays pass through the specimen where they are retarded. From there the rays are collected by the objective which is equipped with a second *Wollaston* prism. The rays are then recombined and allowed to interfere. The *Nomarsky* technique is mandatory in the case of transparent samples (*e.g.* thin oxide grown onto decorated samples), or when submicron height differences in the sample have to be displayed (*e.g.* delineated samples). *Figure 3.3a* shows the emitter bond wire of an *IGBT* device, which lifted off after a power cycling test. This image shows the depth of field provided by a stereoscopic microscope. The failure analysis of the bond wire lift off mechanism is often performed by stereoscopic microscopy. Practically, after package opening and complete removal of the silicone gel, the device is observed under the microscope, while gently blowing onto the bond wires with a compressed air flow. Bond wires, which are completely disconnected, start to vibrate and can be quickly identified, and the information used for statistical purposes. *Figure 3.3b* has been acquired by bright field microscopy after microsectioning an *IGBT* device, which was submitted to a power cycling test. The microsection shows the silicon substrate, the thick oxide on the bottom of the gate contact pad, and the bond between the aluminum metallization and the wire. This sample also shows that the microcrack, which results from the thermomechanical mismatch between the aluminum bond wire and the silicon substrate, clearly propagates within the wire and not at the interface, where it would be located, if the problem were due to a loss of adhesion.

3.4.2 Liquid crystal microthermography

The scope of *liquid crystal microthermography* is to locate hot spots at the chip surface, due to enhanced local power dissipation. This technique makes use of the fact that certain liquid crystals exhibit a transition from the *nematic* to the *isotropic* phase at a very precise temperature (*clearing temperature*). For thermography applications, a thin layer of a selected liquid crystal is deposed onto the surface of the device, which is operated in such a way to reproduce the failure mode to be investigated. In order to increase the sensitivity of the technique, the device is heated up by a thermochuck close to the clearing temperature of the liquid crystal. After having electrically connected the device, if the local power dissipation reaches such a level that the local temperature exceeds the clearing

temperature, a local phase transition occurs. This transition is observed with a microscope by taking advantage of the fact that in the nematic phase the liquid crystals have an anisotropic and birefractive optical behavior, while, when exceeding the clearing temperature, they exhibit isotropic properties. Thus, when linearly polarized light is directed on the liquid crystal thin layer and the reflected light is analyzed by a cross-polarized filter, regions that are in the nematic phase will appear iridescent, while regions in the isotropic phase will appear dark. The fact of using a microscope enables to locate hot spots with a lateral resolution of less than *5 μm*.

The liquid crystals, which can be used for *IGBT* devices and the related clearing temperatures, are listed in *Table 3.1*.

Table 3.1 Liquid crystals for microthermography use

Liquid crystal	Clear. Temp.
$C_{15}H_{19}N$	35.3°C
$C_{20}H_{21}NO_2$	47.5°C
TM 75 A™	53.3°C
$C_{15}H_{20}N_2O_2$	73.2°C

The use of the liquid crystal microthermography is recommended, when the power is mainly dissipated at the defect location. On the contrary, when the failure is produced in the conducting mode of the *IGBT* device (static or dynamic) other techniques, like infrared thermography, should be preferred. Liquid crystal microthermography is not suitable if during the analysis the device requires to be operated at voltages higher than *100 V*. In fact, due to the low dielectric strength of liquid crystals, the application of high voltages may result into a partial discharge. Furthermore, most liquid crystals exhibit a voltage sensitivity, which could impair the readability of the thermal map.

In *Figure 3.4*, the $C_{15}H_{19}N$ liquid crystal has been used for detecting the location of a short-circuit between gate and emitter (R_S = *12.5 kΩ*), which occurred in an *IGBT* device during field operation. *Figure 3.4a* shows the sample with the liquid crystal in the nematic phase, *i.e.* the whole surface of the chip is visible. By forcing a short circuit current of *400 μA*, a dark spot appears (*Figure 3.4b*), indicating an increased local power

dissipation, and thus the occurrence of a local phase transition. The detection of hot spots can be improved by using a pulsed current source at a typical frequency of *2 Hz*, such that the heat source appears as a pulsing dot. The sensitivity of the technique strongly depends on the temperature control of the thermochuck.

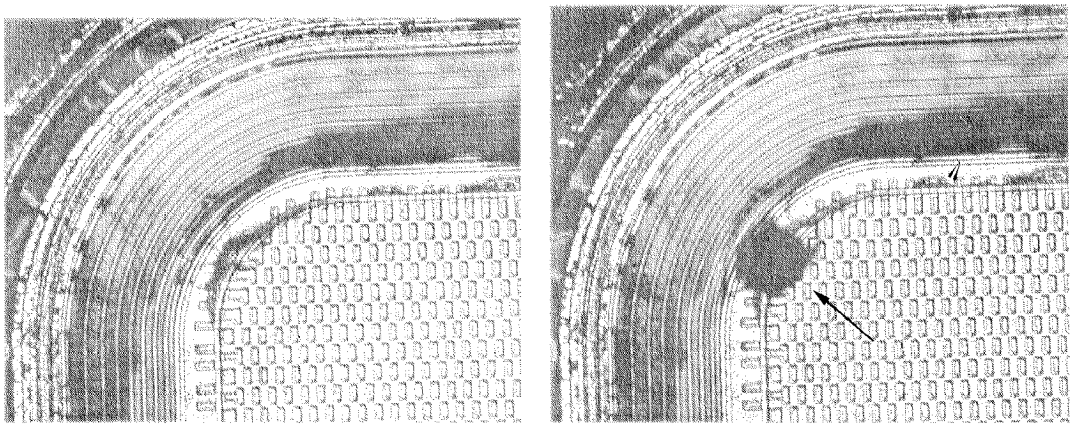


Figure 3.4 (a) IGBT covered with a liquid crystal in the nematic phase (15x). (b) At a dissipated power of 2 mW, the liquid crystal turns locally to the isotropic phase (15x), due to a leakage between gate and emitter.

The best results have been obtained by heating the device *0.5°C* above the clearing temperature of the liquid crystal. When the thermal equilibrium is reached, the surface of the sample is cooled down by a gentle airflow, such that the liquid crystal is driven into the nematic phase. After removal of the airflow, the sample heats up again very slowly towards the thermal equilibrium. As soon the hottest area within the sample crosses the clearing temperature, a dark spot is initiated at that location and propagates across the whole chip surface. By iterating this procedure in order to find out the optimal upper and lower temperatures, sensitivities in the *100 μW* range can be attained. Usually, liquid crystals with low clearing temperatures are preferred, since working at low temperatures minimizes the fluctuations of the surface temperature due to local air convection. The sensitivity of the technique is also strongly dependent on the thickness and on the homogeneity of the liquid crystal film. These properties can be influenced by proper sample preparation techniques. Since during the thermal analysis the device is operated at low current levels, almost all (but one) emitter and gate bond wires can be removed by tweezers. This operation avoids the formation of a meniscus around the bond wire, which could disturb the homogeneity of the liquid crystal film. Furthermore, the wettability of the chip surface can be highly improved by partial removal of the silicon nitride passivation layer with

the solution presented in *Section 3.5.1*, or in any case by avoiding grease contaminations or silicone gel residuals on the die surface. In order to achieve a homogeneous film, liquid crystals are diluted immediately before deposition in highly volatile solvents (*e.g. 1 part* by weight liquid crystal in *10 parts acetone* or *methanol*). As soon the solution is applied to the chip surface, the solvent evaporates, leaving behind a very homogeneous thin film of liquid crystal. The optimum thickness of the layer should be in the *5 μm* range. Two factors, which strongly affect the sensitivity, especially in *IGBT* devices, is the heat spreading and the thermal resistance due to the thick layers (*e.g.* thick oxide, polysilicon, metallization), which can be found between the heat source and the liquid crystal. Unfortunately, there is no technical solution for this problem. Finally, it is essential that the thermochuck provides a very homogeneous surface temperature on the chip. This is the case, if the base plate of the device is mounted onto the thermochuck by the means of a thin layer of thermal grease, in order to reduce and to equalize the contact thermal resistance.

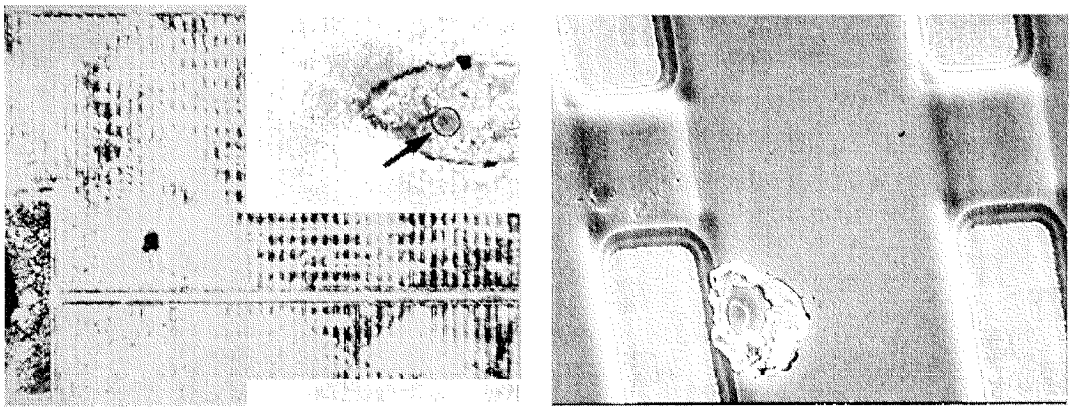


Figure 3.5 (a) Localization of a leakage path between emitter and gate by liquid crystal microthermography (80x). (b) The hot spot is located below an emitter bond wire, and it is due to a mechanical damage of the polyoxide (SEM image 500x).

Figure 3.5 illustrates one among the most insidious cases of latent short circuit between gate and emitter, which occurred during a frequency test on several devices of the same manufacturer. When the liquid crystal analysis has been performed before removing the bond wires, no hot spot could be detected, even at high levels of power dissipation. After carefully removing the emitter bond wires (but one) the failure mode was unaffected, but a hot spot appeared within the footprint of a bond wire (*Figure 3.5a*). Selective etching of the metallization (*Figure 3.5b*) revealed a damaged polyoxide with traces of interdiffusion between

aluminum and polysilicon. Since the failed modules passed the final production test and did not present any problem during the first operating period, one can conclude that the polyoxide was pre-damaged during the assembly phase (probably by a uncalibrated bonding tool). Due to the thermal and thermomechanical stresses, which arise during operation, the microcracks in the polyoxide propagated and the aluminum metallization contacted the polysilicon, such that the pre-existing damage evoluted to a *low-ohmic* short circuit.

3.4.3 Emission microscopy

Emission microscopy is a non-destructive optical technique for failure localization. The most relevant effects, which can be imaged by emission microscopy are junction leakage, contact spiking, hot carriers, junction avalanche, latch-up, oxide current emission, polysilicon filaments, substrate damage. After exposing the chip surface, the device is operated electrically within a dark chamber, in order to reproduce the failure mode to be investigated. The electro-luminescence of the device is acquired by an inspection microscope equipped with a system for light detection, amplification, and for on-line image processing. Traditional emission microscopes image the device onto a photocathode of a microchannel plate. The intensified photoemission map is then imaged onto a phosphor plate, where it is acquired by a video camera. Because the photon count is low, the image from the video camera is averaged and processed through an image capture board. For practical purposes, emission maps are usually electronically superposed onto optical images of the device under investigation. The spectral response of the system depends upon the type of detector is used. Usual emission microscopes are sensitive either in the visible (*400 – 850 nm*), or in the near infrared (*770 – 1500 nm*) range and with a maximum quantum efficiency at about *600 nm*. Spectral analysis can be performed, by inserting *shallow-band* interference filters directly after the objective lens. However, optical filters strongly reduce the transmissivity of the system and can introduce artifacts due to induced fluorescence effects. Further artifacts in spectral resolved measurements can be introduced by the absorption of layers above the emitting site (silicon nitride, polysilicon, and oxide). Recently, equipment has been proposed, which makes use of cooled *CCD arrays* instead of microchannel tubes.

Emission microscopes are usually intended for traditional low-voltage microelectronic applications. Thus, when working with *IGBTs*, special attention must be paid to the galvanic insulation of the acquisition system from the high-voltage bias of the device. Furthermore, extra long working distance objectives are mandatory, in order to avoid sparking through the air even at the maximum magnification (*1000x*).

Sample preparation is a critical issue for the investigation of *IGBT* devices by emission microscopy. In fact, without selective etching (or thinning) the emitter metallization the use of this technique is restricted to the guard ring area (provided that eventual polyimide layers are also removed). Nevertheless, this is sufficient for investigating most instability failures. After removal, thinning, patterning, or windowing of the emitter metallization, the device cannot be operated at the maximum current rating. For instance, if the aluminum layer is completely removed and a semi-transparent gold layer is deposited (*see Section 3.5.4*) no more than *few milliamperes* of emitter current are allowed. There are also some restrictions for the maximum emitter-collector voltage. In fact, the silicone gel must also be completely removed, such that for voltages exceeding *2 kV*, partial discharge phenomena may occur.

An additional approach is backside emission microscopy. It requires the use of infrared optics, infrared detectors, dedicated sample preparation, and special mounting. Sample preparation for backside emission microscopy is cumbersome, since it requires to remove the central portion of the backside metallization (*e.g.* silver on nickel) and to reduce the thickness of the chip down to *80 μm* mechanically or by wet chemistry. Furthermore, unlike in integrated circuits, thinning of the silicon substrate of vertical devices, like *IGBTs*, is a destructive procedure. Thus, this technique can essentially be used for investigating light emission in the channel area at a reduced current injection level.

They are two fundamental mechanisms leading to photoemission in semiconductors, which can be observed by emission microscopy. The are *interband recombination* and *intraband recombination*. Indirect interband recombination provides photons at *1.1 eV*, while intraband transitions generate low-energy phonons (less than *1 eV*). Additionally, *bremsstrahlung* is sometimes invoked for explaining the broad background spectrum, which is often observed.

In forward biased junctions, majority carriers are injected across the depletion region and recombine by an indirect phonon assisted transition. In silicon, the emitted phonons are centered at the bandgap energy of *1.1*

eV (*i.e.* at a wavelength of *1120 nm*). Although detectors designed for the visible range have a low sensitivity in the near infrared tail, such a photoemission can be easily observed either at sufficient injection levels or by increasing the sample temperature.

During reverse bias of a junction the carriers, which cross the depletion region may gain sufficient energy to get hot. A wide spectrum of photon energies arises from the thermalization and recombination of hot carriers; it ranges from the band gap energy up to *3 eV* (*i.e.* at wavelength of *420 nm*). Although the recombination probability of the carriers is by many orders of magnitude lower than for forward biased junctions, photoemission can be observed even if the breakdown current is far below *1 microampere*. This is due to the fact that the power dissipation occurs in a very delimited region and that the main part of the photons are emitted within an energy range, where the quantum efficiency of the detector reaches its maximum. The real nature of the photoemission in a reverse biased junction (and thus also in the pinch-off region of a *MOS* transistor) is still an open issue. Processes, which can be reasonably taken into account, are the radiative transition of holes between the light-hole band and the heavy-hole band, as well phonon-assisted intraband recombination [54].

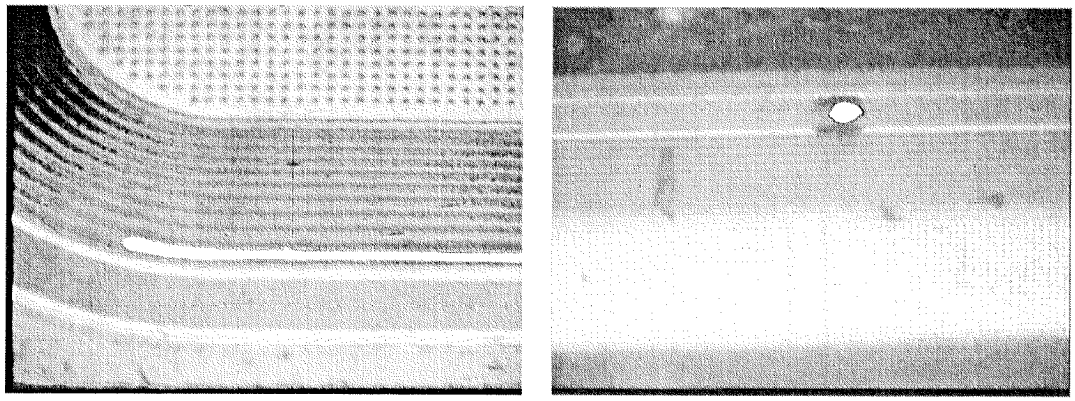


Figure 3.6 (a) Emission image of the guard ring area of an IGBT exhibiting a low blocking voltage (50x). (b) Emission image of an IGBT with the blocking voltage limited by arching between the outer guard ring and the channel stop ring (20x).

Figure 3.6a shows an emission map of an *IGBT* with a low and unstable blocking voltage of *860 V*, which drifted towards higher voltages with time. According to *Section 3.2* the failure analysis started with high temperature storage at *200°C* (unpackaged device) for *6 hours*, without showing any recovery. Surface cleaning with *trichloroethylene* produced

a small increase of the blocking voltage up to *1000 V*. Emission analysis revealed a junction breakdown located at the surface between the penultimate and the last guard ring. This syndrome is typical for surface contamination. This hypothesis has been demonstrated by partial removal of the nitride passivation layer, which resulted into a prompt recover of the blocking voltage up to *1500 V*, and into the complete disappearance of the breakdown signature in the emission map.

Emission analysis is also very helpful in detecting *leaky junctions*. In fact, the major causes of leakages in junctions are either asperities causing premature breakdown or excess recombination centers within the depletion region. In damaged junctions (*e.g.* after spiking or filamentation) the leakage current is mainly due to *Frenkel-Poole conduction* [55]. In this case, the local recombination rate can reach such a level, that it can be easily be detected by emission microscopy.

When electrons tunnel across a thin oxide layer (*e.g.* gate oxide), they produce a photoemission spectrum centered on *2.5 eV* [56]. If there were no scattering with phonons within the oxide, the tunneling electrons would release at the anode *4.3 eV* for *n-well* capacitors and *4.8 eV* for *p-well* capacitors. Elctro-luminescence effects during *Fowler-Nordheim injection* of electrons into thin gate oxides are shown in *Figures 5.3*, *5.4*, and *5.6*. Furthermore, *Figure 5.7* shows an application of the emission microscopy for locating the site where oxide breakdown occurred.

Figure 3.6b, shows an unusual application of the emission microscopy for detecting partial discharge phenomena at high operating voltages. The *IGBT* under investigation exhibited during the pre-packaging electrical test an abrupt decrease of the blocking voltage down to *1400 V* (instead of *1600 V*). The emission microscopy analysis revealed a very intensive emitting spot located between the last guard ring and the external stop channel. In this case, the photoemission was not due to the mechanisms described above, but it was caused by arching at the device surface. The ultraviolet emission produced by the gas plasma within the discharging path was so intensive to be easily detected even with a detector working in the visible range. The unwanted arching path has been caused by an aluminum smear extending from the emitter metallization, which has been produced by a test needle.

Ohmic shorts, shorted metal interconnects, surface inversion, silicon conducting paths, and sub-threshold currents can be observed by emission microscopy only if they produce secondary effects (*e.g.* junction breakdown due to local inversion).

3.4.4 Scanning electron microscopy and EBIC

Since thirty years, *Scanning Electron Microscopy* (*SEM*) became a routine investigation tool for failure analysis of semiconductor devices. Due to its large depth of field and due to its ability to produce high-resolution images, *SEM* is used for microstructural surface topography characterization, metrology, and local elemental analysis. When the primary finely-focused electron beam impinges onto a sample, it generates a variety of secondary radiations, like secondary and backscattered electrons, *Auger* electrons, characteristic *x-rays*, cathodoluminescence, as well additional signals including absorbed current, induced hole-pair generation, and voltage contrast. All these signals address different physical and chemical properties of the irradiated sample and are used as quantitative information for analytical purposes [36,37]. A critical aspect of *SEM* is to adequately prepare the sample for avoiding charging effects and for enabling the visualization of the structure under investigation. These techniques, which include depackaging, selective etching, and delineation are considered in *Sections 3.3*, *3.5*, and *3.7*. Furthermore, when operating devices (especially power and high-frequency devices) within the *SEM* vacuum chamber, it has to be considered that cooling just occurs by conduction and not also by convection as it is usual in air. This situation can result into an excessive junction temperature of the device.

In the following, we focus our attention onto a dedicated detector, we developed [38] for performing Electron Beam Induced Current (*EBIC*) measurements on semiconductors samples, which exhibit high levels of leakage currents as soon the junction under investigation is set in reverse polarization. This is the case of microsections, and in particular of microsections of power devices. Basically, *EBIC* uses electrons injected by the primary beam to produce a map of the local recombination efficiency within the semiconductor. The physical background of signal generation and carrier transport in the space charge regions produced either by junctions or by Schottky contacts is well known. Briefly, the energy of the primary beam impinging onto a probe is mainly dissipated by generation of photons, phonons, secondary radiations, and by the creation of electron-hole pairs. In absence of a local electric field, electron-hole pairs rapidly recombine such that no *EBIC* current is generated and consequently no *EBIC* current is collected if an external amplifier is connected to the sample. On the contrary, if the carrier pairs are generated inside a space charge region, the local field drives electrons and holes in opposite directions, leading to the formation of a current to

be amplified by the external *EBIC* detector. The number of collected *EBIC* electrons is typically three orders of magnitude larger than the primary current, and under certain conditions, a locally reduced *EBIC* signal indicates the presence of recombination centers within the depletion. They are usually due to crystal defects, different doping concentration, contaminants, *etc.*. These recombination centers cause a modulation of the pure *EBIC* signal in the *0.5* to *5%* range. In general, the application of a reverse bias is not mandatory for producing a reasonable *EBIC* map. However, reverse biasing the device under investigation can improve the collection of the free carriers by expanding the space charge region. Polarization greatly improves the signal-to-noise ratio, especially where the dominating contribution is given either by the parasitic surface recombination, or by high levels of the dark current. This is the case of microsectioned devices, and of power devices, respectively.

When a reverse bias is applied to the device, the signal fluctuations due to the *EBIC* current are superimposed onto the contribution due to the leakage current by resulting into a dramatic degradation of the signal-to-noise ratio. Since the noise due to the leakage current is usually a slow function of the time, it could be virtually suppressed through a high pass filter realized by a capacitively coupled amplifier.

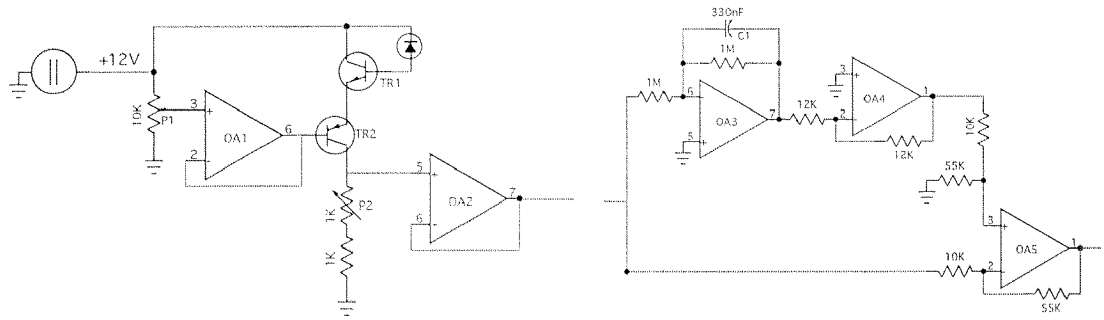


Figure 3.7 (a) Polarization and pre-amplification stage of the EBIC detector. (b) Adaptive DC-filter.

Unfortunately, this solution is not always viable, since in the case of either a large device or of very low scanning rates, the low frequency components of the *EBIC* signal would be completely canceled by the high pass filter making the detector completely useless. On the other side, *DC-coupled* amplifiers would be immediately driven into saturation by the large current offset due to leakage currents, which can exceed by up to an order of magnitude larger the pure *EBIC* signal.

In order to match previous requirements a *DC-coupled* amplifier with an adaptive compensation of the *DC-level,* with reverse bias capability of the sample up to *12 V* and a *cut-off* frequency of *300 kHz* has been designed. The detector consists of three stages: the polarization-preamplifier (*Figure 3.7a*), the *DC-filter* with self-compensation (*Figure 3.7b*), and an impedance transformer with brightness and contrast controls (not shown here). The working principles and the performances of this adaptive *EBIC* amplifier are described in detail in [38].

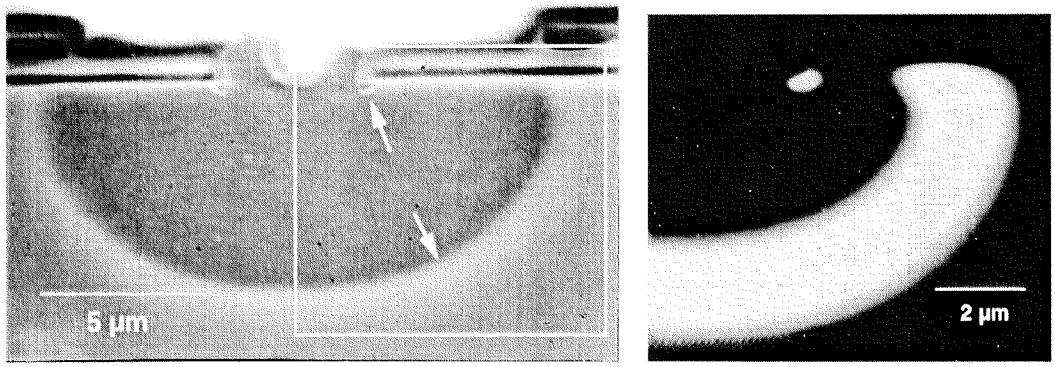


Figure 3.8 (a) Delineated microsection of an IGBT. (b) EBIC map of the region within the insert of the microsection

In *Figure 3.8*, we present crossection of an *IGBT* through the emitter contact. After chemical delineation (*see Section 3.7.2*) the junctions between n^+p (upper arrow) and pn^- (lower arrow) are clearly visible. A similar crossection intended for *EBIC* characterization, has been masked, sputtered with *30 nm* gold, and annealed at *250°C* for *20 minutes*, such that the lowest n^-/p^+ *junction* is short-circuited. The emitter contact has been polarized negatively (*–10 V*) in respect to the n^-*-doped* region, in order to set the p/n^-*-junction* in reverse bias. The *EBIC* map obtained with an acceleration voltage of *9 kV*, a probe current of *100 pA*, and a scanning rate of *0.05 frames per second* is represented in *Figure 3.8b*.

The *EBIC* map clearly shows the very intensive signal originated from the space charge region of the reverse biased pn^- *junction*. In spite of a leakage current of about *40 µA*, the self-compensation of the *EBIC* detector is sufficiently accurate for imaging the small space charge region associated with the n^+p *junction*. Here, the image formation mechanism is different than in previous case. In fact, the n^+p *diode* is forward biased and represents the basis-emitter junction of the parasitic transistor n^+pn^-, which is activated through the electron beam induced carriers. Imaging of

the n^+p *junction* would not be possible without polarizing the sample and without compensating the leakage current.

Unless the metallization is selectively removed and the device is sputtered with a conductive semi-transparent layer, the thick metallization of power devices (*3* up to *5* μm) imposes strong restrictions to the use of emission microscopy for the localization of gate oxide breakdown sites. In this case the most straightforward approach is to use *EBIC*. In fact, at a typical acceleration voltage of *35 kV*, the typical range of primary electrons in light materials, like aluminum and silicon, exceeds *8* μm. During the thermal breakdown of a thin gate oxide the energy stored in the capacitors close to the conducting path is released within several nanoseconds, by causing local melting of the polysilicon, of the oxide, and of the substrate materials. The conducting path resulting from this destructive event can be easily observed in the form of a leakage current through the dielectrics with an *I–V* characteristic, which is usually highly non-linear. It has been often registered that, if the transient is fast enough such that there just a limited doping interdiffusion between a heavily *n-doped* polysilicon gate and a p^+*-region*, the breakdown event results into a relatively well-defined junction with a *Zener*-like diode characteristic with a reverse breakdown voltage close to *4 V* [39,83].

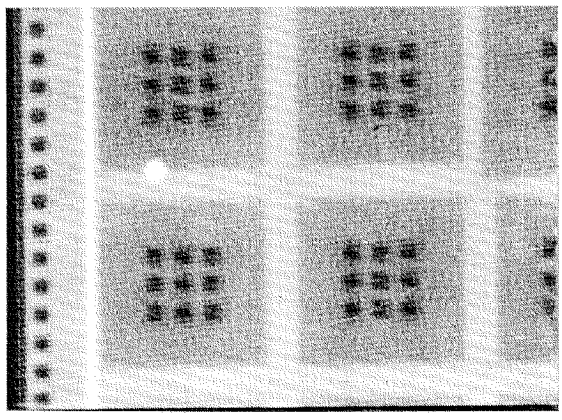


Figure 3.9 EBIC map of a MCT array. The bright spot represents the site of the gate oxide breakdown (100 x)

Figure 3.9 represents the case of a gate oxide breakdown between the n-doped gate and the p-doped region of a *MCT*. The junction (presenting a Zener characteristic) has been reverse-biased at *3 V* with the polarizing amplifier. The map has been acquired through a *2* μm thick aluminum metallization at an acceleration voltage of *35 kV* and with a probe current of *200 pA*.

3.4.5 X-ray microscopy

Failure analysis of *IGBT* devices takes advantage of real time *x-ray* microscopy systems, which provide a way for non-destructive imaging of microscopic package features [64]. Since *IGBT* modules are large and consist of materials with relatively high attenuation (copper, solder alloys), they require systems with a broad field of view and image intensifier tubes with high sensitivity. In a *x-ray* microscope, an electron beam of several milliamperes and with energy up to *160 keV* is focused to a spot onto a suitable target material (copper, tungsten). This spot is *3 μm* in diameter, and the size of the *x-ray* source determines the lateral resolution of the *x-ray* microscope. The projected image of the sample is collected by an image intensifier tube and then processed by a suitable image-processing unit. The contrast in *x-ray* images is due primarily to differential absorption of the radiation transmitted through the sample. Usual image intensifiers can resolve differences in the attenuation, which are in the *1% range*. Since the effect of delaminations or cracks is normally lower, they cannot be detected by *x-ray* microscopy. Denser materials such as copper, gold, lead, and tin have large mass absorption coefficient, and for this reason they can be easily observed, even if they are present in very thin layers. Unfortunately, relevant materials, such as silicon, aluminum, plastic mold compounds, and ceramics cannot be seen, if they are present in the sample together with denser materials.

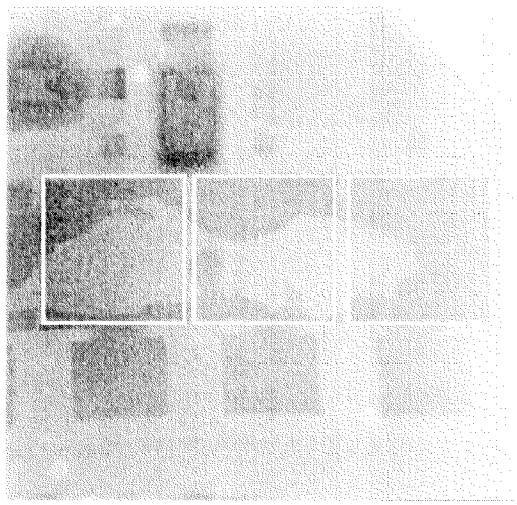


Figure 3.10 X-ray microscopy image of an IGBT module, which shows a large void immediately below three IGBT chips (0.8x). The void is located in the die attach layer.

Figure 3.10 shows a *x-ray* image of an *IGBT*, which failed due to inhomogeneous current sharing. The *x-ray* image shows a large void extending over three *IGBT* chips. Since *IGBT* packages are basically one-dimensional multi-layers, it is virtually impossible to determine by *x-ray* microscopy, only in which layer such a void is located. By successive microsectiong of the device it has been demonstrated, that the large void is located within the die attach layer. Thus, *x-ray* microscopy is a very powerful inspection tool if it is combined with other techniques, like scanning acoustic microscopy.

3.4.6 Scanning acoustic microscopy

Scanning acoustic microscopy (*SAM*) is an imaging technique for non-destructive internal sample inspection of multi-layered devices. In failure analysis of *IGBT* packages, *SAM* is a powerful technique for detecting voids, cracks and delamination problems, which may occur after cycle operation of the device. Scanning acoustic microscopy is based on the analysis (amplitude, phase, time of flight) of the ultrasonic waves reflected at the internal interfaces in the sample. The core element of the microscope is an ultrasonic transducer that works as generator, lens, and receiver of the ultrasonic signal. For power devices the working frequency of the transducer is in the *15 - 25 MHz* range. This represents a reasonable trade off between resolution and penetration depth. The most common operation mode of *SAM* in failure analysis is the *C-mode*, where a two-dimensional map of a given interface is produced.

Figures 3.11a and *3.11b* represent two *C-mode* images of the bottom and of the top interface of the solder layer between base plate and direct copper bonded ceramic (*DCB*) substrate of an *IGBT*, which has been submitted to extensive power cycling. It can be easily seen that the bottom interface is still integer, since the intensity of the reflected signal is almost homogeneous over the whole surface. On the contrary, the periphery of the *DCB* substrate appears as completely delaminated. Also in this case, the delamination has to be attributed to the shear stresses, which arise due to the thermomecanical mismatch between the base plate material (copper) and the ceramic layer (aluminum nitride).

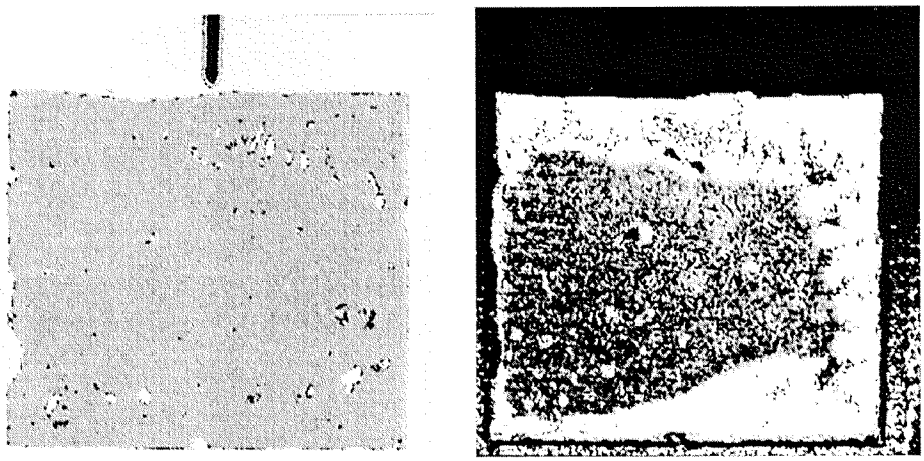


Figure 3.11 (a) SAM map of the interface between base plate and solder showing a homogeneous intermetallic layer (0.5x). (b) SAM image of the same solder layer at the interface with the ceramic substrate. It shows a large delaminated area at the periphery (0.5x, SAM images by courtesy of D. Newcombe).

When investigating *IGBT* devices, the scanning acoustic microscopy suffers from some limitations. The most concerns refer to the fact that some interfaces have three-dimensional features, which can scatter the transmitted acoustic waves in such a way to produce heavy artifacts. Furthermore, the important losses in ceramic materials such as *AlSiC* can restrict the use of the *SAM* in some cases.

3.4.7 Scanning probe microscopy

The scanning probe microscopy (*SPM*) is an imaging tool with three-dimensional topography profiling capability and which enables to measure local physical properties such as carrier concentration, magnetic fields, electric fields, temperature, and surface conductivity [60]. In the failure analysis of power devices, *SPM* is mainly used as a tool for the visualization either of microscopic topographic details or of regions with different doping concentration [52]. The related techniques are the *Atomic Force Microscopy* (*AFM*) and the *Scanning Capacitance Microscopy* (*SCM*). In both cases, the device under investigation requires a dedicated sample preparation procedure.

In the AFM mode a probe tip of few nanometers radius is kept in soft physical contact with the sample by a cantilever and it is scanned across

the sample by a piezo-electric transducer. As the transducer displaces the tip, the contact force with the surface of the sample causes the cantilever to bend in order to accommodate the changes due to the topography. The instantaneous height of the tip is monitored through a laser beam, which is reflected from the cantilever towards a photodiode. Finally, the topography map is reconstructed by an image processing system. A useful alternative to the contact mode is the AFM in tapping mode. This technique provides basically the same information than the AFM in contact mode but is carried out with an oscillating tip [61].

Since in these operating modes the *AFM* acquires the topography only, the semiconductor sample must be prepared in such a way that the properties, which have to be imaged, are translated into height differences. Thus, all the techniques, which are presented, in *Section 3.7,* can be applied also in this case.

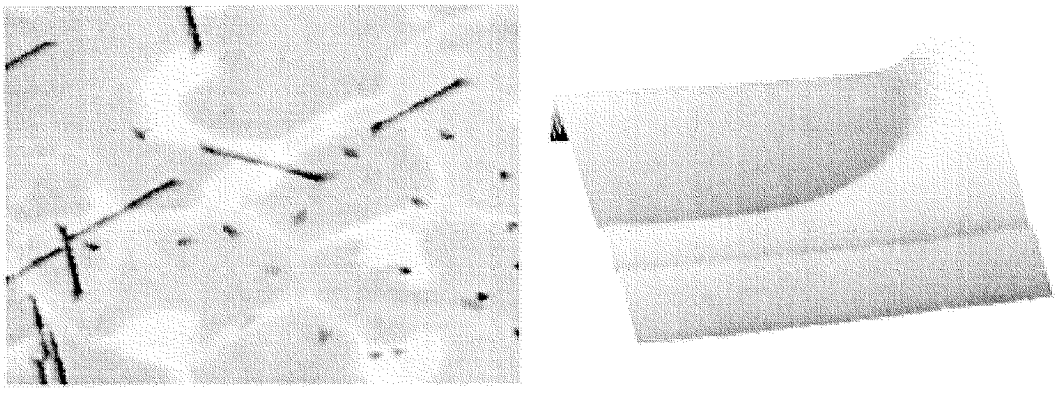


Figure 3.12 (a) AFM image of stacking faults in a power diode after defect delineation with the Wright etch (4000x). (b) Pseudo 3D representation of the n^+ edge termination of a power diode after delineation of the doping regions by the Malbot etch (1500x).

As an example, *Figure 3.12a* represents the topographic map of power diode after decoration of the stacking faults by the *Wright etch* (*see Section 3.7.1*). Different gray levels code the local height of the sample; *i.e.* bright features are in relief, while dark regions are in the depth. The selective doping etch technique has been used for producing *Figure 3.12b*. After being cross-sectioned and polished, the edge termination at the border of a power diode has been delineated by the *Malbot etch* (*see Section 3.7.2*), such that regions with different doping have been etched at a different etch rate. The n^+ *doped well* can be clearly seen, since it has been etched more in the depth than the surrounding *p-region*. It is worth

to note that this preparation also shows more subtle structures. This is the case the thin grove below the diffusion, which is due to local defect creation by proton implantation.

In the *Scanning Capacitance Microscopy* (*SCM*) a metallized tip is scanned across a cross-sectioned sample, where a thin oxide layer (typically *3 nm*) has been grown at low temperature (typically *250°C*) and under ultraviolet irradiation. The contact of the metallic probe with the oxidized semiconductor forms a local *MOS* capacitor. The local carrier concentration of the semiconductor can be extracted quantitatively from inverse modeling of the local capacitance curve, and a two-dimensional map of the sample can be reconstructed qualitatively by monitoring the capacitance variations as the tip scans across the sample surface. The typical lateral resolution, which can be achieved by *SCM* is *10 nm*, while the concentration resolution is in the *10% range* [51,58,59]. *Figure 3.13a* shows the flat topography of a cross-sectioned *IGBT* after oxidation as it is measured by *AFM* in the contact mode.

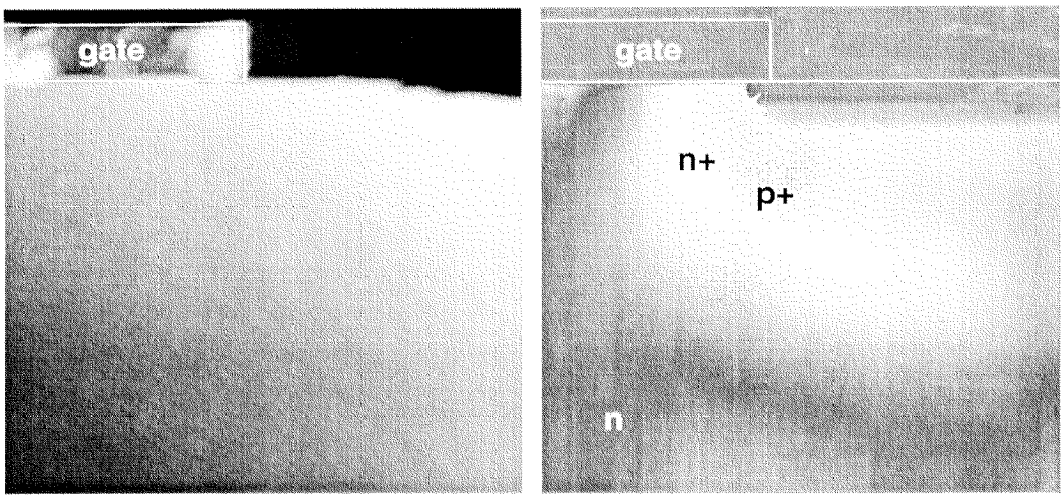


Figure 3.13 (a) AFM image of the channel region of a micro-sectioned IGBT showing a perfectly flat topography (1200x). (b) SCM image of the same sample showing the region of different doping (1200x).

Through combined mechanical-electrochemical polishing, the roughness of the surface has been reduced down to *5 nm peak-to-peak* (rms). *Figure 3.13b* represents an image of the same sample acquired by *SCM* in the *constant-dV mode* [51]. In the *channel area*, one can easily recognize the n^+ *source*, the p^+ *emitter* contact implant, the *p emitter*, and the *n substrate*. Investigations are going on for improving both the quantitative and the imaging capabilities of the technique.

3.5 Selective removal techniques

3.5.1 Removal of passivation and surface layers

The removal of the passivation (and in general of the surface coating layers) is often required in failure analysis of *IGBT* devices with the scope of eliminating residuals of carbonized materials or of surface contaminants. This is usually performed by underetching, *i.e.* by partial removal of the passivation. Total or partial depassivation of the device is also very helpful for improving the visibility of small features when using optical or scanning electron microscopy and for enhancing the adhesion of liquid crystals, when performing liquid crystal microthermography. Finally, the removal of the surface layers is always the initial step for selective layer-by-layer strip processes.

Besides the traditional inorganic passivation layers, most *IGBT* chips are also coated with thin polymeric layers to improve different properties, like the surface discharge resistance. Passivation layers are either of *silicon nitride* (Si_3N_4) or of *silicon oxynitride* if a lower intrinsic mechanical stress is required, while organic chip coatings are usually patterned *polyimide* films. In addition, some manufacturers make use of single or multiple organic coatings for the emitter bond wires, in order to retard failures due to bond wire lift off.

The polymeric chip and emitter bond wire coatings can be removed either by oxygen plasma or by red fuming acid at $50°C$. Both methods are selective and do not affect the inorganic passivation layers. The second technique has to be preferred, since when working with full multichip modules, the device usually does not fit within the reactor chamber of a plasma etcher. Furthermore, when using a system with microwave plasma generation, the device can reach very high temperatures due to the absorption of the microwave energy by the conductive parts of the package, in particular by the base plate. Even after partitioning of the chips, it should be paid attention to the possible plasma damage, which can occur due to the antenna effect produced by the long bond wires. Additionally, re-deposition of heavy metals on the top surface of the device can lead to a generalized short circuit.

Deprocessing multichip modules by wet chemistry requires the different agents to be confined within the area of interest. This can be achieved

either by masking the adjacent regions (by apiezon wax or by silicone), or by dropping low quantities of the agent directly on the layer to be removed after heating of the module with a hot plate. In particular, it should be avoided the contact between copper and nitric acid. Furthermore, one should be aware that, as a consequence of the galvanic potentials, which arise due to the presence of different metals, nitric acid could produce a fast dendrite growth, which may easily result into a short circuit.

Since polyimide is prone to hydrolysis and attack by alkali and concentrated acids, thin polyimide films deposited on the top of a passivated area can also be confortably removed by underetching. In fact, if thin polyimide layers are exposed at *65°C* to the passivation etch presented in the Table 3.2, they are weakened and soaked within some few minutes (typically *3 minutes*), such that the underlying passivation is slightly etched. This results into the loss of adhesion of the thin film, which peels-off and can be easily removed mechanically. This procedure is selective and avoids all the risks involved with plasma etching and nitric acid. The inorganic passivation is left unaffected, since the portion, which is etched during this treatment does not exceed *10%* of the original thickness.

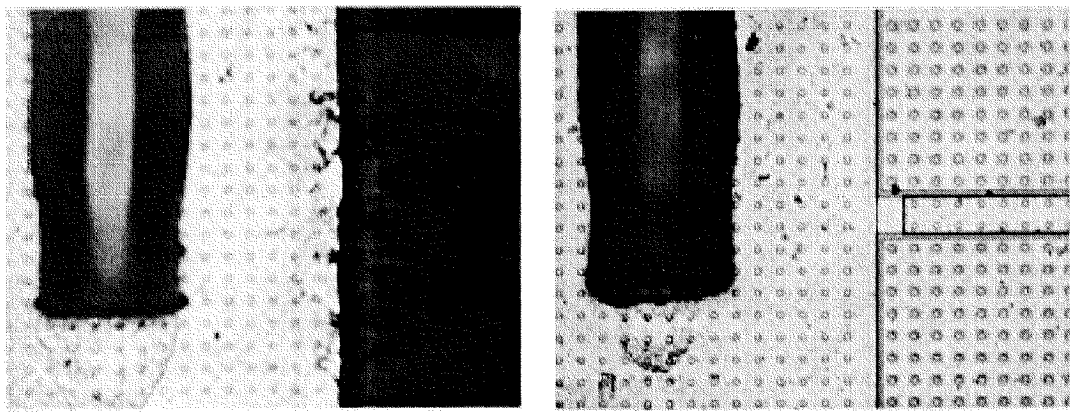


Figure 3.14 (a) Emitter and gate metallization of an IGBT before and (b) after selective removal of the polyimide coating layer (Optical microscope 50 x).

The removal of the passivation in order to expose the underlying connections without affecting the full functionality of the device has always been recognized as one among the most difficult tasks in failure analysis. Until a selective wet etch for silicon nitride and oxynitride has been developed [41,42] this process has been accomplished either by plasma or by reactive ion etching. In failure analysis, plasma etching is

usually performed by carbon tetrafluoride (CF_4), silicon hexafluoride (SiF_6), or mixtures of unsaturated gases (like C_2F_6 and C_2H_4). Silicon hexafluoride etches silicon nitride very selectively against oxide and silicon. However, due to its high reactivity, it requires rather expensive reactors and piping systems. Carbon tetrafluoride exhibits a good etching rate for silicon nitride. Unfortunately, it is not selective, and it etches silicon and silicon oxide even faster. In spite of this, carbon tetrafluoride based plasma reactors are still very popular, and they are usually operated with a *90% CF_4* and *10% oxygen* mixture, in order to reduce the drawbacks associated with the formation of polymers. The main problems related with plasma etching are selectivity, temperature control, etch orientation, reproducibility of the etch rate, and end point detection. During the depassivation process they can result into underetching of thin metal lines, ionic contamination, plasma damage (or even dielectric breakdown), formation of polymers, re-deposition of metals, and enhanced etching rate at the passivation defects. Since all these concerns may have lethal consequences when trying to depassivate *IGBT* devices, the use of a selective wet etch has to be preferred.

Silicon nitride is etched both in hydrofluoric (*HF*) and hot phosphoric acid (H_3PO_4). Unfortunately, phosphoric acid etches very rapidly aluminum and no buffer is known, which can hinder this problem. On the contrary, buffers are known, which make hydrofluoric acid in aqueous solution very selective against aluminum. The buffer, which is normally used for hydrofluoric solutions, is *ammonium fluoride* (NH_4F) and the related chemical reaction have been investigated in [43,44]. In general buffered *HF* solutions etch both silicon nitride and silicon oxide. Unfortunately, the fact that the etch rate for silicon nitride is much more lower than for silicon oxide (*200 nm* per minute at room temperature) makes the usual buffered *HF* etches (*e.g.* the *PAF* etch [41]) unusable for depassivation purposes. This problem also arises in the case of the solution reported by *Shankoff* [45], where *ethylene glycol* is added as diluent. The same occurs if *ethylene glycol* is replaced by *glycerol* [41]. Starting from the *Shankoff* etch, we [42] arranged the concentration of *ammonium fluoride* and of *hydrofluoric acid* in order to realize a *pH* level of the solution, at which aluminum is still passivated. Furthermore, acetic acid was used as diluent instead of distilled water. The result of this preliminary attempt [42] has been a wet etch selective against aluminum in a very narrow temperature range (*45-50°C*), and with an etch rate (*80 nm per minute*) almost comparable for both silicon nitride and silicon oxide.

Further investigations have shown that solutions of ammonium fluoride and ethylene glycol were also effective in removing selectively silicon nitride passivation layers even when the hydrofluoric acid was completely suppressed. However, in absence of an acid activator, the reaction occurs at a reasonable speed in the *120°C* range, only.

Table 3.2 Wet etch for selective removal of silicon nitride passivation layers [41]

Component	Quantity
Ethylene glycol $HOCH_2CH_2OH$, 99+%	60 ml
Acetic acid CH_3COOH, 100% glacial	20 ml
Ammonium fluoride NH_4F, 98+%	12 g
Nitric acid HNO_3, 65%	14 ml

Due to safety reasons and in order to make such a solution appropriate for failure analysis purposes, it was decided to design a wet etch, which operates within a lower temperature range. Since the most suitable activator has been found to be nitric acid, we have proposed the recipe in *Table 3.2* [41].

This etch has an operating temperature in the *45-80°C range*, where the *pH* remains almost constant (*pH=4*). It shows an excellent selectivity over aluminum, polysilicon, and refractory metals. The etch rate increases smoothly with the temperature, and at the optimum operating point of *70°C* it is *70 nm per minute*, both for silicon nitride and silicon oxide. The chemical reactions involving silicon nitride, nitric acid, and ammonium fluoride are still poorly understood. This is also the case of the role played by *ethylene glycol* and similar organic diluents (*e.g. propylen glycol*) in preventing the aluminum from corrosion.

Figure 3.15b represents the emitter and gate metallization of an *IGBT* immediately after the removal of the polyimide coating layer. The arrow in *Figure 3.15a* indicates the rim of the silicon nitride passivation layer. After selective removal of the passivation layer with the solution in *Table 3.2* (*14 minutes* at *70°C*), the rim has disappeared, and the aluminum

metallization is fully exposed. Neither the aluminum nor the insulation layers have been affected by the depassivation process. After proper neutralization of the chemical residuals [41] and *10 minutes* storage in vacuum, no increase in the surface leakage current between gate and emitter is observed. Further storage under vacuum hinders water adsorption and the consequent increase of surface currents.

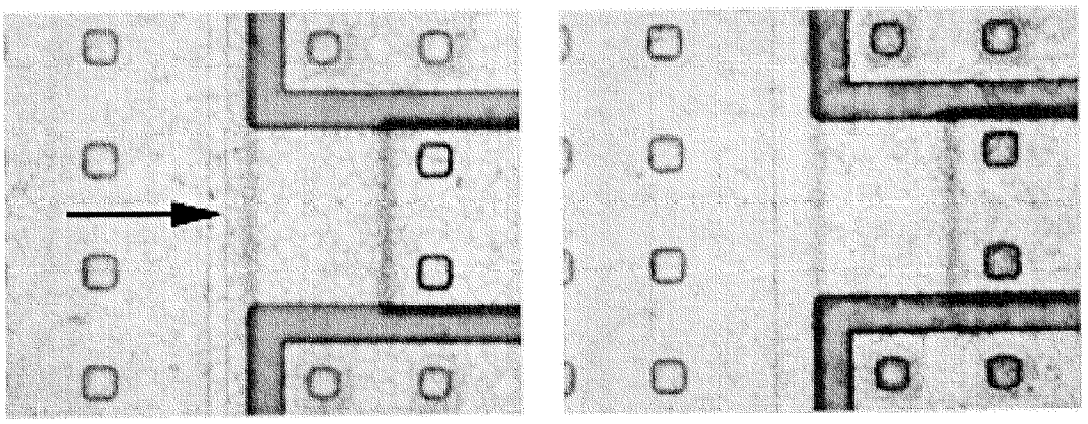


Figure 3.15 (a) Gate and emitter metallization of an IGBT with silicon nitride passivation layer (arrow) and (b) after selective removal (Nomarski interference, 200x).

We finally note, that by adding to the standard recipe of *Table 3.2*, *5 grams* of *potassium hydroxide* (*KOH*), the etch rate for silicon nitride gets at least one order of magnitude more than that for silicon oxide. However, the contamination, which results from the potassium ions hinders the use of such a solution for the most microelectronic applications.

3.5.2 Removal of the oxide layer

Selective removal of oxide layers is not as difficult as selective etching of nitride layers. Among the most efficient etches, one can mention the pure ammonium fluoride etch (*PAF*), which does not make use of hydrofluoric acid, as the common buffered hydrofluoric acid solutions do. The composition of the *PAF* etch is summarized in *Table 3.3*.

At room temperature, the etch rate of the *PAF* solution for phosphor-doped silicon oxide is in the range of *200 nm* per minute, depending on the concentration of the phosphor. The etching process has to be

performed under continuos stirring motion of the solution, in order to avoid the precipitation of ammonium-acetate salts, which could deposit onto the chip surface.

Table 3.3 Wet etch for selective removal of silicon oxide [46]

Component	Quantity
Ammonium fluoride NH_4F, 98+%	20 g
Acetic acid CH_3COOH, 100% glacial	50 ml
DI water	50 ml

In the case of thermal oxides, where the etch rate of the *PAF* solution is much lower than for deposited oxides, it is recommended to use the same etch than for the *HF strip*.

3.5.3 Removal of semi-insulating passivation layers

The guard ring area and the scribe lane of advanced *IGBT* devices can be passivated with semi-insulating layers made either of diamond-like carbon (*DLC*) or of semi-insulating polysilicon (*SIPOS*). These layers, whose thickness is in the *50 nm* range, are contacted with the guard rings and with the edge termination by contact windows through a thick insulation layer (usually silicon nitride, *see Figure 3.18a* and *3.18b*).

DLC layers cannot be removed selectively by wet chemistry, even if very strong oxidizing agents, like hot red fuming nitric acid, are used. Nevertheless, we succeeded in removing selectively *DLC* by plasma etching in pure oxygen atmosphere. At moderate plasma power, the removal process requires about *30 minutes*.

Also *SIPOS* layers are relatively resistant to wet etches, and in particular they are not attacked by the polysilicon solution presented in *Section 3.5.6*. Plasma etching (*e.g.* in CF_4 atmosphere) can remove *SIPOS*, however it is not selective over silicon and silicon oxide. In the case of a

complete delayering of the device, *SIPOS* is removed by a *25%* aqueous solution of hydrofluoric acid at room temperature and in an ultrasound bath. As it can be seen from *Figure 3.18*, the etch rate of *SIPOS* is lower than that of silicon oxide and of silicon nitride, such that the removal mainly occurs by underetching.

3.5.4 Removal of the aluminum metallization

The scope of this deprocessing step is to remove the thick emitter (and eventually gate) aluminum metallization without affecting the underlying layers. Following considerations apply for *IGBT* chips without molybdenum strain buffer plates. The metallization requires to be removed for different purposes than just for inspecting the active area. One among these is to enable emission microscopy for the localization of light emitting sites (*e.g.* thin oxide breakdown sites). In this case, the sample preparation consists into a selective etch of the emitter metallization followed by the deposition of a semi-transparent gold layer with a typical thickness of *20 nm*. In order to keep the functionality of the *IGBT*, all those sections of the device which do not need to be removed or short-circuited (*e.g.* gate contact, gate bridges, *etc.*) have to be masked either with a photoresist layer, or more simply with apiezon wax. When realizing such a semi-transparent electrode it is useful to keep intact a portion of the emitter metallization, in order to provide a contacting area for the needle and for improving the electrical continuity with the gold electrode. This requires a two-mask process. The first (positive) mask defines the portion of the metallization to be removed, while the second mask defines by lift-off the area for gold deposition. When working with non-packaged *IGBT* chips, the sputtering of the gold layer and the removal of the second mask have to be followed by an annealing at *200°C* during *30 minutes*. This step promotes the sinterization of the gold with the residual emitter aluminum metallization and with the residual diffusion barrier at the emitter contacts (*e.g.* silicide). The total current through the gold electrode should never exceed *5 mA*. This technique has been used for the preparation in *Figures 5.3* and *5.4* (*Chapter 5*) for investigating the *Fowler-Nordheim injection* through the gate oxide of real devices. If a larger emitter current is needed, the emitter electrode can be masked with a dense grid pattern. Of course, this requires a more sophisticated photolithography equipment.

The most common wet etch for aluminum is the buffered solution of phosphoric and nitric acid presented in *Table 3.4*.
This wet etch has an optimal operating temperature of *50°C* and provides a two-steps reaction. Aluminum is firstly oxidized by nitric acid and the reaction product ($Al(OH)_3$) is immediately dissolved by the phosphoric acid. The etch rate is of the order of *300 nm per minute*, however it depends on the size of the aluminum grains and it is slowed down by the presence of copper in the metallization.

Table 3.4 Wet etch for selective removal of the aluminum metallization

Component	Quantity
Orthophosphoric acid H_3PO_4, 85%	156 ml
Nitric acid HNO_3, 65%	10 ml
Acetic acid CH_3COOH, 100% glacial	34 ml

The end-point detection of the process can be easily performed by optical microscopy. *Figure 3.16* represents an *IGBT* device before and after selective removal of the aluminum metallization.

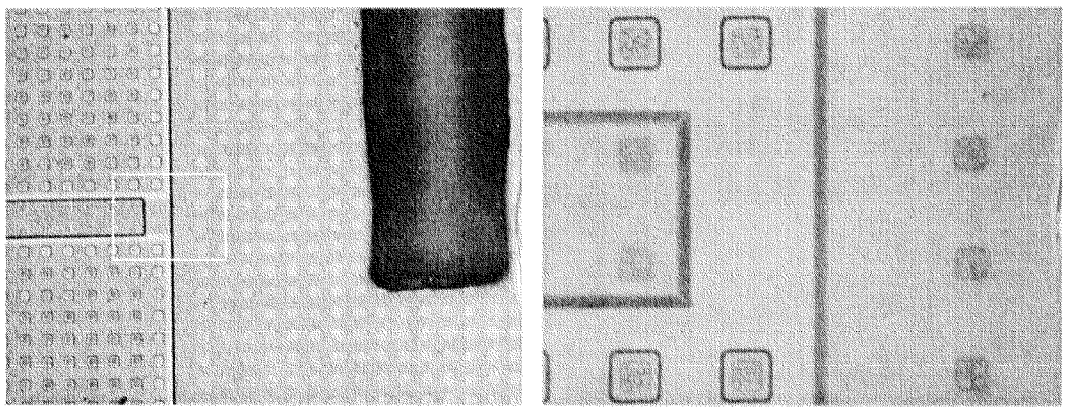


Figure 3.16 (a) IGBT in Figure 3.15b after removal of the aluminum metallization (optical microscope 50 x). Insert of image on the right after removal of the precipitates (Nomarski interference, 250 x).

As it can be observed in *Figure 3.16a*, the oxide insulation is still not transparent even after complete removal of the aluminum layer. This is due to the surface roughness produced by small silicon-rich precipitates, which form during the sintering phase of the metallization. These residuals can be easily removed by slightly etching at room temperature the oxide isolation layer by a *10%* aqueous solution of hydrofluoric acid, as it is shown in the magnification of the insert of *Figure 3.16a*. It must be noted that, even after this cleaning step larger precipitates can still be seen, especially within the aluminum-silicon contact windows.

3.5.5 Removal of the silicides at the contacts

Silicides at the contacts are not removed by the aluminum etch of *Table 3.4*. Since almost all silicides ($TiSi_2$, $PtSi$, WSi_2) are soluble in hydrofluoric acid they can be etched away by a *20%* aqueous solution of *HF*. However, hydrofluoric acid is not selective over silicon oxide and moreover silicon oxide has a faster etch rate than the silicides. Thus, this cleaning process can result into a partial or complete removal of the insulation oxide. Titanium and tungsten silicides are usually removed by *3 parts* ammonium hydroxide (NH_4OH, *58%*) diluted into *100 parts* hydrogen peroxide (H_2O_2, *30%*) at *90°C*. In the case of a failure of previous etch, a more aggressive procedure can be used on samples completely delayered down to the substrate [46]. This process consists in thermal shocks realized by dipping alternatively the sample into a concentrated sulfuric acid (H_2SO_4) bath at *150°C* and into a water bath at room temperature.

3.5.6 Removal of the polysilicon gate

After having etched selectively the passivation, the metallization, and the insulation layers, the next step is the removal of the polysilicon gate without affecting the underlying gate oxide. In *IGBT* devices the polysilicon area is quite large, and sometimes, depending on the scope of the analysis, the gate does not need to be removed completely.

The wet etching process consists in two distinct phases. In the first step the thick thermal polyoxide, which is grown onto the polysilicon is

removed. Then, after having exposed the polysilicon, it is selectively etched. The first phase is realized with a *40%* aqueous solution of hydrofluoric acid (*HF, 40%*) at room temperature. The most critical phase of this process is to determine when the polyoxide has been completely removed. In fact, the etching time strongly depend on the oxidation conditions and the end point detection cannot be performed just by visual inspection. In this case, the etching time has to be measured with a sacrificial device. This measurement is performed by stopping the oxide etch process after a given time and by checking whether the polysilicon is attacked uniformly by the polysilicon etch. If it is not the case, the oxide etch process is resumed until the oxide has been completely removed.

At room temperature, the polysilicon etch solution in *Table 3.5* removes the gate layer of *IGBT* devices in about one minute.

Table 3.5 Wet etch for the removal of the polysilicon after removal of the polyoxide

Component	Quantity
Nitric acid HNO_3, 65%	50 ml
DI water	30 ml
Hydrofluoric acid HF, 40%	1.5 ml

It is important that the polysilicon solution is prepared immediately before the use according to the procedure stated in [40], since it degrades after some few hours. Furthermore, any contamination of the solution by metals, which could be present in the package, should be strictly avoided by masking them with silicone or Apiezon wax.

Three emitter contact windows are represented in *Figure 3.17* before and after selective removal of the polyoxide and of the polysilicon gate with the solution in *Table 3.5*. In *Figure 3.17a,* the silicide strap can still be observed on the bottom and along the wall of the emitter window. During selective etching of the polyoxide the diluted hydrofluoric acid dissolves the silicide, by exposing the n^+ and the *p-doped* regions. *Figure 3.17b*, shows that the silicon within the emitter contact has been slightly etched. The rim, which can be seen around the emitter contact is due to the

complete removal of the polyoxide spacer and to the slightly underetch of the gate oxide, which clearly shows the underlying *n^+-region*.

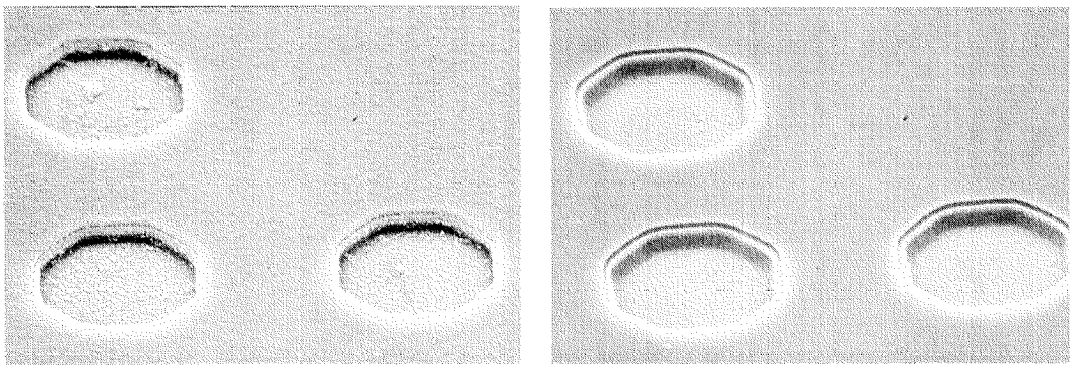


Figure 3.17 (a) Emitter contact windows before and after (b) selective removal of the polyoxide and of the polysilicon gate (SEM images, 1000x).

The central depressed region represents the contact *p^+-implant* of the emitter, which, according to the cell design, lies above the *n^+-region*. The gate oxide has been properly exposed and no residuals of polysilicon can be seen.

3.6 HF strip

The one step removal of all layers (passivation, metallization, dielectrics, polysilicon, and thermal oxide) of a device with the scope to expose the surface of the single crystal is called *HF strip*. This simple deprocessing technique is performed in an ultrasound bath at room temperature by a *25%* aqueous solution of hydrofluoric acid (*HF, 40%*). The *HF strip* is always recommended when the analysis is intended to evidence defects or damages within the silicon crystal. During this aggressive process, materials as aluminum, silicides, and silicon oxide are directly etched, while additional layers as polysilicon and silicon nitride are removed by the combined action of chemical underetching and of the ultrasounds. If ultrasounds alone are not sufficient in removing the underetched layers the result of the *HF strip* can be improved by mechanical rubbing the surface of the device with a cotton swab soaked with liquid soap. After this cleaning step the grease traces due to the soap have to be removed by rinsing in warm water and acetone. If the delayering process is not completed, etching can be resumed again. The *HF strip* always delivers clean samples, where the bare silicon has an almost uniform gray color.

Usually, crystal defects are not decorated, while major damages, like those due to thermal run away or gate breakdown events, are clearly evidenced.

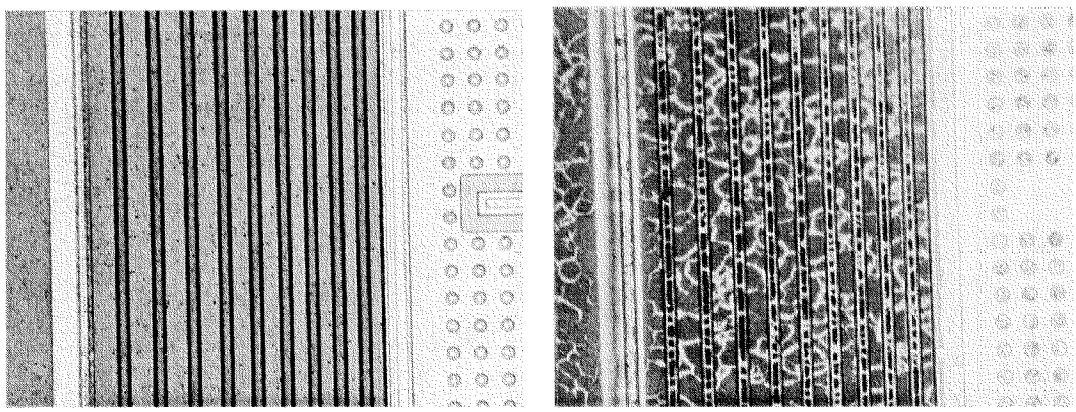


Figure 3.18 (a) Guard ring area of an IGBT before and during (b) the HF strip. The area with the emitter contacts has been completely stripped, while the SIPOS is being removed by underetching (Optical image, 200x).

HF strip also removes layers, like the *SIPOS*, which are very difficult to be removed by alternative techniques. This can be clearly in *Figure 3.18b*, where the *SIPOS* layer deposited onto a thick thermal oxide is being removed by underetching. The final result is a clean silicon surface, which just shows the topography due to different processing steps.

3.7 Delineation techniques

3.7.1 Delineation of crystal defects

Crystallographic defects (dislocations, stacking faults, oxygen precipitates, *etc.*) or minor defects in the silicon substrate can be evidenced by preferential etching, after removal of all the layers (*e.g.* by the *HF strip*). Due to the enhanced etching rate of regions with high doping concentration, defect delineation in *IGBT* devices is not always easy. The most popular recipes used in failure analysis are the *Wright etch* [47], the *Secco etch* [48], the *Sirtl etch* [49], and the *Schimmel etch*

[50]. Excellent results have been obtained in the failure analysis of *IGBT* devices by using the *Wright etch*, whose recipe is described in *Table 3.6*.

Table 3.6 Components of the Wright etch

Component		Quantity
A	CrO_3	15 g
	DI water	30 ml
B	$Cu(NO_3)_2\ 3H_2O$	2 g
	DI water	60 ml
C	Acetic acid (glacial, 100%)	60 ml
	Nitric acid (69%)	30 ml
	Hydrofluoric acid (49%)	60 ml

The solutions *A*, *B*, and *C* in *Table 3.6*, have to be prepared separately. Shortly before the use, solutions *A* and *B* are mixed and let emulsify for some minutes. Finally, the solution *C* is added to mixture *A* and *B*. The obtained solution is a preferential etch, which works optimally for the crystallographic orientations (*100*) and (*111*). In case of defects (or of high doping concentrations of both type) the usual etch rate at room temperature is about *1 μm per minute*. The detailed procedure to be followed for defect delineation in *IGBT* devices is presented in [40].

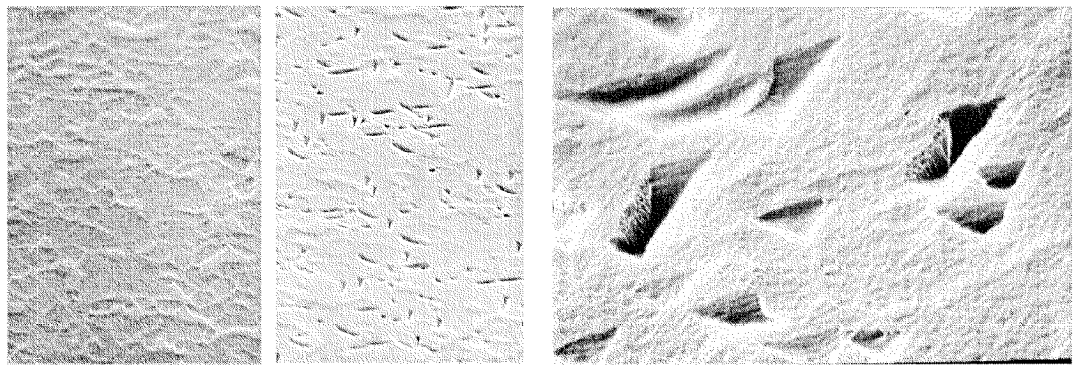


Figure 3.19 Crystallographic defect delineation by the Wright etch of a power diode (a) without and (b) with aluminum shallow implant (SEM, 1000x). (c) Magnification of a region presenting crystal defects, which shows the preferential etching behavior of the Wright etch (SEM, 7000x)

The two first images in *Figure 3.19* show the result of a defect delineation by the *Wright etch* of a diode without and with a transparent anode structure. The preparation clearly demonstrates that the damages produced by the aluminum shallow implant were not completely annealed during the successive thermal treatments. This results into residual dislocations and stacking faults, which are decorated during the sinterization of the metallization. The spiking reduces the reverse breakdown voltage during dynamic switching of the device. *Figure 3.19b* represents at high magnification a region, where the crystal defects have been delineated by the *Wright etch*. The enhanced etch rate of the solution along the crystallographic defects clearly evidences stacking faults and dislocations, which are originated in the bulk.

3.7.2 Delineation of doped areas

One among the most important step in the technological characterization of an *IGBT* device is the delineation of the different doping regions. This phase is of great relevance either when extracting quantitative information about the active area (*e.g.* for simulation purposes), or just for understanding the architecture of a device under investigation. Traditionally, three different procedures are used: *delineation*, *decoration* and *staining*. Delineation is produced by a differential in the etch rates of two or more materials constituting the sample. The difference in the etch rates directly translates into a height difference on the specimen, which can be observed with appropriate microscopy techniques. Decoration is produced by preferential galvanic electroplating one side of the junction of interest. Usually, decoration is realized by using solutions of compounds of transition metals (*e.g.* copper sulfate and copper nitrate), and it is performed under strong illumination for increasing the photovoltage. Stain processes result into the selective growth of a thin oxide film onto silicon doped areas. Stained sample are usually investigated by optical microscopy (bright field, *Nomarsky* interference). The stain films exhibit commonly interference colors (typically brown or violet) and are not soluble in hydrofluoric acid solutions. Stains are usually buffered (acetic acid) solutions of nitric and hydrofluoric acid. Recently, scanning probe based techniques have been used for investigating doping profiles in *IGBT* devices ad in power diodes [51]. These techniques are the atomic force microscopy in the topography mode, the scanning capacitance microscopy, and the scanning spreading resistance microscopy [52]. Applications of these methods are shown in detail in *Section 3.4.7*.

The experience has shown that, excepted scanning probe based techniques, wet chemical delineation is the most adequate process for preparing *IGBT* devices. *IGBT* samples can be delineated either from the top surface after partial or total delayering, or after microsectioning. Because of the etch rate depending on the doping concentration, the *Wright etch* presented in *Section 3.7.1* yields excellent results also for the delineation of *IGBT* cells, provided that the sample is free of crystallographic defects.

Figure 3.20 shows the delineation of the emitter contact area of an *IGBT* by the *Wright etch* after removal of the metallization and of the oxide layers. The differential etch behavior of the solution evidences three regions, which were not visible before the delineation. It is interesting to note that the same n^+ *region* is etched differently, depending on the fact that it is in contact with a p^+ or with a *p well*.

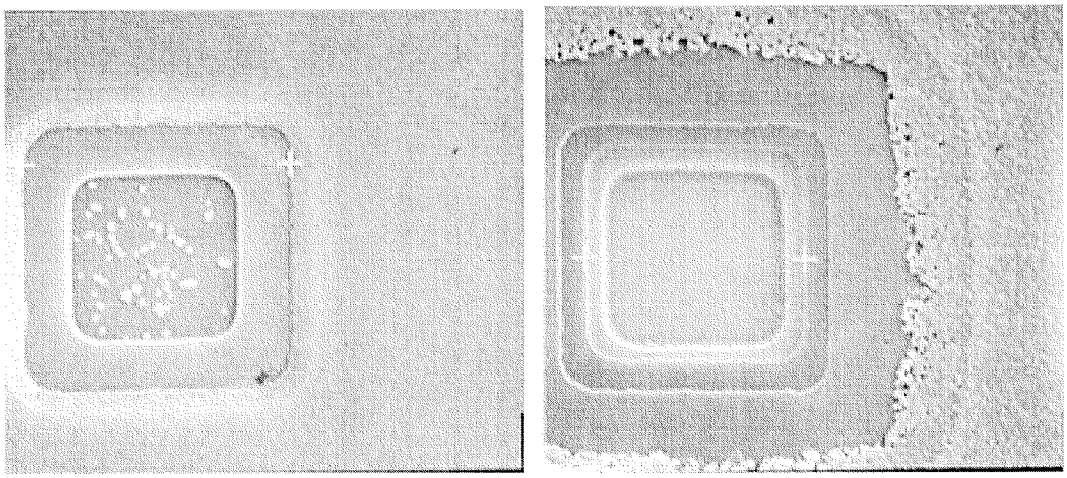


Figure 3.20 (a) Emitter contact window of a device of the manufacturer A after removal of the metallization. (b) Delineation of the emitter contact with the Wright etch after partial removal of the polysilicon gate (SEM, 1500x).

In the case of micro-sectioned devices, the *Wright etch* interacts often with the crystallographic defects, which arise during lapping and polishing, by resulting into deep scratches, which make the preparation unusable. For this reason, the delineation of microsections (or of cleaved samples) should be preferably done by the *Malbot etch* [53], whose recipe is presented in *Table 3.7*. The *Malbot solution* works at room temperature after removing the native oxide layer by a diluted aqueous solution of hydrofluoric acid and has a typical etch time of *10 seconds* (without illumination).

Being a variant of the classical *3-1 etch*, the *Malbot etch* does not exhibit any preferential etching behavior and thus it can be used as an all-purpose delineation etch. For doping concentrations lower than 10^{17} cm^{-3}, the delineation by the *Malbot etch* has to be performed under strong illumination.

Table 3.7 Malbot etch for the delineation of microsections

Component	Quantity
Nitric acid HNO_3, 65%	20 ml
Acetic acid CH_3COOH, 100% glacial	50 ml
Hydrofluoric acid HF, 40%	2 ml

Figure 3.21a identifies the different doping regions of an *IGBT* device of *manufacturer A*, which can be seen in the preparation of *Figure 3.20b*. The delineation of the microsection with the *Malbot etch*, clearly shows the polysilicon gate, which has been completely etched, the *p^+-contact implant*, the *n^+-source*, the *p-emitter* and the *n-substrate*. This preparation also enables to measure the *channel length*.

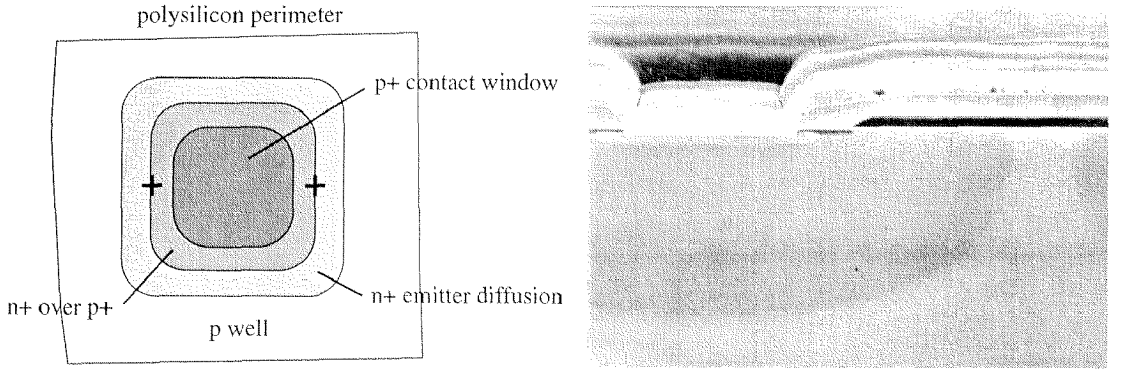


Figure 3.21 (a) Emitter contact window in Figure 3.20b. (b) Cross-section through an emitter window of the same device than in Figure 3.20b, after delineation with the Malbot etch (SEM, 3000x).

Figure 3.22 presents the microsection of an *IGBT* device of *manufacturer B* after delineation with the *Malbot etch*. The most evident difference with the design of *manufacturer A* is the n^+-*emitter*, which is smaller and does not overlap with the p^+-*contact* implant.

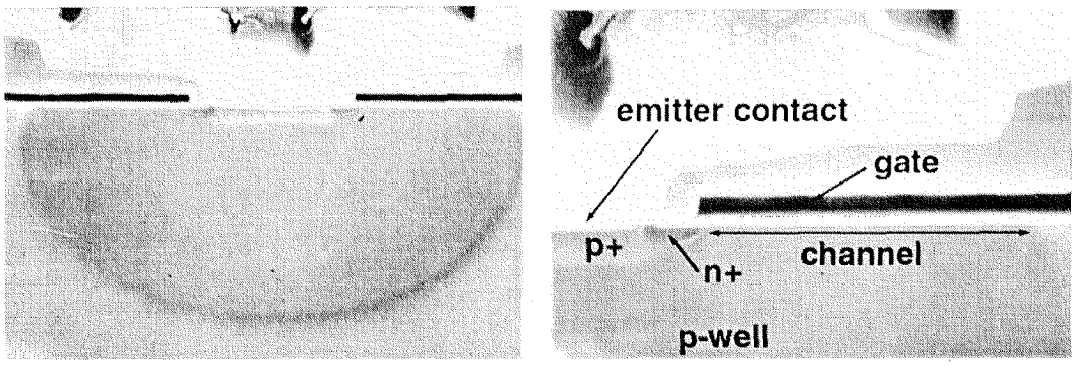


Figure 3.22 (a) Cross-section through an emitter contact of a device of manufacturer B after delineation with the Malbot etch (SEM, 2500x). (b) Detail of the channel area of Figure 3.22a with identification of the delineated regions (SEM, 7500x).

Before microsectioning or cleaving a sample to be delineated, it must be accurately prepared. In fact, we observed that the best results are obtained if all doping regions to be delineated (including the polysilicon gate) are galvanically connected. This situation can be realized by sputtering a thin gold layer on the surface of the *IGBT*, such that emitter and gate are short-circuited. Since the contact to the collector is normally provided by the backside metallization, it does not requires any additional preparation. In order to avoid that the *n-substrate* is left floating, one side of the chip is cleaved, sputtered, and annealed. Finally, during the etching phase, all the surfaces of the device are connected together by means of an aluminum foil, which should not enter in contact with the etching solution.

3.8 Etching of the silicon chip

The silicon substrate can be etched because of multiple reasons. The most simple application of this process is the complete removal of the single crystal as an alternative (destructive) technique to *x-rays* for the inspection of the integrity of the die solder. This procedure is preceded by the complete removal of the passivation, metallization, dielectrics,

polysilicon, and thermal oxide by means of a *HF strip* (*see Section 3.6*). Then the pre-processed device is immersed into a *20%* aqueous solution of *potassium hydroxide* (*20 g KOH* in *100 ml* water) at *80°C* and under stirring motion. The etch proceeds isotropically at an etching rate of about *70 μm per hour* and it is selective over the usual solder alloys. In few devices, it has been observed that the etching process stops at the back metallization, especially if it is very thick and did not completely interact with the solder. In this case the residuals of the metallization should be submitted to microanalysis for chemical identification and then removed by a specific etch (normally based on nitric acid). Once the solder has been exposed, it is straightforward to observe macroscopic voids or inclusions by stereoscopic microscopy.

An additional application is selective device backside etching to provide direct access by optical microscopy to normally hidden structures, such as gate oxides, metal-to-silicon contacts, polysilicon lines, and wire-to-metallization interaction area. Many silicon backside etch processes are used in failure analysis, like those which base on aqueous solutions of *potassium hydroxide* (*KOH*), *choline*, *ethylene-diamine-pyrochatecol*, and *freon plasma*. The main limitation of these wet and dry etches is the poor selectivity over dielectrics and interconnection materials. In particular, none of the mentioned etches is selective at the same time over silicon oxide and aluminum at the contacts. For this reason, a selective back etch for silicon devices based on an aqueous *4%* solution of *TetraMethyl-Ammonium Hydroxide* (*TMAH*) has been developed, which fulfills all the requirements concerning the selectivity [62]. The etch rate of this solution at *80°C* has been measured to be *40 μm per hour*. However, the selectivity of the etch over various dielectrics is improved, if *13.5% by weight* silicon powder is added to the *4% TMAH* aqueous solution. In addition, this doping level keeps the *pH* of the solution within the range where the aluminum is self-passivated. The selectivity of the doped etch has been observed to be in the range of *10000* over thermal oxide and silicon nitride, and in the range of *1000* over deposited oxides. Before being processed, single chips have to be prepared such that they are glued topside down onto a substrate (ceramic or glass). The substrate must be stiff enough and the glue should not introduce tensile stresses, in order to avoid that the thin self-sustaining membrane, which results after removal of the silicon crystal, is destroyed. The backside metallization of the device must be preliminary removed either by wet chemical etching or by mechanical polishing. In the next step the thin native oxide on the top of the freshly exposed backside of the silicon has to be removed by a *40%* aqueous solution of hydrofluoric acid. Then, after rinsing in water, the sample is dipped into a beaker with the *TMAH* solution at *80°C*. Since the

etching process can take several hours, the beaker should be covered, in order to avoid water evaporation. The device should not be removed from the etching bath, before the region of interest of the chip has been completely exposed.

Finally, selective silicon etching can also be used for detecting tiny pinholes, which can occur in thin oxide as a consequence either of gate oxide breakdown events, or of thinning effects. In this case, the sample must be deprocessed down to the gate oxide level (*see Section 3.6*) and immersed into a *20% by weight* aqueous solution of *KOH*. This technique bases on the fact that under these conditions, the etch rate for silicon of such a *KOH* solution is about *1000* faster than for a thermal oxide. Thus, since the thin thermal oxide acts as a positive mask, if pinholes or a thinned area are present, they will result into an underetch of the exposed silicon. After removal of the thermal oxide by *HF strip* (*see Section 3.6*), the pits in the silicon substrate can be observed either by optical or by scanning electron microscopy.

3.9 Microsectioning and Focused Ion beam

Microsections of power devices can be prepared either by cleaving, by mechanical polishing, or by *Focused Ion Beam* (*FIB*) sputtering.

Since *IGBTs* are mostly large periodical arrays of identical cells, single chip cleaving is an excellent technique when spatial resolved microsections are not required. The cleavage is always preceded by notching (or scribing) of the crystal by a tungsten pencil, by a diamond disc saw, or by more sophisticated tools, as they are used in precision cleaving equipment. Manual cleaving is normally performed by placing a thin insulated copper wire under the notch in the direction of to the desired cleavage axis, and by exerting a slight downward pressure with the fingers. Outstanding results have been also been achieved with thicker crystals just by gently pressing the back side of a chip against a rotating diamond disc saw until the sample breaks. A problem normally encountered in manual cleaving is the smearing of soft materials like thick aluminum layers. In this case, cooling the sample down to liquid nitrogen temperature is strongly recommended. All the microsections imaged in *Section 3.7.2* have been obtained by cleavage. Particular features can be localized with a precision of about *500 μm*.

Unfortunately, the simple cleavage procedure is ineffective either for assembled devices, or if a microsection through bond wires, ceramic substrates, and solder layers is needed. Cleaving is also not suitable for *Scanning Capacitance Microscopy* (*SCM, see Section 3.4.7*), since the excessive roughness of the surface does not allow to grow reasonable thin oxides. In those cases mechanical polishing is mandatory.

For packaged samples a diamond disc saw is used to approach the area of interest, and then the sample is encapsulated in epoxy. For grinding and polishing a wide range of procedures are available, which are described in more detail in [40]. In general, packaged *IGBT* samples contain materials with different hardness (ceramic, silicon, copper, aluminum, solder alloys). This leads usually to uneven material removal rates. A case of special interest for *IGBT* is cross-sectioning of aluminum bond wires for the investigation of the initiation and the propagation of cracks during thermal cycles. While cross-sectioning bond wires the process is stopped immediately before the final polishing step, in order to avoid that small features like cracks or voids disappear due to smearing of the aluminum. Small cracks are often only visible after decoration of the aluminum grain boundaries. The composition of an effective wet etch (*Keller etch*) is given in *Table 3.8*. Typically, the half-polished cross-sections are dipped into the *Keller etch* at room temperature for about *15 seconds* and then rinsed in water. Sometimes this process can result into dirty preparations. This is mainly due to the presence in the sample of multiple materials, which can produce galvanic couples with consequent deposition of metals in the solution, like copper.

Table 3.8 Etch for the delineation of the grain boundaries in aluminum

Component	Quantity
Nitric acid HNO_3, 65%	20 ml
Hydrofluoric acid HF, 40%	2 ml
Hydrochloric acid HCl	5 ml
DI water	190 ml

Figure 3.22 represents a micro-section through an emitter bond wire of a device, which has been submitted to thermal cycles. Since aluminum is a very ductile metal, it locally smears during the preparation process such that small features, like cracks are not visible.

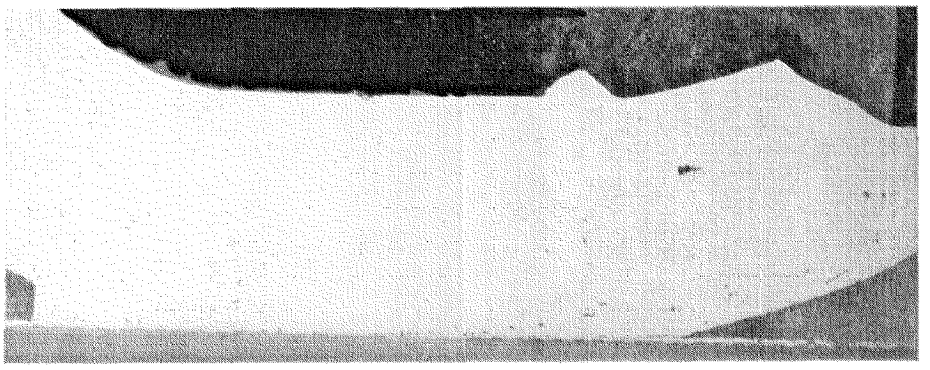


Figure 3.22 Optical image of a cross-section trough an emitter bond wire submitted to thermal cycles (100x). Without chemical delineation no crack is visible.

Figure 3.23 represents the same sample like in *Figure 3.22* after wet chemical etching. This image clearly shows a crack propagating from the bond tail, which has been properly delineated by the etch of *Table 3.8.* The additional pits distributed across the bond wire are due to the decoration of local defects introduced by the polishing with diamond paste.

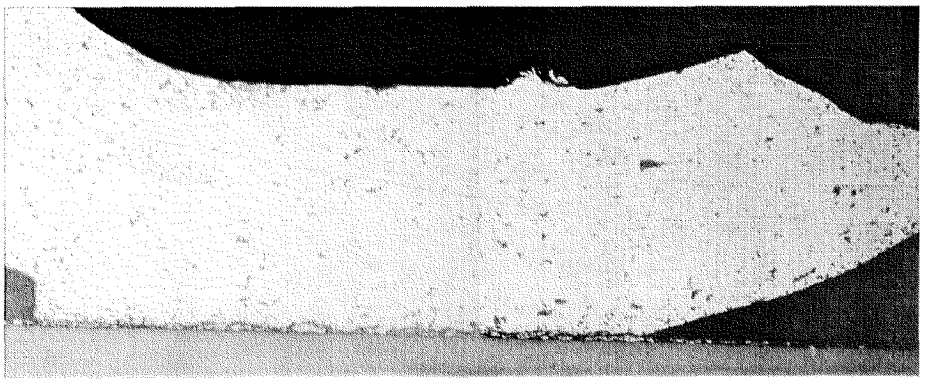


Figure 3.23 Optical image of the same cross-section than in Figure 3.22 after wet chemical delineation (100x). The crack within the aluminum bond wire is clearly visible.

Focused Ion Beam (*FIB*, [63]) becomes an increasingly important tool also for microsectioning *IGBT* devices. In the *FIB*, a focused beam of gallium ions scans across a designated area by milling a rectangular hole

into the sample. *FIB* enables the user to sputter the material, while simultaneously imaging the cross-section either by the *FIB* in the ion microscope mode, or by a scanning electron microscope (dual column system). *FIB* is recommended for precision microsectioning and for yield analysis. However, material re-deposition and gallium implantation (from the primary beam) makes this technique questionable for the preparation of samples either for *SCM* or for delineation. Furthermore, for section lengths exceeding *50 μm* (as it is the case for bond wires), *FIB* can become very time consuming, such that alternative techniques tend to be more effective.

3.10 Advanced characterization techniques

Several characterization techniques can also provide relevant inputs to the failure analysis of *IGBT* devices. This is the case of the *electrolytic metal tracer*, which is used for monitoring the metal contamination and the process induced defects. This technique bases on the measurement of a photo-current and provides a two-dimensional map of the recombination centers (with the associated lifetime) at the surface of processed or unprocessed wafers [65].

An additional technique basing on photo-excitation is the *Optical Beam Induced Current* (*OBIC*) microscopy. *OBIC* is essentially the optical counterpart of *EBIC* (*see Section 3.4.4*), which uses a scanning laser beam instead of an electron beam. As in the case of *EBIC*, *OBIC* maps provide important information about the distribution of the recombination centers [37]. Since as most of the surface of power semiconductors is metallized, *OBIC* microscopy from the chip back side is more effective. In this case, infrared lasers have to be used, in order to enable carrier generation even in the depth of the semiconductor. An interesting technique, which makes possible to acquire the free carrier density and the temperature profile within a power semiconductor under realistic operating conditions, is the *Laser Deflection Microscopy* (*LDM*, [66]). The *LDM* quantifies the local variations of the refractive index of the silicon as function of the temperature and of the free carriers density by measuring the deflection (or the absorption) of a finely focused infrared laser beam, which is transmitted through a specially prepared semiconductor sample.

Laser Speckle Interferometry (*LSI*) has been successfully applied for imaging thermomechanical deformations in power devices [69]. This technique takes advantage of the speckle structure produced by illuminating an object with a laser source. *LSI* enables to measure the displacement of each speckle dot for both in-plane and out-of-plane movement with a resolution close to *0.01 μm*.
When operating multiple *IGBT* chips in parallel within a power module, it is important that during the turn-on, turn-off, and the on state, the current is equally shared among the chips. A non-invasive technique, which is used for measuring the current switched by every single chip, is based on miniaturized air-cored field coils fitted around the emitter or collector conductors close to the chip under investigation [67]. In spite of the simplicity of the measurement principle, this technique requires accurate calibration and a lot of precautions for avoiding common-mode noise and stray pick-up from adjacent chips.

In infrared thermal imaging systems, the surface temperature of the device is inferred from the radiant energy at a particular wavelength (*see Section 4.8*). Unfortunately, under normal operating conditions the surface of *IGBT* devices is not visible since it is coated by a thick silicone gel layer. Even after removing the silicone gel, the infrared radiation originated from the chip surface is often perturbed by the shadow or the emission produced by the emitter bond wires. This problem can result into artifacts causing wrong temperature readings [7]. An alternative (invasive) technique for surface temperature measurements in *IGBT* devices is based on the temperature dependence of the time constant of the fluorescence decay in materials, like the *europium thenoylfluoroacetonate* (EuTTA), when they are stimulated by ultraviolet irradiation. A thin layer of such a material is applied on the top of a glass fiber, which is then placed in contact with the chip surface [68]. The active layer is stimulated every *250 milliseconds* by an ultraviolet source and the resulting transient of the fluorescence is acquired by a photodiode. This technique has an accuracy of *1°C* and provides a lateral resolution of *250 μm*. It also enables to perform time-resolved measurements with a typical repetition frequency of *4 Hz*.

Chapter 4

Experimental thermal characterization of IGBT devices

4.1 Introduction

Since power semiconductors are temperature-sensitive components a sophisticated thermal management is required in order to remove the heat produced during operation, and to keep the device within the safe operating area.

From the thermal point of view *IGBT* modules are multilayers made of different materials, with different thickness, and different thermal conductivity. Though the relative complexity of such a multilayered structure the produced heat is mainly dissipated by conduction along an almost one-dimensional thermal path. The thermal behavior of an *IGBT* is strongly influenced by elements beyond the thermal interface between the

device and the environment. The first factor affecting the thermal response is the thin layer of thermal grease, which is normally used for decreasing the thermal resistance due to the air gap between base plate and heat sink. This layer can include voids, can dry out during the operation degrading its thermal conduction properties, or can be also depleted due to the squeeze-out effect produced by the cyclic bowing of the base plate during operation. Therefore, the thermal grease introduces some uncertainty while estimating the thermal resistance of the device, and in addition it may introduce a time-dependent degradation.

An additional problem, which is often underestimated, is the role of the heat sink. Normally, in thermal simulations the heat sink is considered as an ideal isothermal boundary condition and the related thermal capacitance is usually neglected. In fact, a real heat sink is a complex and bulky mechanical system, whose dissipation properties change from location to location (*e.g.* locations close to the water inlet or not). All these non-idealities may be the cause of different local temperature levels within a module, leading to an uneven current distribution among the different chips and finally to premature device degradation. Furthermore, the thermal mass and the size of a heat sink often exceed those of the device. As a consequence, it largely contributes both to the overall thermal resistance, as well to the overall thermal impedance. As an example, a water-cooled copper heat sink, whose contact interface plate with the device is *5 millimeters* thick, represents up to *50%* of the overall thermal resistance, and may have a thermal capacitance, which is about a *factor of 5* larger than that of the whole module. Therefore, the slow thermal characteristics of a module are largely dominated by the thermal properties of the heat sink.

New generations of dedicated *IGBT* packages are under development, which are intended to match special requirements as they are imposed by particular applications like railway or automotive traction. Solutions, which are envisaged in the next future are integrated packages, which make use of very efficient solutions like direct chip cooling or cooling channels realized directly within the base plate. The main scope pursued with these new technologies is to quench as much as possible the temperature excursion due to switching operation, in order to reduce the related thermomechanical stresses. Although the development of these new package technologies is slightly delayed due to the fact that standard *IGBT* packages offer the advantage of high modularity (*i.e.* of the economy of scale), innovative module designs will appear, as soon new applications will represent a perspective for high-volumes (*e.g.* electrical or hybrid vehicles).

In summary, thermal management of *IGBT* modules is a main issue, which impacts both device performances and reliability. Therefore, accurate and efficient experimental methods are required to ensure that real devices meet all requirements imposed by present and future applications.

In this Chapter, after recalling some fundamentals, two characterization techniques are presented for the measurement of the chip temperature in *IGBT* modules. In the first part, the use of a new coating layer is demonstrated, which improves the temperature and the lateral resolution of measurements by infrared thermography. The method considered in the second part bases on the principle of the *time-resolved terminal voltage measurement.* A dedicated circuitry is presented, which enables to compensate the long-term drift of the peak junction temperature and that delivers quantitatively the conjugate of the transient thermal impedance.

4.2 Effect of the temperature on IGBT devices

The heat generated by a device when it is in operation raises the temperature of the component where it originates, the temperature of the surface on which it is mounted and the temperature of neighboring components. Although, this work just deals with failure mechanisms in *IGBT* modules, it is worth to mention that a large amount of failures observed in the field must be attributed to passive components, which are mounted close to semiconductor devices, like capacitors [70]. The limited maximum temperature rating of these devices (typically $85°C$) is among the factors, which slows the development of integrated packages for power semiconductors. This can also be the case of the gate control unit, which is normally mounted close to (or even integrated into) *IGBT* modules.

Large temperature swings mainly result into thermomechanical failure mechanisms (*see Chapter 2*), while high static temperature levels accelerate almost all failure mechanisms. Besides of time-dependent degradation mechanisms, silicon exhibits a maximum operating temperature, which can be estimated basing on the density of the intrinsic carrier density. In fact, when the intrinsic density reaches the doping level of the device, several electrical parameters are expected to change drastically. Among these there are the *multiplication factor for avalanche*

breakdown, the *carrier mobility* and *diffusion constants,* the *generation* and *recombination lifetimes*, the *thermal conductivity*, and the *MOS threshold voltage.* The maximum allowable temperature for a device also depends on the blocking voltage, which is specified for a given application. For instance, the estimated maximum junction temperature in silicon for a blocking voltage of *1 kV* is about *150°C*, while for *10 kV* it decreases down to *100°C*. Generally, the function of *IGBT* devices is limited by the leakage current, which increases exponentially with the temperature. As an example the typical reverse leakage current in a *1.2 kV IGBT* at a V_{CE} of *1200 V* is *60 μA* at *125°C*. By increasing the junction temperature up to *200°C* the leakage current causes the device to fail within some few seconds due to thermal runaway. Like in the case of power MOSFETs, the breakdown voltage of *IGBT* increases with the temperature.

4.3 Heat generation

Unlike in integrated circuits, in *IGBT* devices the heat is not generated at the surface of the device, only, but also in the silicon bulk. During pulsed operation, heat generation occurs due to both the voltage drop across the device and due to the switching power dissipated during the turn on and turn off phases. For a collector current density above $10A/cm^2$, the typical collector to emitter voltage is in the *4 V* range, which corresponds to a power dissipation of about *300 W* for each chip at the maximum rated current. The power dissipated during the turn on and turn off averaged over one period is in the range of *100 mWs*. However, especially during the turn on phase, the dissipated *peak power* may reach *several hundreds kilowatts*. It is interesting to note that for typical switching frequencies up to *3 kHz*, the junction temperature follows the switching frequency.

Under normal operating conditions the heat produced by ohmic dissipation within the bond wires can be almost neglected. However, in the case of severe bond wire lift off, the current can concentrate within some few bond wires such that the wire temperature may exceed the junction temperature. As it has been demonstrated by numerical simulation [71], the temperature distribution is raised mainly within the footprint of the bond wire and rapidly decreases down to the chip temperature level within a range corresponding to some few bond wire diameters. Therefore, since this effect is very localized, it mainly impacts

the thermomechanical degradation of the bond wires, rather than the chip surface temperature distribution.

For reliability purposes, the thermal coupling between *IGBT* chips and freewheeling diodes is usually neglected. This simplifying assumption bases on the fact that the heat flow within the module is approximately one-dimensional, and that the lateral spreading is non-negligible within the thick base plate, only. However, this assumption may be wrong in particular module designs where the few diodes are surrounded by multiple paralleled *IGBT* chips.

4.4 Thermal equivalent circuits

In *IGBT* devices, the heat generated within the semiconductor is mainly dissipated by conduction through the ceramic substrate and through the base plate from where it is released to the heat sink and finally to the environment by conduction, convection, and radiation. The junction temperature (T_j) in the simple case of a constant power dissipation P in the active layer is

$$T_j = T_a + P R_{th_{j-a}} \tag{4.1}$$

where T_a is the ambient temperature and $R_{thj\text{-}a}$ is defined as the thermal resistance (*junction-to-ambient*).

The dynamic thermal conduction behavior of the different package components is generally described for technological purposes by the means of equivalent thermal models. Similarly than in *Equation 4.1*, this concept is based upon an analogy between the electrical and the thermal properties of materials, with temperature, heat flow, and thermal impedance being analogous to voltage, current, and electrical impedance, respectively. In particular, the model of the transient thermal impedance is based on the analogy between one-dimensional heat conduction and distributed *resistor-capacitor* (*RC*) networks.

The typical structure of an *IGBT* multichip module is basically reported in *Table 2.1*. The multilayer can be subdivided into the elementary slabs, each being characterized by the material (density ρ_i, thermal conductivity σ_i, specific heat c_{thi}) and by the thickness t_i. The thermal resistance

associated with the whole stack of n layers in the lumped element equivalent circuit approximation is given by

$$R_{th_{j-a}} = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} \frac{1}{\sigma_i} \int_0^{t_i} \frac{dx}{A(x)} \tag{4.2}$$

where $A(x)$ is the effective conduction area discussed in *Section 4.6*. The equivalent circuit in the dynamic case is determined by the series connection of the impedances Z_{thi} related to each elementary plate, with

$$R_i = \frac{1}{\sigma_i} \int_0^{t_i} \frac{dx}{A(x)} \tag{4.3}$$

and

$$C_i = \rho_i \, C_{th_i} \, V_{effi} \tag{4.4}$$

V_{effi} in *Equation 4.4* is the effective conduction volume, which takes into account the lateral heat spreading effects (*see Section 4.6*). This yields the equivalent lumped thermal circuit in *Figure 4.1*.

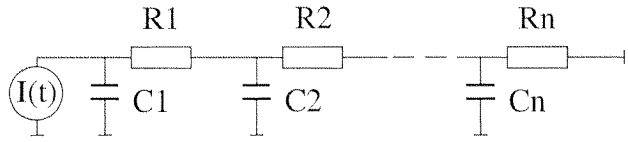


Figure 4.1 Equivalent thermal circuit of a multilayer structure

Each *RC pair* in *Figure 4.1* represents a physical layer and the time-dependent voltages at the different nodes represent the instantaneous temperatures at the interfaces. From a mathematical point of view the circuit in *Figure 4.1* is equivalent to a system of coupled differential equations, which can be integrated either numerically or *e.g.* by *PSPICE* simulation. In the last case, a current source of *1A* would correspond to *1 W* power dissipation in the upper layer, *i.e.* in the *IGBT* chip.

In practical applications the alternative equivalent thermal *RC circuit* in *Figure 4.2* is used.

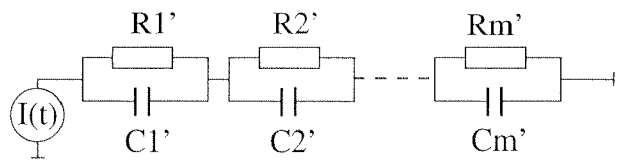


Figure 4.2 Behavioral model of the circuit in Figure 4.1

The circuit in *Figure 4.2* represents a *behavioral model* of the circuit in *Figure 4.1*, which mimes the temperature evolution in the node close to the power source, only. Since all *RC* sub-circuits are independent, there is a closed analytical form for the total transient impedance

$$Z_{th}(t) = \sum_{i=1}^{m} Z_{th_i} = \sum_{i=1}^{m} R_i' \left(1 - e^{-\frac{t}{R_i' C_i'}}\right) \tag{4.5}$$

The parameters R_i' and C_i' and the number of *RC* sub-circuits in *Figure 4.2* are not the same than in *Figure 4.1*. The extraction of the parameters R_i' and C_i' is performed by fitting the function $Z_{th}(t)$, which results either from the simulation of the circuit in *Figure 4.1*, or from the experimental heating (cooling) curve measured according to the procedure in *Section 4.9*. The optimum number of terms in the sum of *Equation 4.5* is usually determined from the number of linear regions occurring in a semilog plot of the transient thermal impedance as function of the time. The heating curve $T_H(t)$ at constant power injection P and for $T_H(0) = 0$ is defined by

$$T_H(t) = P\, Z_{th}(t) \tag{4.6}$$

If $Z_{th}(t)$ is known, the time evolution of the chip temperature T_{chip} due to an arbitrary power injection $P(t)$ and with $T_{chip}(0) = 0$, is computed by following convolution integral [72]

$$T_{chip}(t) = \int_0^t P(t) \frac{d}{dt} Z_{th}(t - \tau)\, d\tau \tag{4.7}$$

4.5 Evaluation of the heating curve

The extraction of Z_{th} from experimental data according to the procedure reported in *Section 4.4* requires the measurement of the heating curve of a

device. The measurement of heating curves presents some experimental problem. In fact, due to the temperature-dependent variation of sensitive parameters like V_{CEsat}, the injected power cannot be easily kept constant. In addition at high injection levels, the temperature-sensitive parameters used for thermometry purposes are poorly accurate. Therefore, the heating curve is inferred from the cooling curve basing on the linearity of the heat equation (superposition principle). As it will be shown in more detail in *Section 4.9*, the measurement of the cooling curve is made by heating the device to the steady-state, switching the power off, and monitoring the junction temperature as the device cools down. By assuming that the cooling curve $T_C(t)$ is the conjugate of the heating curve $T_H(t)$ delivers

$$T_H(t) = T_{steady} - T_C(t) \tag{4.8}$$

where T_{steady} is the steady state junction temperature immediately before switching off the heating power. This approximation is valid, if the thermal conductivity of the silicon and of the package materials do not vary excessively within the considered temperature interval. For junction temperatures in the $20°C$ to $125°C$ range, this is the case of all commonly used materials excepted silicon, which changes its thermal conductivity by about a factor of two. Nevertheless, the contribution of the silicon chip to the total thermal impedance is almost negligible.

4.6 Equivalent area and volume

The computation of a realistic transient thermal impedance requires estimating the equivalent area A_{eq} and the equivalent volume V_{eq} (*Equations 4.2*, *4.3*, and *4.4*). In fact, the one-dimensional equation for the thermal resistance

$$R_{th} = \frac{t}{\sigma A} \tag{4.9}$$

is no longer valid, since it does not take into account lateral heat spreading effects, which may occur in the thick base plate of an *IGBT* module, or in the heat sink. *Equations 4.2* and *4.3* already include the differential definition of the thermal resistance. However, in general the function $A(x)$ is not known in closed form.

The most common engineering estimates assume that the lateral heat spreading occurs under a constant spreading angle. One among the most invoked is the *45 degrees* model. This simple model enables to solve the integrals in *Equation 4.2* and *4.3*. For a square-shaped heat source of area A on the top of an homogeneous slab of thickness t the *45 degrees* model delivers

$$R_{th} = \frac{t}{\left(\sqrt{A} + 2t\right)\sigma\sqrt{A}} \tag{4.10}$$

and the equivalent volume

$$V_{eq}(t) = At + 2t^2\sqrt{A} + \frac{4}{3}t^3 \tag{4.11}$$

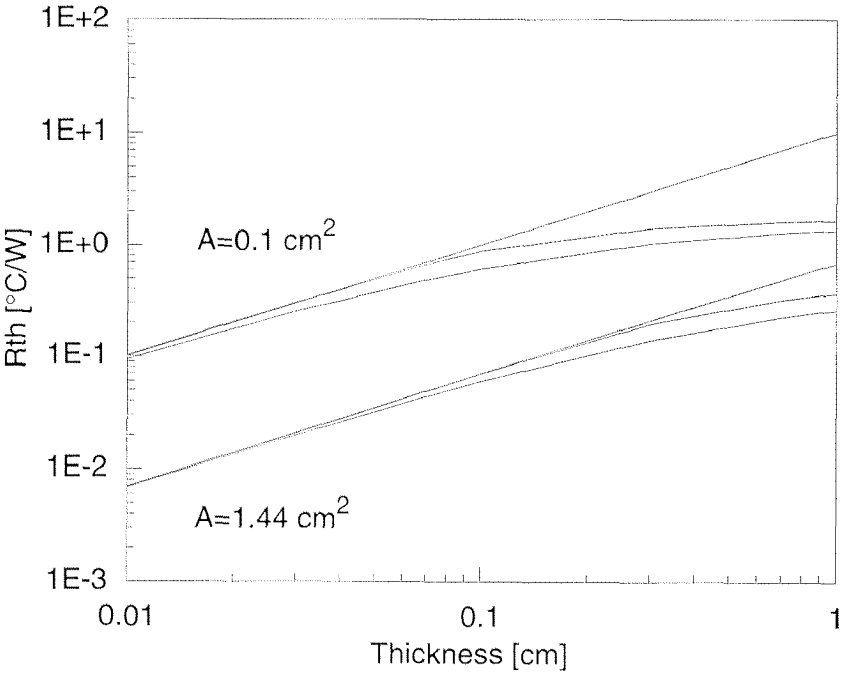


Figure 4.3 Thermal resistance of a single slab with σ = 1 W/cm°-C (temperature independent) for two heat sources with different area. In both cases the upper curve represents the one-dimensional estimate of Equation 4.9, the lower curve the estimate with the 45 degrees model of Equation 4.10, and the central curve the value obtained by the solution of the three-dimensional heat equation.

Figure 4.3 shows the comparison of the thermal resistance of a single slab obtained with the one-dimensional model of *Equation 4.9*, with the *45 degrees* model of *Equation 4.10*, and with the numerical solution of the *three-dimensional steady-state heat equation* (*3D*, [73]). By

considering the case of a square-shaped heat source with area *1.44 cm²* (*i.e.* the size of a typical *IGBT* chip), it can be seen that the heat conduction can be considered as one-dimensional up to a thickness of the slab up to *3 millimeters*. The *45 degrees* model dramatically underestimates the thermal resistance starting from a layer thickness of *1 millimeter*. Since the one-dimensional model delivers heavy overestimates of R_{th} for slab thicknesses larger than *3 millimeters*, alternative solutions are needed for modeling devices, which make use of thick base plates or of bulky heat sinks.

As shown in *Figure 4.3* for a source area of *0.1 cm²*, the boundaries of the validity regions of the different models strongly changes with the lateral size of the heat source. For sources with a lateral size of the same order of magnitude than the thickness of the slab, the three-dimensional effects become more evident. Unfortunately, also in this case the *45 degrees* model underestimates R_{th} by at least *25%* of the value obtained from the *3D-model.*

The *3D-curves* computed in *Figure 4.3* have general validity for a square source of area *A*, located on the top of a slab with a lateral size much larger than the *square root of A*. Since the curves have been evaluated for σ = *1 W/cm°-C*, the plotted values R_{th1} can be easily scaled for obtaining the thermal resistance R_{thx} for an arbitrary material with an arbitrary thermal conductivity σ_x

$$R_{th_x} = \frac{1}{\sigma_x} R_{th_1} \tag{4.12}$$

The equivalent conduction area can be defined starting from *Equation 4.9*

$$A_{eq}(t) = \frac{t}{R_{th_1}} \tag{4.13}$$

The equivalent volume can be computed basing on the equivalent spreading angle α_{eq}

$$\tan \alpha_{eq}(t) = \left[\frac{\sqrt{A_{eq}(t)} - \sqrt{A}}{2t} \right] \tag{4.14}$$

yielding

$$V_{eq}(t) = At + 2\sqrt{A}\,t^2 \tan\alpha_{eq} + \frac{4}{3}t^3 \tan^2\alpha_{eq} \qquad (4.15)$$

For the heat source of *1.44 cm*^{2} in *Figure 4.3*, *Equation 4.14* provides an equivalent angle of *0°* up to *1 mm* thickness, of *4°* at *3 mm*, of *7°* at *7 mm*, and of *12°* at *10 mm*. For a slab thickness of *1 cm*, the equivalent volume is about *40%* larger than the value provided by the one-dimensional model. On the other side the thermal resistance is *50%* lower than predicted by *Equation 4.9*.

More sophisticated techniques based on *CAD* assisted extraction of thermal parameters are illustrated in [74,75].

4.7 Experimental techniques for temperature measurement

The junction temperature of power semiconductor device can be measured *invasively* or *non-invasively*.

Invasive methods require that the surface of the semiconductor chip be exposed either for direct observation, or for depositing temperature-sensitive layers. This is the case of the liquid crystals microthermography (*see Section 3.4.2*) and of the induced fluorescence decay thermography (*see Section 3.10*).

Although infrared thermography (microradiometry) is a contactless technique, it has to be considered an invasive technique, since in the case of *IGBT* devices it requires special sample preparation. In fact, before acquiring any thermal map the device has to be depackaged, the silicone gel dissolved, and all the power lines within the optical sight field of the objective removed, in order to expose the chips to be observed (*see Section 3.3*). Furthermore, if quantitative measurements are required, the surfaces to be imaged must be coated with a special thin layer for equalizing the thermal emissivity of the different materials. All these factors hinder the device to be operated at voltage higher than *1.2 kV* or to be characterized in-situ when operated within a power system. Nevertheless, infrared thermography is still the only technique, which allows establishing the temperature distribution over large areas, as they occur in *IGBT* modules. The most radiometers provide steady-state measurement only, or they are limited to scan frequencies in the TV range.

Non-invasive techniques use a temperature-sensitive electrical parameter of the device as an integrated thermometer. For this reason the characterization can be performed on packaged devices, and can provide steady state or transient temperature information.

Basically, these techniques consist in two phases: the calibration, and the measurement itself. In this context the concept of junction temperature is ambiguous. In effect, *IGBT* chips do not exhibit just a single temperature, but they have a temperature distribution depending on different factors, like device design, power injected, current crowding, *etc.*. Since techniques using temperature-sensitive parameters indicate a single device temperature, this has to be intended as an average value over the temperature distribution. Typically, temperature-sensitive electrical parameters have a variation of few millivolts per centigrade, hence fast and accurate measurement circuits are required.

The terminal voltage method applied to *IGBT* devices is based on the peculiarity that, when injecting a constant (small) reference current into an *IGBT*, V_{CE} is a linearly decreasing function of the junction temperature. Being completely non-invasive, the terminal voltage method is often used as a diagnostic tool at the end of fabrication process, but it is also particularly suitable for failure analysis purposes [77]. In failure analysis this technique may give important semi-quantitative information about the integrity of the different interfaces inside and outside the module. For this reason the measurement of the thermal impedance by the terminal voltage technique has been proposed as a possible reliability indicator for the level of wearout of the device during the scheduled maintenance of power systems.

4.8 Characterization by IR thermography and calibration

Temperature mapping of the device is performed contactless by a *10-elements InSb high-resolution argon-cooled infrared radiometer Hughes TVS-2000* working in the *3-5.4 μm* spectral range. The on-line image processor converts the surface emission into an absolute temperature scale and provides a digital map of *100* pixels by *256* pixels at a scan rate of *30 frames per second*. At usual magnifications the lateral resolution is *100 μm* in the horizontal direction and *160 μm* in the vertical direction.
The accuracy of radiometric measurements made by infrared thermography is always questionable when the device under test presents

unknown or multiple local emissivity levels due to different materials. This can cause severe concerns about the validity of the results.

To overcome the problem of the different emissivity levels and to make quantitative measurement possible, the surface emissivity of the silicon chip and of all module materials has to be equalized by using a coating layer with appropriate physical characteristics. In the case of *IGBT* devices, the coating layer must be electrically insulating up to reasonably high voltages, have an emissivity close to one (black body), exhibit good wettability properties on the surface of usual electronic materials, and coat conformally the sample.

In present measurements, a coating layer has been used that we expressly developed for power device applications [20]. When preparing this coating layer, *one part* (vol) of a *10-wt.% polystyrene resin, 70-wt.% n-Butylmethacrylate polymer* and *20-wt.% carbon black* mixture is dissolved in *five parts* (vol) *acetone*. Once finely sprayed this solution behaves as a low-viscosity paint with excellent wettability over silicon and metal surfaces. Drying and curing for *4 hours* at $85\,^{\circ}C$ result in a compact and conformal layer of approximately *5 μm* thickness, which does not presents noticeable leakage currents at least up to *50 V*.

In order to determine the emissivity of the coating layer the *IGBT* module is mounted on to a thermochuck, which is set at a known temperature. Once the reference temperature is reached, the emissivity control of the radiometer is adjusted in order to obtain the correct radiometric temperature. The resolution is improved, if during this phase the temperature is set close to the junction temperature, which is expected during operation. In order to realize the required temperature accuracy (± *1%* up to $110\,^{\circ}C$) the temperature controller of the thermochuck has been calibrated referring to the phase transition temperatures of the set of liquid crystals reported in *Section 3.4.2*. Since the module is properly mounted on the thermochuck by using a thin layer of thermal grease, the temperature drop across the package while adjusting the emissivity is negligible.

During the previous phase a constant small reference collector current of *100 ± 0.5 mA* is injected while applying a gate voltage of *15 V* and the related V_{CE} is acquired.

Once coated and properly cured, the *IGBT* under test is mounted onto the heat sink without applying any power and after activating the water cooling system. After the thermal equilibrium is reached, the device

temperature is sensed at the same time by the terminal voltage technique and by the infrared thermograph as a last accuracy test. If the coating layer has been properly deposited, both values should be coincident within *1%* and every pixel of the infrared map should indicate the same temperature within the overall radiometric resolution of *1%*.

During the characterization, the constant injected power is monitored by a calibrated shunt resistor in series with the device, and by a voltmeter directly connected between the collector and emitter terminals of the *DUT*.

The typical emissivity of the coating layer as deposited and cure is *0.91* and does not depend on the temperature. In the case that the coating layer has not been properly cured, the emissivity may decrease over the time up to *5%* of the initial value. For this reason, it is recommended to re-calibrate the emissivity before every measurement. Temperatures exceeding the *110°C* have been noticed to induce visual and emissivity changes of the coating layer.

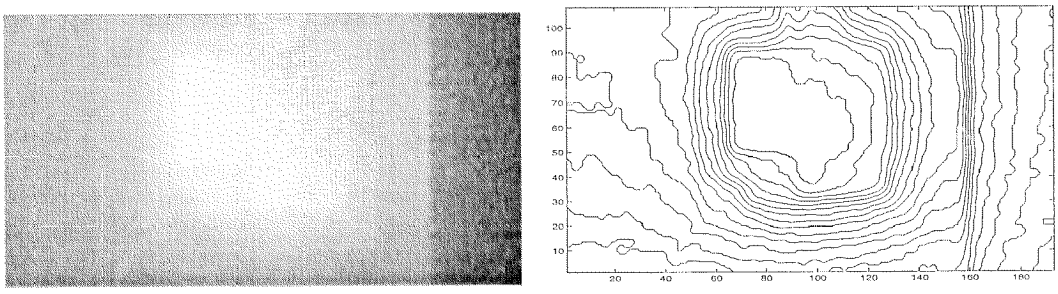


Figure 4.4 (a) Temperature map (gray levels) of a single IGBT chip dissipating 260 W (Infrared image 1.5x). (b) Isotherm plot of Figure 4.4a. The central area is at a temperature of 40°C, while the difference between two adjacent isothermal line corresponds to about 1°C.

Figure 4.4a shows an infrared map of a coated *IGBT* chip with a molybdenum strain buffer on the top. While measuring, the presence of such a plate has two beneficial effects. The first advantage is that it equalizes the temperature over a large surface, by eliminating all unwanted topography effects that may result in measurement artifacts. Secondly, the use of a molybdenum plate, enables to place the bond wires along a side of the chip, such that the surface of interest is left completely exposed as it can be clearly seen in the isothermal plot of *Figure 4.4b*. This fact increases the overall temperature resolution of the technique. Unfortunately, the surface of *IGBT* chips that do not use molybdenum

plates is almost completely hidden by numerous bond wires. The resulting shadowing effect is responsible for the large discrepancies between infrared and electrical measurements, which have been reported in previous works [7].

The black strip imaged on the right side of *Figure 4.4a* is a portion of the base plate. Since it is at a lower temperature than the chip and the ceramic substrate it is imaged as a darker region. The associated temperature gradient can be clearly recognized in the isothermal plot of *Figure 4.4b*. The thermal resistance of the different layers within the module can be roughly estimated by the temperature differences measured in corresponding flat areas of the image. On the contrary, temperature measurements at the edges are less reliable, since at these locations the infrared radiation is emitted under a different solid angle. However, the isothermal plot in *Figure 4.4a* indicates at least qualitatively where the main temperature gradients occur.

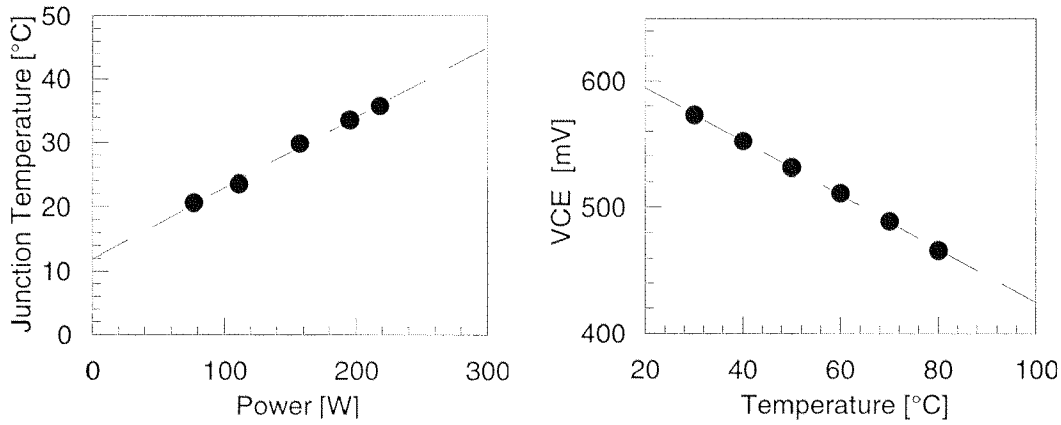


Figure 4.5 (a) Extraction of the thermal resistance of a single IGBT chip from the measurement of the junction temperature by infrared thermography. The regression delivers R_{th} = 0.11 ± 0.01 °C/W. (b) Temperature calibration of a single IGBT chip. V_{CE} is measured at a gate voltage of 15 V and at a constant collector-emitter current injection of 100 mA. The regression delivers a calibration factor κ = -2.1 mV/°C.

Figure 4.5a shows the procedure for extracting the thermal resistance of a single *IGBT* chip from the measurement of the junction temperature performed at different power levels. As expected, the measurements are distributed along a straight line, indicating that the technique is sufficiently accurate. The thermal resistance calculated from the slope of the straight line yields R_{th} = *0.11 ± 0.01 °C/W*. This value also includes the contributions of the thermal grease layer and of the heat sink. The

intercept indicates the temperature of the cooling water within the heat sink.

Figure 4.5b shows the calibration curve of a single *IGBT* chip, performed according to the procedure described above. The measurements are distributed along a straight line with negative slope, as one would expect from a simple diode. The regression yields the calibration factor $\kappa = -2.1$ *mV/°C*. Although this value is often measured (also in bipolar transistors), there is no universal calibration curve for *IGBT* devices. Therefore the calibration procedure has to be performed for every device under investigation. The calibration factor normally decreases by increasing the calibration current. Furthermore, for sensing currents exceeding *1 A* the relation between V_{CE} and the temperature may become non-linear and non-unique [76]. In general, the determination of the sensing current level is the result of a trade-off between resolution and device self-heating. Usual values of the sensing current are within *50* and *500 mA*. Unlike other electronic devices, *IGBT* modules do not require to be calibrated in an oven or in a thermostatic bath. In fact, the thermal contact provided through the base plate is good enough for conditioning the device just by a calibrated thermochuck. This is a great advantage, since it speeds up the calibration procedure and avoids the contamination of the device by thermal fluids, while keeping high levels of accuracy.

4.9 Measurement of the thermal impedance

The experimental set up presented in the following is the evolution of a similar equipment that was firstly developed for measuring the thermal impedance in optoelectronic devices [77] and later extended for characterizing low power *IGBT* devices [7]. The realized system acquires the cooling curve of a single high power *IGBT* chip, basing on the principle of the terminal voltage measurement. The used temperature sensitive electrical parameter is the forward voltage drop between collector and emitter V_{CE} at a constant injected current. The system provides both the time-resolved measurement of the junction temperature and the static thermal resistance.

Unlike in *MOSFETs* or in bipolar transistors, in *IGBT* chips there is no intrinsic diode directly available for internal temperature sensing. However, since the *IGBT* is a *MOS*-based device, additional temperature sensitive electrical parameters can be exploited as the gate to emitter

voltage V_{GE} or the saturation voltage at high current injection levels V_{CEsat}. V_{CEsat} is normally used for acquiring heating curves. Different techniques based on the transient measurement of V_{GE} are very common in industrial applications, because standards are available [78].

Although these techniques provide about the same accuracy, the selected terminal voltage method can be realized by less experimental efforts, since no sophisticated control systems are required. An extensive comparison of the different techniques applied to power transistors is given in [79].

The realization of the terminal voltage technique requires a dedicated circuitry, which performs four fundamental functions: generation of the power pulse, generation of the reference current, processing of the acquired voltage drop, and sequencing of the different phases. The layout of the experimental set up, which has been realized is schematically represented in *Figure 4.6.*

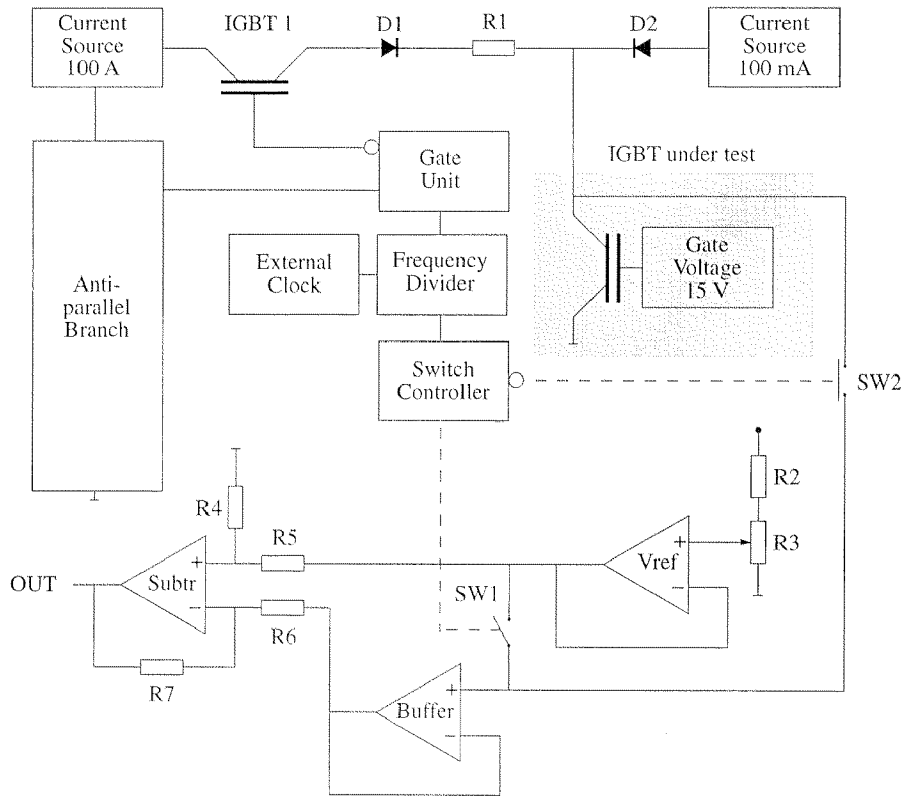


Figure 4.5 Schematic layout of the experimental set up for realizing the terminal voltage technique.

The pulse generation section consists of a *10 V* power supply, which delivers up to *100 A*. The voltage is chopped by *IGBT 1* (rated for *300 A*),

which is in conduction during the heating phase and blocking during the acquisition of the cooling curve. When *IGBT 1* is switched-off a similar *IGBT* (not shown in *Figure 4.5*) is synchronously turned on, which sinks the current into an anti-parallel branch that is perfectly symmetric to the measuring branch but thermally decoupled (shown as a box in *Figure 4.5*). The scope of the anti-parallel branch is to keep almost constant the current provided by the power supply in order to avoid overdriving effects. All switching devices are mounted on water-cooled heat sinks and the cables are kept as short as possible for minimizing the inductance. The power pulse generator can be operated with a repetition frequency of about *15 Hz*. The heating current is monitored by the voltage drop across the calibrated shunt resistor *R1* (*1 mΩ*). A water-cooled blocking power diode *D1* is connected in series with the *IGBT 1* in order to avoid the back flow of the sensing current during the turn-off (*D1* is optional if the anti-parallel branch is used).

The constant current generator is a critical component of the measuring set up, since it has to provide within *1%* accuracy the sensing current through the *DUT* shortly after the *IGBT 1* has turned off. In the system of *Figure 4.5*, it has been realized by a voltage regulator (*LM 317*) in constant current source configuration. In order to improve the time response of the regulator, the sensing current is provided continuously to the *DUT*, *i.e.* even during the heating phase. The sensing current has been measured to settle to the regulated value about *80 μs* after the *IGBT 1* turns off. The current output is set through a potentiometer to the same level that has been used for the calibration of the device (*i.e. 100 mA*, *see Section 4.8*). The diode *D2* avoids the backdriving of the regulator output during the turn on of *IGBT 1*.

As shown in the timing diagram in *Figure 4.6*, during the cooling phase the switch *SW1* is conducting (while *SW2* is open), such that the transient V_{CE} across the *DUT* is sensed by a buffer amplifier (impedance transformer). At the same time the reference voltage V_{ref} (adjusted by the voltage divider *R3* and *R4*) is sent together with V_{CE} to the inputs of a difference amplifier with *gain G* (typically *G = 100*), which delivers the output signal

$$V_{out}(t) = G\left(V_{ref} - V_{CE}(t)\right) \tag{4.16}$$

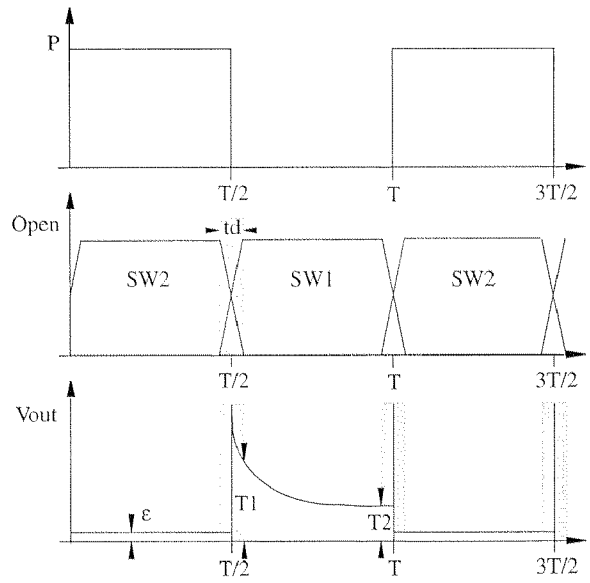


Figure 4.6 Schematic timing diagram of the experimental set up in Figure 4.5. From the top: heating power, timing of the switches SW1 and SW2, and output signal.

Now, if the set up in *Figure 4.5* is operated at $P = 0\ W$ (power supply off), at the thermal equilibrium the *DUT* will reach the same temperature T_a of the cooling water in the heat sink (ambient temperature). Therefore, if at the equilibrium V_{ref} is adjusted such that $V_{out} = 0\ V$, it will result into $V_{ref} = V_{CE}(T_a)$. Then, by applying after this adjustment a power $P \neq 0\ W$ yields

$$V_{out}(t) = G\left(V_{CE}(T_a) - V_{CE}(t)\right) = -\kappa G\left(T_j - T_a\right) \tag{4.17}$$

Combining *Equation 4.9* with the definition of the (conjugated) transient thermal impedance $Z_{ja}(t)$ for a cooling transient delivers

$$\overline{Z}_{ja}(t) = \frac{T_j - T_a}{P} = \frac{V_{out}}{-\kappa\, G\, P} \tag{4.18}$$

Thus, the transient $V_{out}(t)$, which can be monitored with an oscilloscope, is proportional to the conjugated transient thermal impedance. The proportionality constants κ, G are determined by the thermal calibration procedure in *Section 4.8* and by the gain of the difference amplifier, respectively. P is the power dissipated at the steady state during the heating phase and can be easily computed from V_{CEss} and I_{CEss} at the steady state as $P = V_{CEss}\, I_{Cess}$. I_{CEss} is measured across the resistor *R1* in *Figure 4.5*, while V_{CEss} is provided by a voltmeter connected at the collector and emitter terminals of the *DUT* (not shown in *Figure 4.5*).

The *CMOS* switches *SW1* and *SW2* in *Figure 4.6* have two functions. Firstly, *SW1* is opened during the heating phase, in order to hinder that the final stage of the amplification is brought in saturation by the large V_{CEsat}. This avoids introducing artifacts and delays due to the desaturation of the amplifiers during the following cooling phase. Secondly, *SW2* is closed during the heating phase and the voltage V_{ref} is applied at both differential inputs of the amplification stage. This enables again to avoid saturation and at the same time it enables to visualize and to correct the effect of eventual offset voltages of the amplification stage (ε in *Figure 4.6*).

In order to enable noise reduction by integration of successive transients, the measurement set up as been realized as a synchronous circuitry controlled by an external clock signal. The clock frequency is stepped down by a programmable frequency divider, which also defines the duty cycle of the heating and cooling phases (usually *0.5*). The generated signals are then sent to the switch controller, which generates the logic signals for *SW1* and *SW2 CMOS* switches, and to the gate unit, which provides proper gate signals for switching the *IGBTs*. The occurrence of different parasitic effects (*e.g.* inductive ringing, tail currents in *IGBT*, settling times in the amplifiers, *etc.*) introduces a dead time in the response of the set up (t_d in *Figure 4.6*). In the system configuration used for acquiring the cooling curves of *Figure 4.7*, the time required from the turn-off of *IGBT 1* until V_{out} is valid can be estimated in *100 μs*.

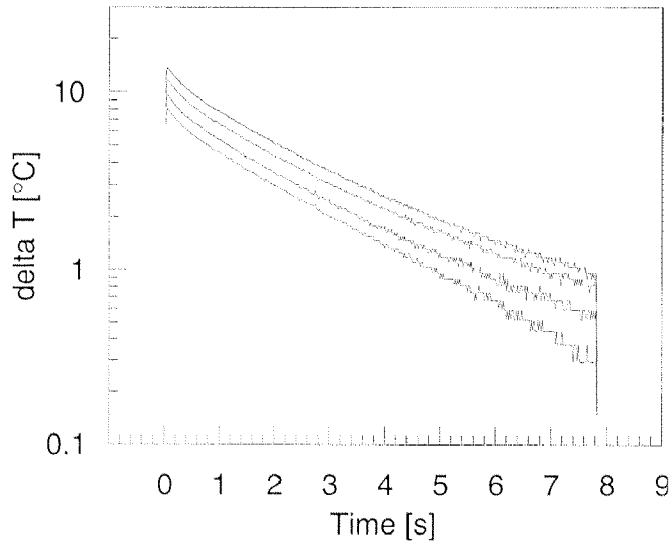


Figure 4.7 Cooling curves of a single IGBT chip at low power injection: starting from the top P = 124 W, 114 W, 98 W, and 86 W.

Figure 4.7 shows four typical cooling curves, which have been acquired with the experimental set up of *Figure 4.5* and converted to temperature transients by the calibration factor from *Figure 4.5b*.

The measurements have been performed at low power injection levels in order to check the accuracy of the technique. It can be clearly seen that the *signal-to-noise ratio* is in the *13 dB* range. The semilog plot of *Figure 4.7* indicates that the cooling curves are not pure exponentials.

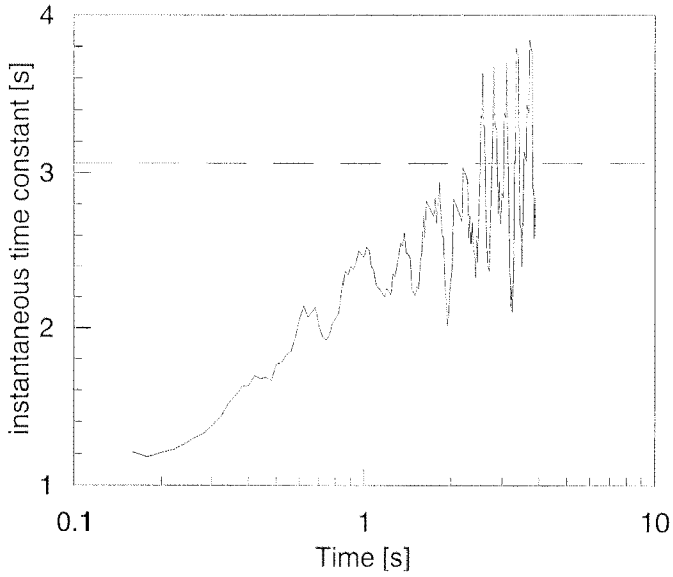


Figure 4.8 Instantaneous time constants from the cooling curve measured at 124 W in Figure 4.7.

Figure 4.8 shows the instantaneous time constant of the cooling curve, which has been acquired at *124 W* in *Figure 4.7*. After about *100 ms* from the removal of the heating power, the cooling curve exhibits a time constant of about *1 second*. The fact that the time constant increases with time and saturates after about *2 seconds* indicates the cooling process slows down until it becomes exponential. This behavior can be explained by a first phase, where the layers within the modules reaches relatively fast a thermal quasi-equilibrium, and by a second phase where the cooling is mainly governed by the total thermal capacitance and by the resistance of the heat sink.

Extrapolating the cooling curves back to $t = 0\ s$, it is possible to estimate the junction temperature T_0 at the steady state, *i.e.* immediately before that the heating power P was removed.

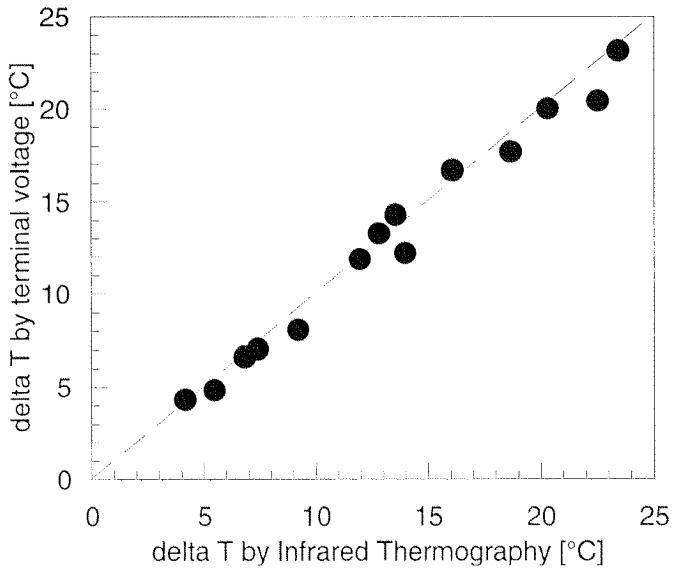


Figure 4.9 Correlation plot representing the junction temperature raise ΔT_j at the steady state (and at different P) measured by infrared microscopy and by the terminal voltage technique.

The correlation plot in *Figure 4.9* shows that the extrapolation of T_0 according to the slope at the very beginning of the cooling curve (in the *100* to *200* μs interval) agrees within *10%* with the measurement performed at the steady state by infrared thermography (at the temperature peak). This demonstrates that in present case the temperature drop during the dead time of the measurement set up is negligible. Furthermore, the fact that the measurements are statistically straggled around the diagonal is a clear indication that the average temperature value delivered by the terminal voltage technique is not systematically lower than the peak temperature measured by infrared thermography. This is can be explained by the averaging effect due to the molybdenum plate on the top of the *IGBT* chip.

Chapter 5

Modeling the Gate Oxide Reliability in IGBT Devices

Time dependent dielectric breakdown (*TDDB*) is a wearout mechanism of thin dielectric films stressed by an electric field, which consists in the sudden loss of the insulating properties of the dielectric. It can be distinguished into two phases. During the first phase oxide damage is accumulated within the thin film. The second phase is triggered as soon a critical level of oxide damage is reached, leading to a thermal runaway process, which results into the local destruction of the thin oxide layer. Under normal circuit operation conditions, *TDDB* is not observed in defect-free (intrinsic) gate oxides. However, if in oxide layers thicker than *20 nm* an electric field in excess of *8-9 MV/cm* is applied, the breakdown of the intrinsic oxide can occur. The critical reliability issue for gate oxides is the defect-related (extrinsic) breakdown. In fact, extrinsic breakdown causes random failures of the thin oxide even under normal operating conditions. In both cases, the time required by an oxide

to breakdown (t_{BD}) has a very strong dependence on the applied oxide electric field and also depends on the temperature.

Usually, *TDDB* is investigated under accelerated conditions, *i.e.* at high field stress and at elevated temperature. In order to extrapolate the high-stress data down to the normal operating conditions of a device, both the degradation mechanism and the statistical occurrence must be accurately known.

The scope of present Chapter is to develop an analytic model for predicting and for designing the field reliability of gate oxides used in *IGBT* devices basing on the results of accelerated tests. The usual questions of technical relevance to be answered are: How to condition samples for performing lifetime measurements ? How to perform accelerated tests under voltage acceleration ? How to describe and parameterize experimental data ? How to predict the lifetime of devices under normal operating conditions ? How to design a screening for realizing a pre-defined failure rate while minimizing the yield loss ?

For answering these questions, we will firstly review the main physical models for intrinsic breakdown, with particular emphasis on the *1/E*-model. Then we will introduce the field acceleration factors, basing on the effective thickness model and on additional empirical models. After that the *TDDB* will be investigated from the probabilistic point of view, in order to take into account both the effects of intrinsic and extrinsic breakdown. *TDDB* will be described with a statistical model, which bases on the *Weibull* distribution and we will derive analytically the most relevant reliability parameters. Moreover, lifetime models will be developed, which make use of experimental data provided by accelerated constant voltage and linear voltage ramp stresses. Finally, the developed models are will be used for computing the set on of the thin oxide wear out and for designing optimal screening procedures.

5.1 Phenomenology

The physics behind the wear out process, which leads first to the degradation and then to the breakdown of gate oxides submitted to high electric fields is rather complex, and up to now no model is universally accepted. However, there is general consense on the fact, that breakdown is the result of a continuous degradation of the volume-interface

characteristics of the dielectric [82]. The degradation behavior of gate oxides is strongly correlated with the local non-idealities of the dielectric. For this reason, the failure-free operating times of devices due to the *TDDB* has a statistical character highly dependent on the nature and on the density of oxide defects, which are either native or created during the stress.

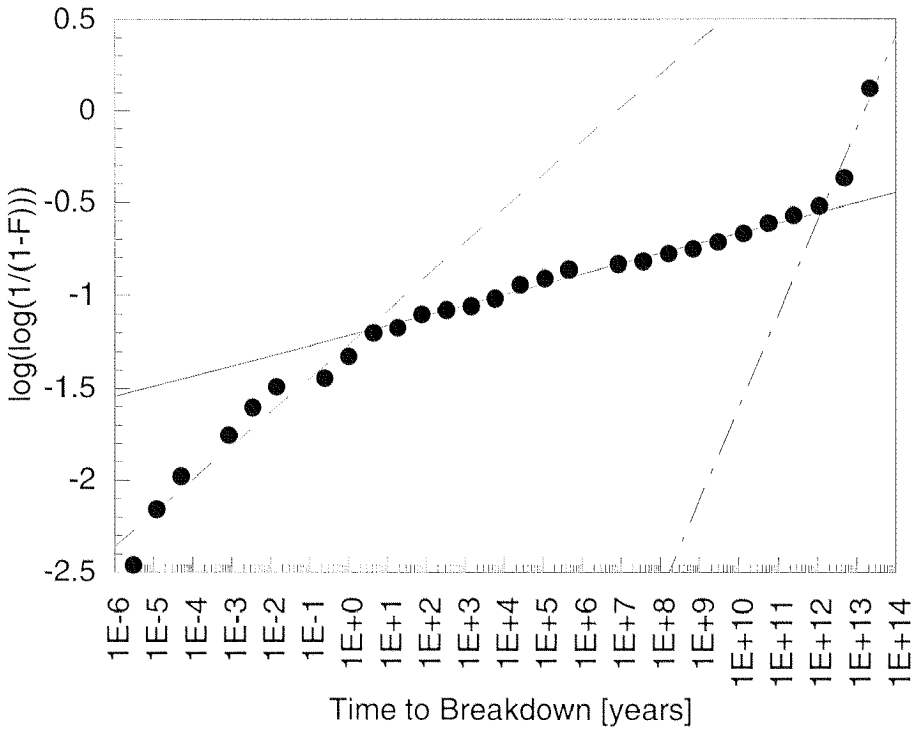


Figure 5.1 Cumulative distribution in Weibull representation (see Equation 5.31) of the failure-time of a thin gate oxide with t_{ox} = 50 nm, submitted to constant voltage ramp stress at 1V/s from [83] and converted to constant voltage stress at 3 MV/cm with Equation 5.5.

In *Figure 5.1*, we can distinguish three types of statistical populations, as suggested by the presence of three linear regions. The first sub-population with the lowest time to breakdown is associated with severe process flaws. The intermediate sub-population is due to less severe point defects caused again by technological processes. The last sub-population in *Figure 5.1* is related to the intrinsic properties of the dielectric (*e.g.* bonding energy, lattice strength, *etc.*). The technological causes associated with these types of oxide defects in power *BiMOS* structures have been investigated in very detail in [83], and generally for silicon technologies in [84,85]. In the case of voltage ramp tests, like that represented in *Figure 5.1*, defects are classified according to the maximum field strength. Three classes are usually defined. *Class A* includes those oxides, which fail for $E_{BD} < 1$ *MV/cm*, and which would

result into early failures in the field, if they were not screened. *Class B* includes those oxides, which fail at an intermediate E_{BD} lower than the intrinsic value and which are commonly called extrinsic oxides. *Class C* includes those oxides, which fail at the highest E_{BD}, due to intrinsic mechanisms. In the following, it is implicitly assumed that all oxides belong to the classes *B* and *C*, while *Class A* oxides have been preventively eliminated by an appropriate screening procedure.

5.2 Intrinsic Oxide Breakdown

Wearout phenomena are observed to occur in oxides prior to breakdown. There are four main physical parameters, which degrade during oxide stressing at high electric fields, and that are usually correlated with the occurrence of *TDDB*. They are the *interface trap density*, the *trapped oxide charge density*, the *hole fluence*, and the *density of neutral traps*. The relevance of every parameter and the related experimental evidence are discussed exhaustively in [86]. The time dependence of the *TDDB* is usually modeled by assuming that during stress, one among these parameters increases up to a fixed critical level. At this time, the local density of traps is sufficiently high to build a conductive path through the oxide and to lead to thermal destruction of the dielectric. Thus, the problem of the time degradation of the dielectric properties of an oxide is translated into the investigation of charge trapping phenomena and of the generation of traps. There are numerous models, which are in use nowadays. The most popular are:

Anode hole injection model [87] – This model assumes that, when the electrons injected from the cathode reach the anode with enough energy, they relax by impact ionization creating energetic holes, which can be injected back to the cathode. The hole current j_p is related to the electron current j_e by [88]

$$j_p = \alpha(E)\, j_e \tag{5.1}$$

where α is the probability for a tunneling electron to create a hole, which is injected back to the cathode. The holes that are generated can either be trapped [89], or create new traps within the oxide [90].

hydrogen release model [91] – In this model the electrons emitted by the cathode impinge in the anode with sufficient energy to release hydrogen at the anode-oxide interface. The free hydrogen ions diffuse towards the

cathode, creating charge traps within the oxide. This model is supported by experimental results, which demonstrate that oxides with high content of hydrogen show a reduced lifetime [92].

thermochemical model [93] – This model postulates that the trap generation mechanism does not depend on the carriers injected by the electrodes, but that traps are generated by the breakage of Silicon-Silicon bonds caused by the local electric field.

Electron trapping [94] – The model assumes that during stress no new trap is created and the pre-existing traps are just filled by electrons. The breakdown occurs as soon a critical amount of negative charge is reached. The model bases on very simple assumptions, but it neglects the fact that the density of the different trapping centers increases during stress.

Although these models seem to be very different in nature, almost all (except the thermochemical model) can be related more or less directly to a damage generated by hole transport through the oxide.

Dependence of t_{BD} on the electric field of intrinsic oxides – All previous models, excepted the *anode injection model*, results into a time-to-breakdown (t_{BD}) of the gate oxide, which is proportional to the electric field (E-model) [95]. On the contrary, the *anode injection model* provides a dependence, which is inverse proportional to the electric field ($1/E$-model). The E versus $1/E$ controversy continues for many years, due to the fact that both models can fit *TDDB* data rather well over a limited range of the electric field. Since the discussion about the choice of the $1/E$ or of the E model is outside the scope of this work, and since it is mostly arbitrary, we will consider in the following those aspects of both models, which can be used for heuristic modeling of the experimental data. Both models are deterministic, *i.e.* they deliver expressions for the time t_{BD}, without providing any information about the statistical distribution of the breakdown events.

Dependence of t_{BD} according to the 1/E-model – Basing on the fact that t_{BD} and hole generation rate have very similar dependencies on E, the model assumes that the rate of oxide damage (D) is proportional to the hole generation rate and breakdown occurs when a critical amount of damage (D_{crit}) has been sustained by the oxide [96]. The holes are assumed to be generated by impact ionization within the oxide, when

electrons tunnel from the cathode to the anode. The generation probability α has following field dependence [97]

$$\alpha(E) = K_0 \, e^{-\frac{H}{E}} \tag{5.2}$$

where K_0 is a constant and H is a parameter, which depends on the oxide thickness. Furthermore, the *Fowler-Nordheim* current j_{FN}, which is injected through an oxide if an electric field E is applied, is

$$j_{FN} = AE^2 \, e^{-\frac{B}{E}} \approx A' \, e^{-\frac{B}{E}} \tag{5.3}$$

where A and B are two constants related to the electron effective mass in the oxide conduction band and to the barrier height between Silicon and oxide. Since the quadratic term in *Equation 5.3* is a slower function of E than the exponential function, it has been approximated by a constant. The failure criterion can be rewritten in integral form as

$$D_{crit} = \int_0^{t_{BD}} j_{FN}(E)\,\alpha(E)\,dt \tag{5.4}$$

If we assume that the charge trapping is negligible during stress, *i.e.* j_{FN} and α do not depend on the time, the integration of *Equation 5.4* is straightforward and yields

$$t_{BD} = \frac{1}{D_{crit}} e^{\frac{B+H}{E}} \equiv \tau_o \, e^{\frac{G t_{ox}}{V_{ox}}} \tag{5.5}$$

In *Equation 5.5* the field oxide has been expressed through the applied voltage V_{ox} and through the oxide thickness t_{ox}. This approximation is correct, if V_{ox} is much larger than the flat band voltage. *Equation 5.5* represents an accurate model of t_{BD} for $t_{ox} < 7\ nm$, *i.e.* when charge trapping during the stress can be reasonably neglected. In this case, the room temperature values of the parameters τ_0 and G are in the *10 ps* and *350 MV/cm* range, respectively [98]. Although in thicker oxides, electron trapping plays a non-negligible role, it has been observed empirically [97], that t_{BD} still depends exponentially on the inverse applied electric field. However, in this case, the parameter G in the exponent of *Equation 5.5* can exceed *500 MV/cm*, and it has no direct physical interpretation. Thus, to predict the lifetime at low-field, one must rely to accelerated testing for extracting experimentally τ_0 and G. The usual procedure to be

followed in this case is to perform firstly a constant voltage ramp test (at a typical rate of *0.5 V/s*) for the determination of the maximum field strength E_{BD}.

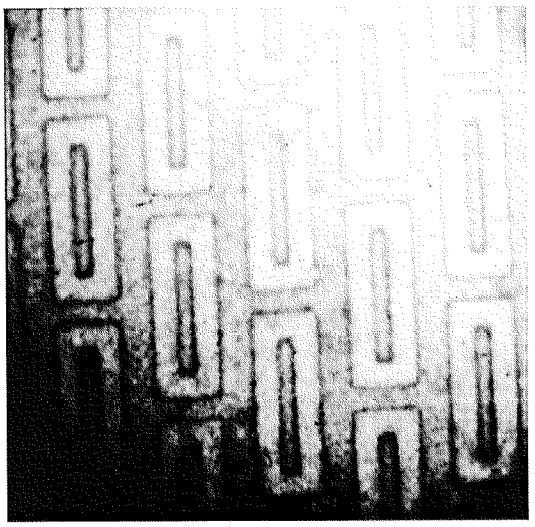


Figure 5.3 Emitter contacts of IGBT cells after selective removal of the aluminum metallization and deposition of a transparent gold layer. (Optical image 400 x).

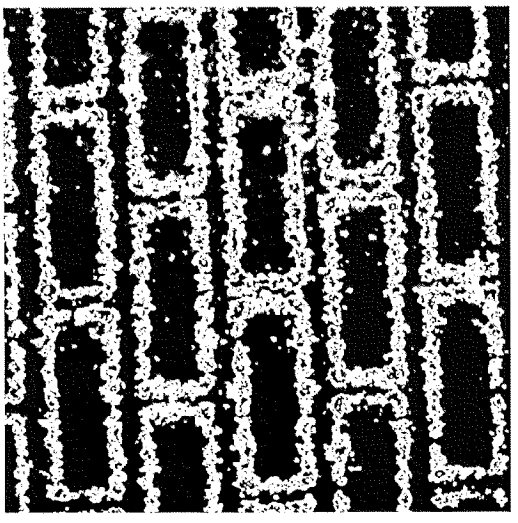


Figure 5.4 Light emission at the gate-emitter overlapping region of the IGBT cells represented in Figure 5.3 during Fowler-Nordheim injection through the gate oxide, at an injection current of 2 μA. (Emission microscopy image, 400 x)

The second step consists in a constant voltage stress realized at an initial electric field, which does not exceed *80%* of E_{BD}. τ_0 and G should be then extracted from the sub-population, which belongs to *Class C*, according to previous definition. Both tests should be carried out preferably at a

temperature, which is representative for the operation conditions of the device. In this way, incertitude due to unknown temperature acceleration factors is simply avoided. Measurements on gate oxides with $t_{ox} = 35\ nm$ indicate values of τ_0 in the *60 ps* range and G in the order of *330 MV/cm* [99].

In *Figure 5.4* and *5.5* the emitter metallization of an *IGBT* device has been removed selectively and a semi-trasparent gold window has been deposited according to the procedure presented in *Section 3.5.4*. After applying a field oxide of *6.5 MV/cm* across the gate oxide, the electro-luminescence arising from *Fowler-Nordheim* injection in the channel regions can be easily detected.

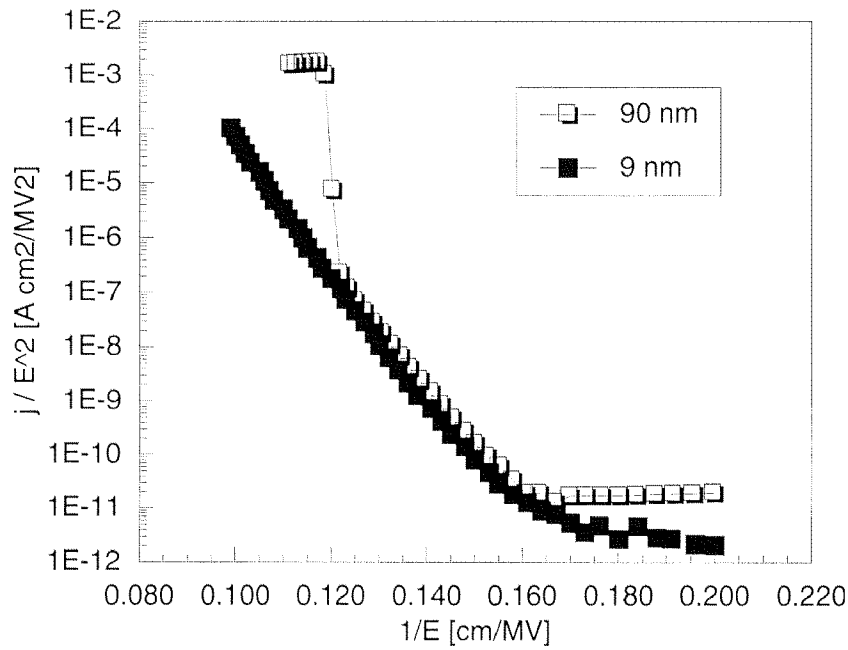


Figure 5.2Fowler-Nordheim characteristics for a thin (9 nm) and a thick (90 nm) oxide. At high electric field, the characteristic of the thick oxide deviates from the behavior predicted by Equation 5.3, due to enhanced hole trapping.

The model of *Equation 5.5* holds, if the injected current follows the *Fowler-Nordheim* behavior indicated in *Equation 5.3*. *Figure 5.2*, which represents the *Fowler-Nordheim* injection characteristic of a *9 nm* oxide compared with that of a *90 nm* oxide, shows that this is not the case for thick oxides with $t_{ox} > 14\ nm$. In fact, we can observe, that for inverse fields ranging between *0.12* and *0.16 cm/MV*, both curves have the same slope, showing that both samples are in *Fowler-Nordheim* injection

regime. For higher fields, the curve related to the thicker oxide exhibits a steeper slope, while the injection in the thinner oxide continues to be represented by *Equation 5.3*, (*i.e.* a straight line in the *Fowler-Nordheim* representation of *Figure 5.2*). This deviation from the *Fowler-Nordheim* behavior, which is observed in thicker oxides, is usually attributed to enhanced trapping of positive charges (holes) [80,81]. Thus, it is evident, that the use of *Equation 5.5* for extrapolating the lifetime down to low field conditions of gate oxides, which have been stressed at a regime where hole trapping dominates, will result into a pessimistic estimate of t_{BD}. For this reason, constant voltage stresses of thin oxides must be performed at a field that is below the threshold field E_{th} at which such an effect occurs [81]. This is the reason of the field scaling recommended above. The value of E_{th} for highly intrinsic oxides as a function of the thickness can be approximated by the maximum field strength E_{BD}. By using an empirical expression for $t_{ox} > 20\ nm$, and basing on the data in [100], we have

$$E_{th} \approx E_{BD} = 8.4 + \frac{10^{-5}}{t_{ox}} \quad [MV/cm] \tag{5.6}$$

Equation 5.6 describes the case of a constant voltage stress. However, t_{BD} can also be extracted with faster stresses, like linear or logarithmic voltage ramps. In fact, within the range of validity of *Equation 5.3*, we can rewrite the integral of *Equation 5.5* in order to compute t_{BD} for an arbitrary stress $E(t)$, which depends explicitely on the time. For this scope, we introduce again a normalized damage function $D(t)$, which is one when the cumulated damage reaches D_{crit} (*i.e.* at t_{BD}). The assumption of linear damage cumulation yields

$$1 = \int_0^{t_{BD}} dD = \int_0^{t_{BD}} \left(\tau_o \, e^{\frac{G\, t_{ox}}{V_{ox}(t)}} \right)^{-1} dt = \frac{1}{\tau_0} \int_0^{t'_{BD}} e^{-\frac{G\, t_{ox}}{V_{ox}(t)}} \, dt \tag{5.7}$$

In the elementary case of a linear voltage ramp with slope r we have $V_{ox}(t) = r\,t$, The parameter measured during this test is the voltage to the breakdown $V_{BD} = V_{ox}(t'_{BD}) = r\,t'_{BD}$. The integral in *Equation 5.7* has no analytical solution. The approximated solution of *Appendix 2* yields

$$\ln(t_{ox}) + \frac{G}{V_{BD}} t_{ox} - \ln\left(\frac{V_{BD}^2}{r\,\tau_0 G} \right) = 0 \tag{5.8}$$

If V_{BD} is measured, the numerical solution of the transcendental *Equation*

5.8 delivers t_{ox}. The computed value of t_{ox} can be then inserted into *Equation 5.5* for obtaining the lifetime t_{BD} of the oxide at constant voltage stress.

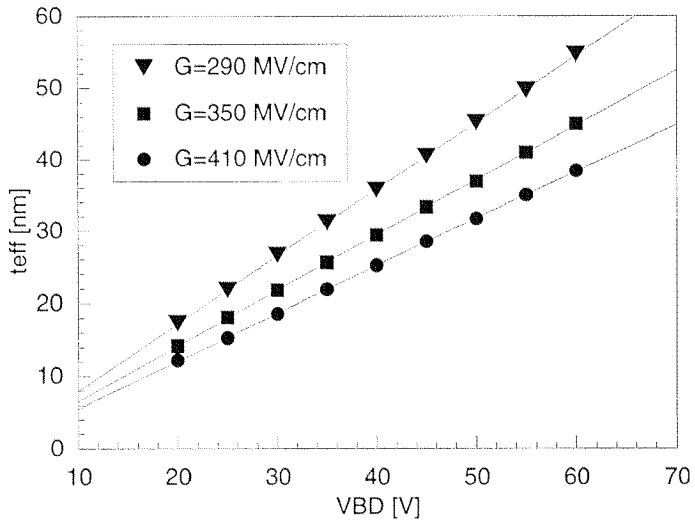


Figure 5.5 Numerical solution of Equation 5.8 for different values of G and for τ_0 = 10 ps

Figure 5.5 represents the solutions of *Equation 5.8*, as a function of V_{BD}. We can see that for $t_{ox} > 16$ *nm*, t_{ox} increases linearly with V_{BD} over a broad range of V_{BD}.

It is important to notice, that in the case of extrinsic oxides, the value of t_{ox} computed according to *Equation 5.8* is smaller than the nominal oxide thickness. However, if the most severe defect that caused the breakdown event is interpreted as a local oxide thinning, t_{ox} can be regarded as an effective oxide thickness t_{eff}. The effective oxide model is investigated in more detail in *Section 5.11*.

Temperature acceleration factor – Oxides are more vulnerable to hole transport at higher temperatures. This is due to the fact that holes produce enhanced damage at higher temperature [99]. t_{BD} depends on the temperature over the parameters τ_0 and G, that is

$$t_{BD}(T) = \tau_0(T)\, e^{\frac{G(T)\, t_{ox}}{V_{ox}}} \tag{5.9}$$

where T is the absolute temperature, k the Boltzmann constant ($8.6\ 10^{-5}$ *eV/K*) with [101]

$$\tau_0(T) = 5.4 10^{-7}\ e^{\frac{0.28}{kT}} \quad [s] \quad G(T) = 120 + \frac{5.8}{kT} \quad [MV/cm] \tag{5.10}$$

The activation energy E_a of oxides thicker than *10 nm*, has been found to increase with t_{ox} (and thus with t_{BD}) according to the empirical relation

$$t_{BD}(T) \propto \quad e^{\frac{E_a}{kT}} \tag{5.11}$$

where

$$E_a = 5.8 \frac{t_{ox}}{V_{ox}} - 0.28 \quad [eV] \tag{5.12}$$

Dependence of t_{BD} according to the E-model – The E-model, as it has been introduced by [102], is based on empirical observations and postulates that the time t_{BD}, which is necessary for a given quantile of the distribution to be failed, depends exponentially on the applied electric field E

$$t_{BD} \propto e^{-\gamma E} \tag{5.13}$$

Thus, if t_{BD0} is the time to the failure measured at field E_0, t_{BD1} at a field E_1 can be extrapolated by

$$t_{BD1} = t_{BD0}\ e^{-\gamma(E_1 - E_0)} \tag{5.14}$$

The field acceleration factor γ strongly depends on the oxide thickness. An empirical relation for γ at room temperature as a function of t_{ox} (valid for $t_{ox} > 15$ *nm*) can be derived empirically from the measurements cited in [103]

$$\gamma(t_{ox}) = 5.3 \ln(t_{ox}) + 76.1 \tag{5.15}$$

where the oxide thickness is expressed in *cm*.

Like in the case of the *1/E* model, the field acceleration factor depends on the absolute temperature *T*. Empirical relations describing the behavior of t_{BD} as a function of *T* are given in [95].

5.3 Breakdown of extrinsic oxides

Oxide degradation depends largely on the distribution of the oxide breakdown voltage, which is dependent on the defect density. The defect density is affected by the gate electrode material, by the thickness of the film, by the defects in the substrate, by the oxidation method, by the implantation steps, and by contamination. They are also affected by the three dimensional structure of the gate area that is based both on design and manufacturing process. The technological causes associated with these types of oxide defects have been investigated for power *BiMOS* structures in [83], and generally for silicon technologies in [84,85].

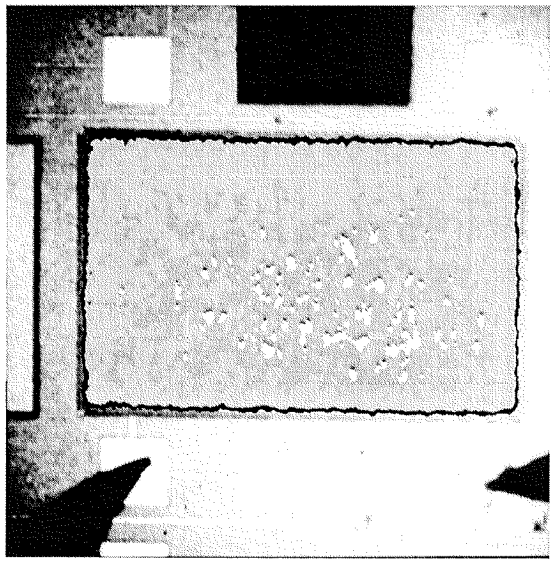


Figure 5.6 Emission image of a MOS capacitor with t_{ox} = 50 nm during Fowler-Nordheim injection before breakdown. The electro-luminescence within the bright regions is more intense, indicating enhanced current injection. (Emission microscope image, 100x)

In the *MOS* capacitor of *Figure 5.6*, electrons are injected from the substrate to the polysilicon anode by *Fowler-Nordheim* tunneling. The brighter locations in the electro-luminescence image indicate regions with enhanced *Fowler-Nordheim* injection (probably due to the thinner oxide). *Figure 5.7* shows the same capacitor after breakdown. The light emission

just occurs at the breakdown site, which does not correlate with the region of enhanced emission in *Figure 5.6*. This illustrate the fact that, if there is no systematic oxide thinning effects (*e.g.* bird's beak) or heavy metal contaminations, the most severe oxide defects, which lead to the breakdown are very localized weak points.

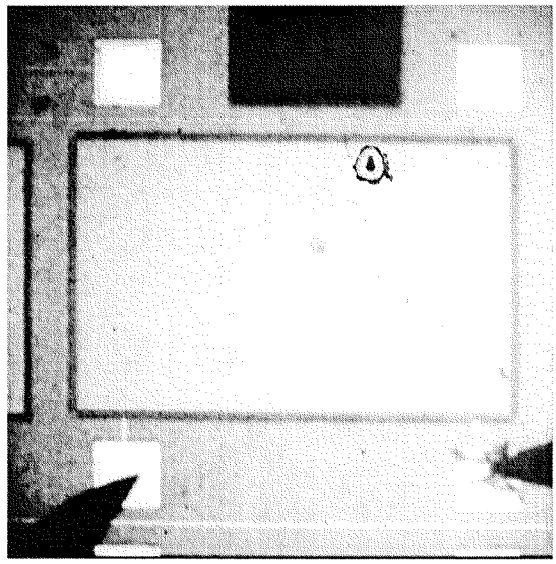


Figure 5.7 Emission image of the same MOS capacitor than in Figure 5.6 after breakdown (emitting spot). The breakdown site is indicated by the emitting spot, which is not located where enhanced electro-luminescence has been observed (emission microscope image, 100x).

5.4 Probabilistic Model

In recent years, a probabilistic model of *TDDB* that combines intrinsic and extrinsic breakdown has been proposed on the base of heuristic considerations [104]. In the following, we will reconsider critically the probabilistic background of this model, in order to derive analytically relevant reliability parameters, like the failure rate and the survival probability. In particular, we will clearly distinguish between the formalism for the *mixture of distributions* and the distribution arising from *competing risks*.

If we consider a population of n devices, which are operated in the field, we can distinguish two cases. In the first case, the n devices are subdivided in m sub-populations, each of them failing due to different failure mechanisms. In the second case, every device is affected by m failure mechanisms, and the operating life of every device is terminated

by the failure mechanism, which occurs at first. Although these two situations seem very similar, from a statistical point of view they are substantially different. In the first case one speaks of *mixture of distributions*, while in the second case of *competing risks*.

Mixture of distributions – If the failure-free operating time due intrinsic and the extrinsic breakdown is described by the probability density function f_i and f_e, respectively, and the occurrence probability of both breakdown types are p_i and p_e, respectively, we can write the resulting probability density function as

$$f(t) = p_i f_i(t) + p_e f_e(t) \tag{5.16}$$

where

$$p_i + p_e = 1 \tag{5.17}$$

Basing on the relations in *Appendix 1*, we have for the cumulative distribution function

$$\begin{aligned} F(t) &= \int f(t)\,dt = p_i F_i(t) + p_e F_e(t) = p_i\left(1 - R_i(t)\right) + p_e\left(1 - R_e(t)\right) \\ &= 1 - \left(p_i R_i(t) + p_e R_e(t)\right) \end{aligned} \tag{5.18}$$

for the reliability function

$$R(t) = p_i R_i(t) + p_e R_e(t) \tag{5.19}$$

and for the failure rate

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{R(t)} = \frac{p_i f_i(t) + p_e f_e(t)}{p_i R_i(t) + p_e R_e(t)} \\ &= \frac{p_i}{p_i R_i(t) + p_e R_e(t)} \lambda_i(t) + \frac{p_e}{p_i R_i(t) + p_e R_e(t)} \lambda_e(t) \end{aligned} \tag{5.20}$$

From *Equation 5.19*, it can be seen that the survival probability decreases proportionally to the slower reliability function, while from *Equation 5.20*, it is clear that the resulting failure rate is a weighted sum over the failure rate of every failure mechanism.

Competing risks – Let now consider a single device. If the failure-free time due to the intrinsic breakdown is the random variable τ_i and that due to the extrinsic breakdown τ_e, the device will fail for $\tau = \min(\tau_i, \tau_e)$. In other words, the reliability block diagram of the device is constituted of two items connected in series, each of them with its own distribution. The reliability function of such a series structure is given by

$$R(t) = R_i(t)\,R_e(t) = 1 - F(t) \tag{5.21}$$

the failure rate by

$$\lambda(t) = \lambda_i(t) + \lambda_e(t) \tag{5.22}$$

and the probability density function

$$f(t) = R(t)\,\lambda(t) = R_i R_e\,\lambda_i + R_i R_e\,\lambda_e = R_i(t) f_e(t) + R_e(t) f_i(t) \tag{5.23}$$

where f_i and f_e are the density functions associated with the intrinsic and the extrinsic mechanism, respectively. Both models can be easily generalized to the case of k different failure mechanisms with their associated probability density functions f_j.

The *mixture of distributions* is suitable for describing the time evolution of an inhomogeneous set of devices composed by k different lots with different quality. On the contrary, the *competing risk* formalism is more adequate to model a homogeneous set of devices characterized by a statistical defect distribution. Since the last is the most interesting case in reliability testing of thin oxides, it will be treated in more detail in the following.

5.5 The Statistical Model

The scope of this section is to describe quantitatively the cumulative distribution of the failure-free lifetime of an homogeneous population of thin oxides, which has been measured under accelerated stress conditions (*e.g.* enhanced gate voltage). The selection of the distribution for modeling the measured data is often arbitrary. Once the model has been selected the related parameters are firstly estimated by the maximum likelihood technique and then the goodness of the fit is quantified by an

appropriate hypothesis (*e.g. Kolmogorov-Smirnov* test). Nevertheless, it is common to use physics-related distributions. There is enough experimental evidence that the empirical cumulative distribution of the failure-free times measured during *TDDB* experiments can be approximated at least piecewise by a *Weibull* distribution (*Figure 5.1*) as it is defined in *Appendix 1*, or in the equivalent alternative form

$$F(t)=1-e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{5.24}$$

where η is the reciprocal of the scale factor λ.

From a physical point of view, this is an expression of the fact that the *TDDB* is a process governed by extreme values, *i.e.* by the weakest oxide at a given time. Since the weak sub-populations can be filtered out either through appropriated data censoring or by adequate screening of the devices, we can restrict our discussion to the case of a bimodal distribution, without loss of generality. Thus, we can assume that the overall empirical cumulative distribution is the result of the competition of intrinsic and extrinsic breakdown, being both processes described by their own *Weibull* distribution of the form

$$F_e(t)=1-e^{-\left(\frac{t}{\eta_{e0}}\right)^{\beta_{e0}}} \tag{5.25}$$

for the extrinsic part, and

$$F_i(t)=1-e^{-\left(\frac{t}{\eta_{i0}}\right)^{\beta_{i0}}} \tag{5.26}$$

for the intrinsic part.

The point estimate of the scale and shape parameters for both distributions can be extracted analytically from the experimental data, according to the procedure presented in *Section 5.6*.

5.6 Application of the properties of the Weibull distribution

The *Weibull* distribution as it has been introduced in *Equation 5.24* is the generalization of the exponential distribution. For a shape factor β less than one (more than one), it results into a monotonically decreasing (increasing) failure rate. The *Weibull* distribution has very useful properties, especially when investigating series models whose elements have independent failure-free operating times τ, which are *Weibull*-distributed. In fact, it can be shown [105,106] that the distribution of the smallest τ related to every series elements is again a *Weibull* distribution. In particular, the reliability function of n identical series elements can be written as

$$R_n(t) = \left(R_1(t)\right)^n = \left(e^{-\left(\frac{t}{\eta}\right)^{\beta}}\right)^n = e^{-\left(\frac{t}{\eta'}\right)^{\beta}} \tag{5.27}$$

where

$$\frac{1}{\eta'} = \frac{1}{\eta} n^{1/\beta} \tag{5.28}$$

Thus, referring the failure rate λ_n of the n series element to that of a single element λ_1 yields (as expected also from *Equation 5.22*)

$$\frac{\lambda_n(t)}{\lambda_1(t)} = n \tag{5.29}$$

Similarly, building the ratio of the survival probability R_n of n identical elements in series with the survival probability of a single element R_1 results in

$$\frac{R_n(t)}{R_1(t)} = \left(R_1(t)\right)^{n-1} \tag{5.30}$$

Moreover, since the shape factor β is the same for a single element and for n series elements, the related cumulative distributions F_1 and F_n are parallel straight lines in the *Weibull* representation

$$W_j(t) = \log\left(\log\left(\frac{1}{1-F_j(t)}\right)\right) \tag{5.31}$$

The vertical shift between the two cumulative distributions in the *Weibull* representation is easily computed in

$$W_n(t) - W_1(t) = \log(n) \tag{5.32}$$

In a similar way, we can compute the horizontal shift occurring between the time t_n, required for realizing the f quantile of failure in the series structures of n elements, and the time t_1, required by a single structure for reaching the same quantile of failure. In the *Weibull* representation we have

$$\log(t_1) - \log(t_2) = \log\left(n^{1/\beta}\right) \tag{5.33}$$

Equation 5.33 yields the time transformation

$$\frac{t_n}{t_1} = n^{-1/\beta} \tag{5.34}$$

Equations 5.28 through *5.34* can be directly used for extrapolating the reliability of n identical *IGBT* chips operated in the field, if the reliability of a single chip is known. We assume here, that the cumulative distribution is a *Weibull*. In the case of competing extrinsic and intrinsic risks, the resulting cumulative distribution is reasonably approximated by a single *Weibull* distribution (the extrinsic distribution) up to the turning point expressed by *Equation 5.54*.

A further application is the extrapolation of the reliability of a real device with gate oxide area A_D from accelerated tests performed on gate oxide test capacitors of area A_C. For this scope, the variable n has to be re-defined as

$$n \approx \frac{A_D}{A_C} \tag{5.35}$$

The area A_C is the result of a trade off, such that it is small enough for satisfying reasonably *Equation 5.35*, and it is large enough to produce a defect-related distribution, which is representative for macroscopic devices. By inserting *Equation 5.35* into *Equation 5.34*, we have

$$\frac{t_n}{t_1} = \left(\frac{A_C}{A_D}\right)^{1/\beta} \tag{5.36}$$

A similar dependence has been observed experimentally [83] to occur also in the case where the oxide defects are distributed along the perimeter of the gate. In this case, the gate oxide area has to be replaced in *Equation 5.20* by the perimeter L, yielding

$$\frac{t_n}{t_1} = \left(\frac{L_C}{L_D}\right)^{1/\beta} \tag{5.37}$$

Analytical extraction of the scale and of the shape parameters – If the scale and the shape parameters of the extrinsic and intrinsic distributions of *Equations 5.25* and *5.26* are well separated, we can compute analytically the related point estimate $\hat{\eta}$ and $\hat{\beta}$ from the initial (extrinsic) and the final (intrinsic) tails of the measured cumulative distribution. The proposed procedure is based on the fact that, if represented in the appropriate co-ordonates system, a *Weibull* distribution appears as a straight line with slope A_W and intercept B_W. In this co-ordinates system A_W and B_W can be estimated according to the least-squares criterion, and thus they can be expressed in terms of the empirical correlation and of the regression coefficients. From *Equation 5.31* we have that the searched co-ordinates transformation is

$$x_i = \log(t_i) \text{ and } y_i = W(F_i)$$

and after introducing the mean values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.38}$$

we have

$$\hat{\beta} = A_W = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5.39}$$

and

$$\hat{\eta} = 10^{-\frac{1}{\hat{\beta}}\left(B_W - \log(\log(e))\right)} \tag{5.40}$$

where

$$B_W = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5.41}$$

5.7 Lifetime Prediction

In the following we will discuss the statistical description of data from accelerated experiments.

The model *Equations 5.21* through *5.23* combined with *Equation A1.5* through *A1.9* in *Appendix 1* delivers the exact form of the resulting cumulative distribution

$$F(t) = 1 - e^{-\left(\frac{t}{\eta_e}\right)^{\beta_e} - \left(\frac{t}{\eta_i}\right)^{\beta_i}} \tag{5.42}$$

density function

$$f(t) = \left(\frac{\beta_e}{\eta_e}\left(\frac{t}{\eta_e}\right)^{\beta_e - 1} + \frac{\beta_i}{\eta_i}\left(\frac{t}{\eta_i}\right)^{\beta_i - 1}\right) e^{-\left(\frac{t}{\eta_e}\right)^{\beta_e} - \left(\frac{t}{\eta_i}\right)^{\beta_i}} \tag{5.43}$$

reliability function

$$R(t) = e^{-\left(\frac{t}{\eta_e}\right)^{\beta_e} - \left(\frac{t}{\eta_i}\right)^{\beta_i}} \tag{5.44}$$

and failure rate

$$\lambda(t) = \frac{\beta_e}{\eta_e}\left(\frac{t}{\eta_e}\right)^{\beta_e - 1} + \frac{\beta_i}{\eta_i}\left(\frac{t}{\eta_i}\right)^{\beta_i - 1} \tag{5.45}$$

Equations 5.42 through *5.45* are of practical interest, since they enable to compute almost all relevant reliability parameters starting from experimental data. The computation is purely analytical and avoids the well-known numerical instabilities due to the numerical processing of very large number required by the traditional approaches.

5.8 Lifetime prediction by using an invariance principle

For sake of simplicity, we focus our attention to the case of a single-mode distribution (*e.g.* the extrinsic population). We assume that while stressing this population at a given gate voltage, both the weaker and the stronger oxides are degraded by the same failure mechanism and that the time to the failure of each individual only depends on the gravity of the most severe defect by which it is affected. As it has been shown in [105], this process is described by the extreme value statistics and the failure-free times associated with every individual are *Weibull*-distributed. If this experiment is repeated under the same experimental conditions but at a lower gate voltage than before, we will obtain in general higher values of the failure-free times. However, since the failure mechanism is assumed to be always the same, the extreme value statistics will apply also to this case and the failure-free times will be again *Weibull*-distributed. This fact correlates with the experimental evidence. The invariance of the statistics is used here for deriving the general form of the transformation, which enables to extrapolate the lifetime measured with accelerated tests down to operating conditions. The statistical considerations above imply that there is a function, which transforms the time t_a measured at accelerated stress conditions into the time variable t_{op} at operating conditions, such that the distribution function is kept invariant. If we take into consideration the non-linear time function

$$t_a \mapsto t_{op} = t_0 \left(\frac{t_a}{\tau} \right)^p \tag{5.46}$$

where t_0 an τ are both constant (with the dimension of the time variable), one can easily show that it is an invariant co-ordinate transform, which, once applied to a *Weibull* distribution, provides again a *Weibull* distribution. The transform in *Equation 5.46* is quite general and corresponds to a linear time transform in the *Weibull* representation.

For $p=1$ the transform *5.46* delivers a time scaling by the constant factor

t_0/τ, while keeping unchanged the shape factor of the distribution. In the *Weibull* representation, applying such a transform to a *Weibull* distribution yields a straight line, which is parallel to the initial distribution. *Equation 5.46* can be rewritten in differential form as

$$dt_{op} = \frac{t_0}{\tau} dt_a \equiv q\, dt_a \tag{5.47}$$

such that the constant q can be interpreted as a time acceleration (deceleration) factor. Being q a constant, the acceleration factor is the same for all gate oxides belonging to the population under consideration. Thus, it does not depend on the failure-free time of the individual, as it would be the case for an arbitrary value of p. In fact, the general case yields

$$dt_{op} = \frac{p\, t_0}{\tau}\left(\frac{t_a}{\tau}\right)^{p-1} dt_a \tag{5.48}$$

In *Equation 5.48* the acceleration factor depends on the failure-free time measured under accelerated conditions. In other words, the acceleration factor is a function of the gravity of the defect causing the failure of each individual during the accelerated test. In addition to a time scaling, the general form of the time transform produces a change of the shape factor of the distribution. In the *Weibull* representation, the transformed distribution is again a straight line, but it is no more parallel to the representation of the initial distribution. Finally, it can be easily shown that, being η_a and β_a the scale and the shape factor of the distribution measured under accelerated conditions, the scale and the shape factor η_{op} and β_{op} of the transformed distribution are expressed as

$$\eta_{op} = t_0\left(\frac{\eta_a}{\tau}\right)^p \tag{5.49}$$

and

$$\beta_{op} = \frac{\beta_a}{p} \tag{5.50}$$

Thus, the survival probability $R_{op}(t)$ at operating conditions can be computed from *Equation 5.44* and be expressed in terms of the distribution parameters η_a and β_a

$$R_{op}(t) = e^{-\left(\left(\frac{\tau}{\eta_a}\right)^p \frac{1}{t_0} t\right)^{\frac{\beta_a}{p}}} \tag{5.51}$$

In a similar way, we obtain the failure rate at operating conditions $\lambda_{op}(t)$

$$\lambda_{op}(t) = \frac{\beta_a}{p} \frac{1}{t_0} \left(\frac{\tau}{\eta_a}\right)^p \left(\frac{1}{t_0}\left(\frac{\tau}{\eta_a}\right)^p t\right)^{\frac{\beta_a}{p}-1} \tag{5.52}$$

Both *Equations 5.51* and *5.52* have been derived under the assumption that the contribution due to the intrinsic failure can be neglected, *i.e.* the intrinsic distribution has be assumed to be constant and equal to one.

There is experimental evidence that under normal conditions the shape parameter of the extrinsic distribution is less than one, whereas that of the intrinsic distribution exceeds the unity. Thus, basing on *Equation 5.45*, we can conclude that the failure rate of the whole population decreases up to a turning time t_T (where it reaches a minimum) and after that it rapidly increases due to the occurrence of wear out failures. The most convenient way to determine the turning time t_{Top} at operating conditions is to start from t_{Ta} measured at accelerated conditions. Thus, after differentiation of *Equation 5.45* and by introducing the parameters η_{ea}, β_{ea}, η_{ia}, and β_{ia} as they have been measured by accelerated testing, we have

$$t_{Ta} = \left(-\frac{\beta_{ia}(\beta_{ia}-1)\,\eta_{ia}^{-\beta_{ia}}}{\beta_{ea}(\beta_{ea}-1)\,\eta_{ea}^{-\beta_{ea}}}\right)^{\frac{1}{\beta_{ea}-\beta_{ia}}} \tag{5.53}$$

Finally, the time transform of *Equation 5.46* delivers the requested result

$$t_{Top} = t_0 \left(\frac{t_{Ta}}{\tau}\right)^p \tag{5.54}$$

It should be noticed that, if *TDDB* were the dominant failure mechanism in *IGBT* devices, *Equation 5.54* would represent also the upper limit of the useful life of a device. In fact, operating a device beyond t_{Top} would result into a rapid decrease of the survival probability of the device. Numerical estimates of this parameter indicating that under normal stress conditions t_{Top} exceeds by several orders of magnitude the useful lifetime of a traction system, will be provided in the following, in conjunction with the investigation of the different acceleration models. From a

theoretical point of view, *Equation 5.54* has a more general significance than described above. In fact this equation applies to all failure mechanisms, which obey to a bimodal extreme values distribution. In this case, t_{Top} would represent the optimum time for replacing a device, in order to retain the lowest failure rate (preventive maintenance).

5.9 Optimization of screening procedures

The scope of a screening procedure is to eliminate the weakest tail of a distribution in order to realize a pre-defined failure rate during operation. This task is a typical optimization problem. In fact, one aims to eliminate the minimum amount of weak devices such that the reliability requirements are fulfilled. If we consider that this phase takes usually place when the *IGBT* chips are full-featured, we have that the uncontrolled elimination of an excessive amount of finished devices would reduce the process yield and thus it would turn into an economic loss. On the contrary, the occurrence of early failures due to insufficient screening would negatively impact the availability of the system.

In the following model, we assume again that early failures are just produced by the first quantiles of extrinsic tail of the bimodal distribution. This condition is fulfilled if the most severe process-related flaws have been eliminated in a preliminary phase and the intrinsic failures occur much later than the useful life of the system. In this case, the overall distribution of the failure-free times can be approximated by the extrinsic *Weibull* distribution, having a shape factor less than one (*i.e.* a failure rate decreasing with time until t_{Top} in *Equation 5.54*). The screening procedure is assumed to be a storage for a time t_S at an enhanced gate voltage V_S and at a temperature, which is representative for the operating temperature. Thus, the only acceleration factor considered here is that due to V_S, which exceeds the nominal operating voltage V_0 of the device. In summary, the problem to be solved is to find the minimum storage time t_S, such that the failure rate of the survivor devices operated at the nominal voltage V_0 is lower than a pre-defined value λ_0 over the whole useful life. Since the failure rate of the population under consideration always decreases with time, this problem has a solution.

As a first step, the time t_F is computed, which is needed by the unscreened population for reaching the pre-defined failure rate λ_0. By using *Equations 5.45*, *5.49* and *5.50*, we have

$$t_{Fop} = \eta_{op}\left(\lambda_0 \frac{\eta_{op}}{\beta_{op}}\right)^{\frac{1}{\beta_{op}-1}} = t_0\left(\frac{\eta_a}{\tau}\right)^p\left(\lambda_0 \frac{p}{\beta_a} t_0\left(\frac{\eta_a}{\tau}\right)^p\right)^{\frac{p}{\beta_a - p}} \tag{5.55}$$

The inverse time transform of *Equation 5.46*, yields the requested solution

$$t_S = \tau\left(\frac{t_{Fop}}{t_0}\right)^{\frac{1}{p}} \tag{5.56}$$

At the end of the useful life t_u (typically *30 years*), the logarithm of the survival probability $R(t)$ of a device, which survived the screening is expressed through

$$\ln R(t_u) = -\int_{t_{Fop}}^{t_u} \lambda(t)\, dt = -\int_{t_{Fop}}^{t_u} \frac{\beta_{op}}{\eta_{op}}\left(\frac{t}{\eta_{op}}\right)^{\beta_{op}-1} dt \tag{5.57}$$

Depending both on the oxide quality and on the pre-defined failure rate (typically *100 FIT*), the screening procedure can result into a dramatic reduction of the process yield. The percentage Y of the devices which will survive the screening as a function of the pre-defined failure rate λ_0 is given by

$$Y(\lambda_0) = e^{-\left(\frac{\eta_{op}}{\beta_{op}} t_{Fop}\right)^{\frac{\beta_{op}}{\beta_{op}-1}}} \tag{5.58}$$

and it is represented in *Figure 5.8* for different values of the scale parameter of the extrinsic distribution. It can be seen that for a η_e in the range of 10^{18} *s* (which is representative value for a reasonable extrinsic oxide operated at *3 MV/cm*) and for a failure rate of *100 FIT*, the yield of the screening is better than *0.95*.

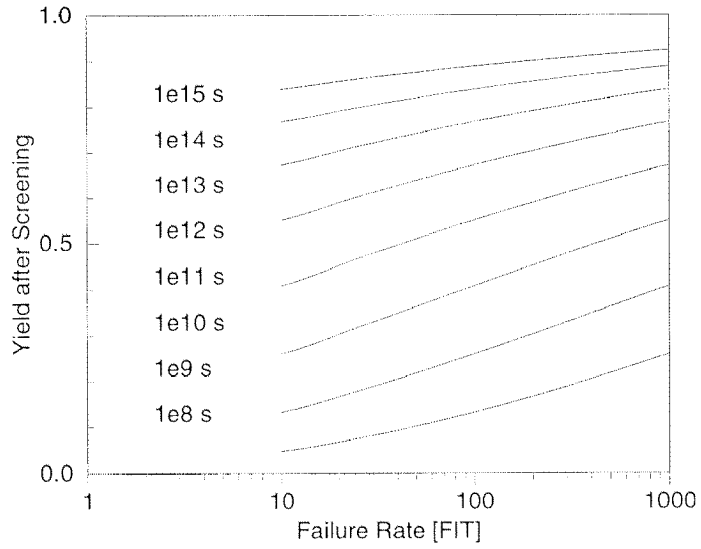


Figure 5.8 Yield of the screening process for realizing a pre-defined failure rate computed for $\beta_{op} = 0.15$ and η_{op} as parameter.

As practical example referred to *IGBT* devices, we consider the cumulated distribution of *Figure 5.1*. In *Figure 5.9*, we represented separately the three different sub-populations.

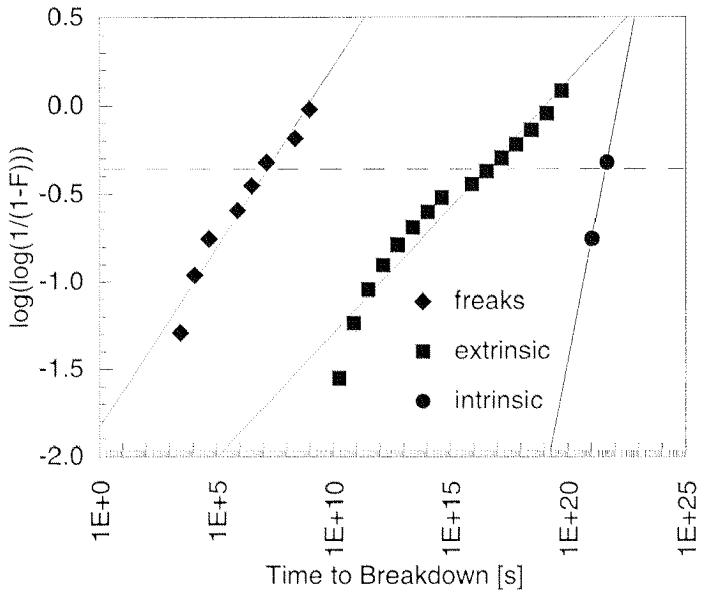


Figure 5.9 Weibull representation of the three sub-populations in Figure 5.1.

The procedure in *Section 5.6* delivers $\eta_f = 10^7$ *s*, $\beta_f = 0.2$, $\eta_e = 10^{17}$ *s*, $\beta_e = 0.14$, $\eta_i = 10^{21}$ *s*, and $\beta_i = 1.5$. It is value to notice that the scale and

shape parameters measured for this *50 nm* oxide are in very good agreement with the data provided in [104] for *10 nm* oxides. Thus, combining the results of the extrapolation in *Figure 5.9* with the curves of *Figure 5.8*, we can conclude that performing a screening for realizing a failure rate lower than *100 FIT* will produce the elimination of the whole freak sub-population and of about *10%* of the extrinsic sub-population. This would result into an overall screening yield of *0.8*.

5.10 Interpretation of the quasi-intrinsic model

If we assume that the whole population of gate oxides subjected to the voltage stress behave in the same way as an intrinsic oxide of nominal thickness t_i, we can use the degradation law of *Equation 5.5*, which yields

$$\frac{t_{BD_a}}{\tau_0} = e^{G\frac{t_i}{V_a}} \tag{5.59}$$

and

$$\frac{t_{BD_{op}}}{\tau_0} = e^{G\frac{t_i}{V_{op}}} \tag{5.60}$$

After building the ratio of previous equations, building the differentials, and rearranging the terms, we have

$$\frac{dt_{BD_{op}}}{dt_{BD_a}} = e^{Gt_i\left(\frac{1}{V_{op}}-\frac{1}{V_a}\right)} = q \tag{5.61}$$

Basing on *Equation 5.47* one can conclude that the acceleration factor is a constant. Since this equation corresponds to the case where $p = 1$, the transformed distribution is just shifted along the time axis and it has the same shape factor than the initial distribution.

5.11 Interpretation of the effective thickness model

In order to apply the effective thickness model for describing the degradation with time of extrinsic oxides, we can reasonably assume that the effective thickness of the samples which survive the pre-screening phase exceeds *70%* of the nominal gate oxide thickness. Thus, if an accelerated stress is performed at an electric field, which is *75-85%* of the critical electrical field defined in *Equation 5.6*, almost all samples are stressed in the *Fowler-Nordheim* regime, since they are reasonably far away from the condition for enhanced hole accumulation. Under this assumption, the differences in the aging behavior of the different samples refer to the differences in the effective thickness, only and the parameters τ_0 and G in *Equation 5.9* can be considered as constant. In the case of the accelerated stress, *Equation 5.9* can be rewritten as

$$t_{BD_a} = \tau_0 \, e^{G\frac{t_{eff}}{V_a}} \tag{5.62}$$

Since all parameters in *Equation 5.62* are known excepted t_{eff}, the equation can be solved over the effective thickness t_{eff}

$$t_{eff} = \frac{V_a}{G} \ln\left(\frac{t_{BD_a}}{\tau_0}\right) \tag{5.63}$$

By inserting *Equation 5.63* into *Equation 5.62*, expressed for the operating conditions, we have

$$t_{BD_{op}} = \tau_0 \left(\frac{t_{BD_a}}{\tau_0}\right)^{\frac{V_a}{V_{op}}} \tag{5.64}$$

The acceleration factor can be determined by building the differential

$$dt_{BD_{op}} = \frac{V_a}{V_{op}} \left(\frac{t_{BD_a}}{\tau_0}\right)^{\frac{V_a}{V_{op}}-1} dt_{BD_a} \tag{5.65}$$

Finally, by comparing *Equation 5.64* with *Equation 5.48*, we can conclude that

$$p = \frac{V_a}{V_{op}} \tag{5.66}$$

and

$$t_0 = \tau = \tau_0 \tag{5.67}$$

Thus, the time transform associated with the effective thickness model transforms the *Weibull* distribution measured under accelerated conditions into a *Weibull* distribution with scale and shape factors, which depends on the relative applied stress, that is

$$\eta_{op} = \eta_a^{\frac{V_{op}}{V_a}} \tau_0^{1-\frac{V_{op}}{V_a}} \tag{5.68}$$

$$\beta_{op} = \frac{V_{op}}{V_a}\beta_a \tag{5.69}$$

In *Figure 5.10* we represent the experimental scale factor of the extrinsic distribution as a function of the oxide field, as they have been measured in [104] and the model of *Equation 5.68* with $\tau_0 = 10\ ps$ and $E_{ox} = V_{ox}/t_{ox}$. All values of the model have been extrapolated from a single measurement at $E_{ox} = 10.4\ MV/cm$.

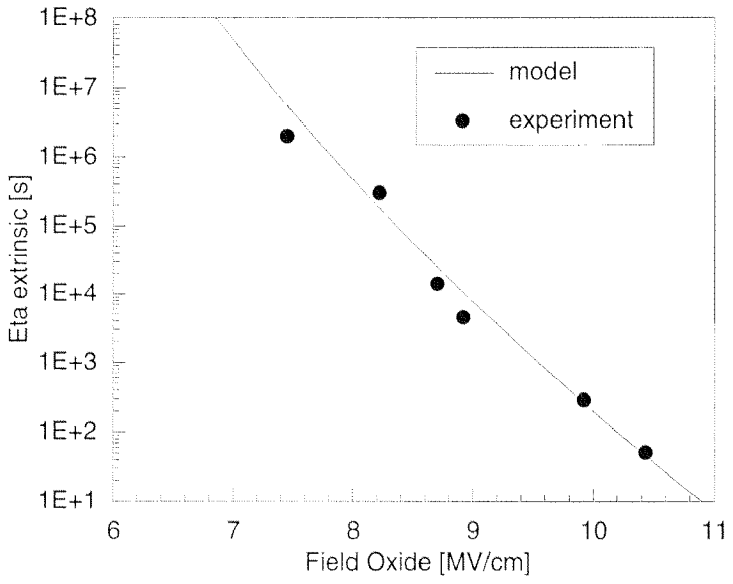


Figure 5.10 Scale factors of the extrinsic distributions. The experimental values are taken from [104], while the model is that of Equation 5.68 with $\tau_0 = 10$ ps.

Figure 5.10 shows that the proposed model predicts the experimental scale factors within a factor of two over at least five decades.

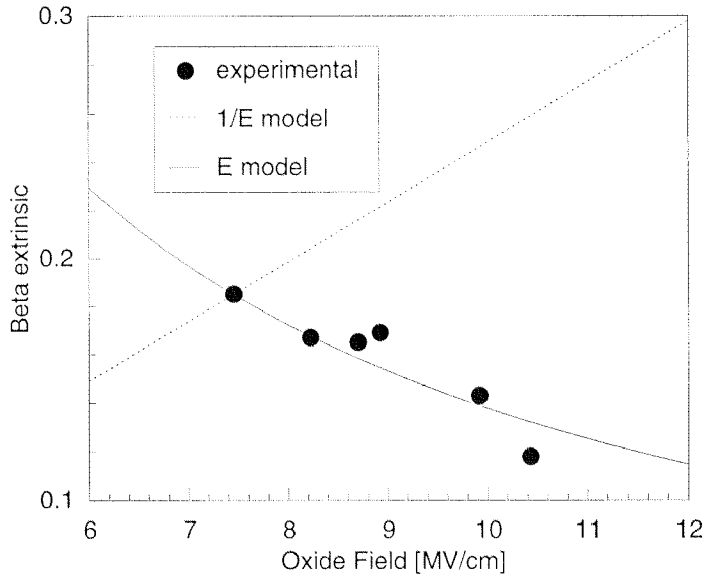


Figure 5.11 Shape factors of the extrinsic distributions. The 1/E-model of Equation 5.5 does not correctly fit the experimental values from [104], which are accurately predicted by the E-model

Figure 5.11 represents the experimental shape factor of the extrinsic distributions as they have been measured in [104]. The model of *Equation 5.69* predicts that the shape factor increases with increasing field oxide, that is the opposite behavior as it is observed experimentally.

On the contrary, if we assume a dependence of t_{BD} according to the *E*-model of *Equation 5.13*, we have

$$t_{BD_a} = \tau_o' \, e^{-\gamma E_a} \tag{5.70}$$

where τ_0' and γ are constants, which depend on the oxide thickness. By keeping constant the oxide thickness and by repeating the procedure in *Equation 5.62* through *5.69*, we obtain following time dependence

$$t_{BD_{op}} = \tau_0' \left(\frac{t_{BD_a}}{\tau_0'} \right)^{\frac{E_{op}}{E_a}} \tag{5.71}$$

which yields

$$\beta_{op} = \frac{E_a}{E_{op}} \beta_a \tag{5.72}$$

The field dependence of *Equation 5.72* predicts accurately the experimental data of [104], as it is shown in *Figure 5.11.*

Furthermore, *Equation 5.71* yields for the scale factor

$$\eta_{op} = \tau_0' \left(\frac{\eta_a}{\tau_0'} \right)^{\frac{E_{op}}{E_a}} \tag{5.73}$$

Since the time to breakdown for an arbitrary field oxide is always smaller than τ_0' (*see Equation 5.70*), and being η_a the time to breakdown of the *0.63*-quantile of the distribution, the base of the power in the right side of *Equation 5.73* is always smaller than one. This results into a scale factor decreasing with increasing field oxide, as it is expected from the experimental data in *Figure 5.10.* A least-square fit of the experimental data in *Figure 5.10* delivers $\tau_0' = 1.2\ 10^{18}\ s$ (and $\gamma = 3.6\ cm/MV$).

5.12 Comparison with the IMEC model

Recently, an empirical model has been proposed [104] for predicting the reliability of thin oxides in the *10 nm* range. Basing on experimental observations, the model assumes that the thin oxides within a given population fail due to the competition of both intrinsic and extrinsic breakdown. In spite of this assumption, the experimental data are processed in [104] according to the formalism for the mixture of distributions (*see Section 5.4*). This leads to some contradictory results, which nevertheless confirm the validity of the competing risk model. The model assumes a different wearout behavior for intrinsic and extrinsic breakdown. In fact, the distribution of the intrinsic population is assumed to depend on the applied electric field over a constant acceleration factor, which just affects the scale factor. In other words, we have

$$\beta_{iop} = \beta_{ia} = 8.76 \tag{5.74}$$

$$\eta_{iop} = \left(\frac{E_a}{E_{op}}\right)^2 e^{G\left(\frac{1}{E_{op}} - \frac{1}{E_a}\right)} \eta_{ia} \tag{5.75}$$

On the contrary, scale and shape parameters of the extrinsic distribution are described by the following empirical dependencies

$$\beta_{eop} = \frac{E_a}{E_{op}} \beta_{ea} \tag{5.76}$$

$$\eta_{eop} = \left(\frac{E_a}{E_{op}}\right)^2 e^{B\left(\frac{1}{E_{op}} - \frac{1}{E_a}\right)} e^{-\gamma_e\left(E_{op} - E_a\right)} \eta_{ea} \tag{5.77}$$

It is interesting to notice that the field dependence of the shape factor is exactly the same, as one would expect in the case of the *E*-model. In opposite, the complex field dependence of the scale factor has been approximated here by combining the *1/E* and the *E*-model through the fitting parameter γ_e, which is defined as

$$\gamma_e = \frac{\ln\left(C_0 A_{cap}\right)}{C_1} \tag{5.78}$$

where A_{cap} is the gate oxide area, while $C_0 = 86\ cm^{-2}$ and $C_1 = 1.436\ MV/cm$ are fitting parameters [104].

The scale and shape parameters of the cumulated distribution arising from a stress at an arbitrary oxide field can be extrapolated from the experimental fact that for *MOS* capacitors with $t_{ox} = 10\ nm$ and $A_{cap} = 0.1\ cm^2$at $10\ MV/cm$ $\eta_i = 599\ s$, $\eta_e = 206\ s$, $\beta_i = 8.76$, and $\beta_e = 0.144$.

Being the oxide area of an *IGBT* is in the $0.1\ cm^2$ range and by assuming the same degradation behavior for oxides thicker than *10 nm*, we can estimate all relevant reliability parameters by using the probabilistic model developed in *Section 5.4*. As an example, we represented in *Figure 5.12* the failure rate due to the extrinsic tail of the distribution for different oxide field strengths. It can be observed that over the whole useful life of a traction system (typically *30 years*) the failure rate is decreasing monotonically. Furthermore, in the case of oxide fields lower than *3 MV/cm*, it is far beyond *1 FIT* even at the very beginning of the operation.

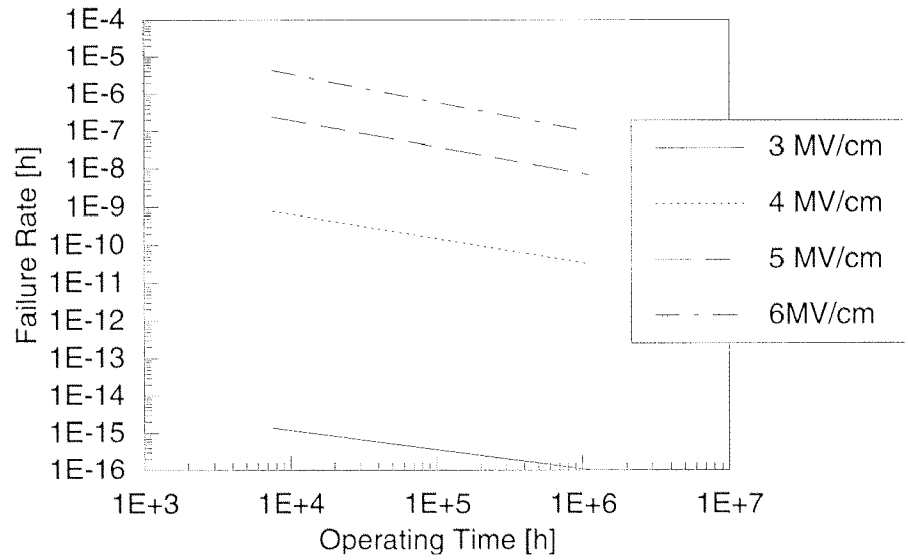


Figure 5.12 Failure rate of the extrinsic population as a function of the time and of the electric field according to the field dependence introduced by [104]

The failure rate due to the extrinsic and to the intrinsic components has been represented in *Figure 5.13* for high values of the oxide field. As it is predicted by *Equation 5.45*, the failure rate decreases until the wearout mechanisms get dominant. For oxide fields below *6.6 MV/cm*, the turning point occurs after *11000 years* operation, *i.e.* far beyond the useful operating lifetime of a traction system.

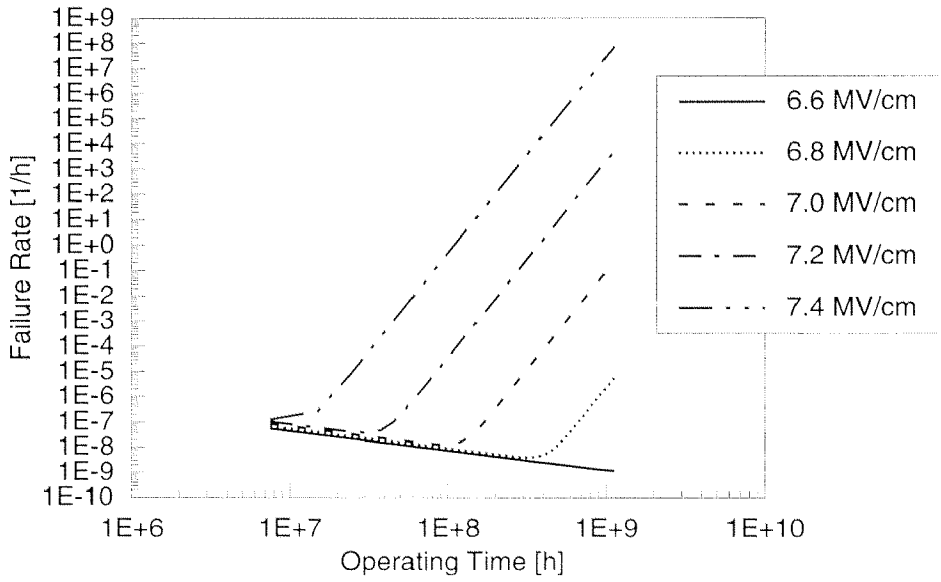


Figure 4.13 Failure rate of the intrinsic population as a function of the time and of the electric field according to the field dependence introduced by [104]

5.13 Final remarks and summary

After reviewing the most popular models for time dependent breakdown and summarizing the main empirical relations, we answered in this Chapter to five questions of technical relevance:

- How to condition samples for performing lifetime measurements
- How to perform accelerated tests under voltage acceleration
- How to describe and parameterize experimental data
- How to predict the lifetime of devices under normal operating conditions
- How to design a screening for realizing a pre-defined failure rate while minimizing the yield loss

The discussion of these subjects has been focused onto the case of extrinsic oxides thicker than *20 nm*. Since the formalism we developed is based on a statistical description of the failure mechanism, it enables to derive the main reliability parameters in analytical form, even for the very first quantiles of the cumulative distribution. This property is of great relevance, especially when developing lifetime models for traction systems. In fact, the traditional deterministic models just predict the average value of the lifetime. The concepts of mixture of failure mechanisms and of competing risks have been formally defined, and the related probabilistic models have been derived. The probabilistic model for competing risks has been applied to the case of a cumulated distribution consisting of an intrinsic and of an extrinsic part. The formalism used for modeling these components of the distribution is the *Weibull* statistics. The scaling rules, which apply either to the case of multiple paralleled devices or to the case of devices exhibiting a different oxide area, have been strictly derived from the properties of *Weibull* distributions. Finally, for the first time the model for reliability prediction has been derived basing on the principle of the invariance of the statistics. Following this criterion, we have defined the family of time transformations, which, once applied, transform a *Weibull* distribution into a *Weibull* distribution. The resulting model has been validated with experimental data. We also demonstrated that the time transformations associated with the *1/E*-model and with the *E*-model are both invariant transformations for *Weibull* distributions. However, the only model, which correctly predicts the experimentally observed dependence of the distribution parameters on the field oxide is the *E*-model.

Chapter 6

Lifetime Modeling of the Bond Wire Lift Off in IGBT Modules

6.1 Introduction

Reliability assessment with the use of full-featured *IGBT* modules is quite expensive and in several cases even unfeasible. For this reason, testing is usually carried out basing on dedicated test structures trimmed on a given failure mechanism. The use of suitable models enables the extrapolation of these data to real devices, under consideration of their complexity. In this Chapter, we start from the results of accelerated tests performed at discrete temperatures swings on single-emitter bondwire devices and we develop a consistent model for extrapolating t_f and the *mean time-to-failure* of devices with a single emitter bond wire under an arbitrary application profile. Additionally, we propose a model which takes into account the complexity of real devices, in particular the parallel structure of the bond wires. Finally, these models will be demonstrated in the case

of a realistic application profile of a locomotive operated on the Swiss railway network, as it has been proposed in the work of *Zehringer* and *Stuck* [1].

6.2 Characterization of the failure mechanism

Since the early *'80s*, bond wire lift-off (*Figure 6.1*) has been recognized as a major failure mechanism affecting power devices submitted to thermal cycles. The driving force of the mechanism is the mismatch in the thermal expansion coefficients of aluminum (*24 ppm/K*) and silicon (*3 ppm/K*). During operation in locomotives, the junction temperature may increase by *70 K*. The maximum junction temperature T_{jmax} is in the *370 K* range [1], and it is reached at a typical temperature change rate of *50-100 K/s*. This produces a differential elongation of the bond wire in respect to the substrate of about *0.2%*, and it results into a plastic flow of the wire material, especially at the periphery of the bonding interface where the shear stress reaches its maximum strength.

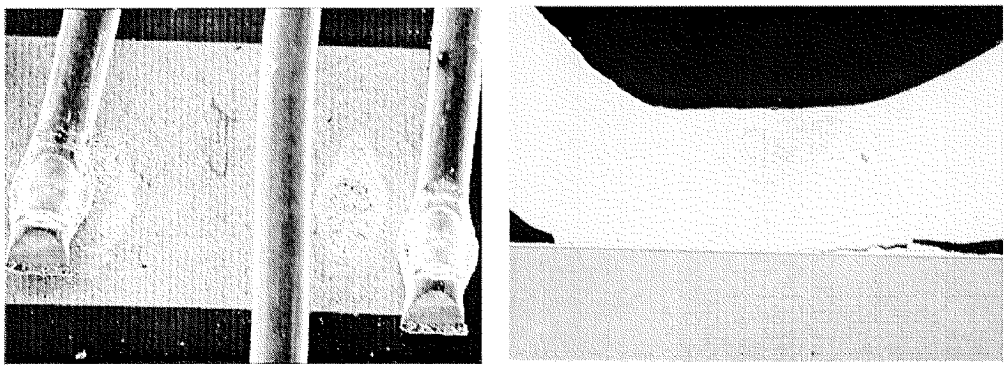


Figure 6. 1 Emitter bond wires after lift-off (left, 15x). Cross-section of an emitter bond wire after thermal cycles (right, 110x). The crack propagates within the aluminum bond wire, starting from the tail.

Cyclic application of such thermomechanical stresses results in low-cycle fatigue followed by progressive degradation of the contact resistance of the bond. Due to the strong deformation experienced by the bond during the welding process, the aluminum grain size in the outer bond region is smaller than in the center, where the bond strength is lower. Crack initiates at the periphery of the bonding interface and propagates along

the transition region between coarse and fine-sized aluminum grains, until the weaker central bond area is reached and the bond wire finally lifts-off.

In devices with multiple bond wires this failure mechanism affects preferably those wires which are located close to the center of the chip, where the junction temperature reaches the maximum. Thus, central emitter bond wires normally fail at first, then they are followed by the survivor bond wires which have to carry the full load. When the current density within the survivor bond wires reaches a critical value, they melt producing an open circuit. This failure mode represents the usual end-of-life behavior of devices submitted to accelerated tests like *power cycling*. In fact, during *power cycling* devices are operated quasi-statically and at on-voltages which normally do not exceed *10 V*. On the contrary, failure analysis of field failures [17] has shown that the terminating mechanism of devices operated at high voltages may be represented by the triggering of parasitics due to the current crowding resulting from the bond wire lift-off.

Bond wire lift-off affects both emitter and gate bond wire. This can lead to the conclusion that the level of the current flowing through the bond wire does not play a dominant role in activating the failure mechanism. In fact, measurements performed by infrared thermography have shown that at least up to a current density of about *8 kA/cm*^{2}, the temperature of *300 μm* aluminum bond wires under normal operating conditions does not exceed the chip temperature. Of course, this is not true when almost all emitter bond wires lifted-off. In this case, the current density through the survivors may increase by a *factor of 3* or more, resulting into a positive feed back which highly accelerates the failure mechanism. However, it must be noticed that such a level of degradation is rarely encountered in devices operated in the field.

In the case of large temperature swings, for which T_{jmax} exceeds *380 K*, bond wire lift-off is usually observed in conjunction with the reconstruction of the aluminum metallization of the device [17]. Both bond wire lift-off and aluminum reconstruction are thermomechanical phenomena, which are characterized by relaxation time constants in the order of some few seconds. This does not apply to solder cracking, which is an additional relevant failure mechanism of *IGBT* modules [26]. In fact, the relaxation of thermomechanical stresses arising in solder joints takes a time in the order of a minute. This difference in the relaxation time constants is normally used for activating selectively these failure mechanisms during accelerated tests [107].

From the electrical point of view, the bond wire degradation is followed by an increase of the collector-to-emitter voltage (V_{CE}) which is measured at low collector current (typically *100 mA*) [7]. When V_{CE} is represented as a function of the number of thermal cycles, it shows a continuous increase (ΔV_{CE}) until the bond wire is completely lifted-off (*see Figure 2.6*). Reaching a pre-defined threshold of ΔV_{CE} is commonly used as failure criterium (typically *5-20%* of the initial value). Besides this electrical indicator, they are also mechanical failure criteria defined on the base of a *pull test*. In this case, a population of bond wires submitted to thermal cycles is said to fail, if the pull strength of given amount of individuals (typically *20%*) decreases below a pre-defined pull force limit. Electrical and mechanical criteria are not equivalent and they are not easy to be correlated. ΔV_{CE} is normally used for accelerated test as *power cycling*, where a pulsed collector current is forced into the full-featured device (permanently set in the on-state) in order to realize temperature cycles of constant amplitude at a constant mean temperature [7]. On the contrary, the mechanical failure criterium is applied in the case of *thermal cycling* [107], where large samples of single bond wires (mounted as they were in a real device) are submitted to passive thermal cycles of constant amplitude and at a constant mean temperature.

6.3 Accelerated Testing

The investigated test structures are *IGBTs* with a single emitter bond wire. The aluminum bond wire is *350 μm* in diameter and it is bonded on aluminum bond pads directly deposed on silicon (without strain buffer). Since creep has been shown to be negligible during fatigue testing of bond wires [107], the devices are submitted to power cycles with a cycle period of *3 seconds*. The working point of the device is set such that T_{jmax} never exceeds *373 K*, in order to selectively activate the bond wire lift-off mechanism [7,107]. The degradation of the contact resistance of the bond wire is characterized by on-line monitoring of ΔV_{CE}, during the switch off transient. The failure criterium, which has been assumed here, is a *10%* drift of V_{CE} with respect to the initial value. The details of the measurement set-up are described in [108]. The results of the power cycle experiment are represented in the *log-log plot* of *Figure 6.2*.

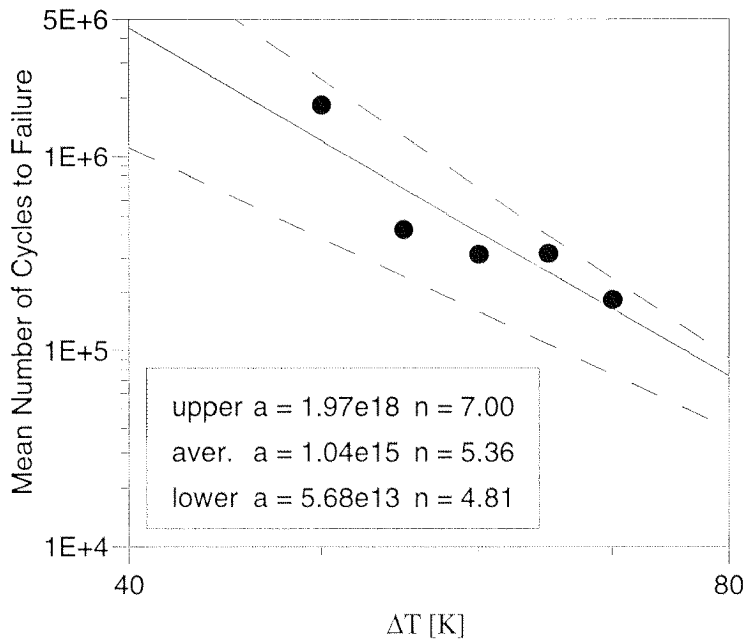


Figure 6.2 Log-log representation of the power cycling data. The Coffin-Manson coefficients for the maximum, the lowest and the average values are reported in the inset. Failure criterium $\Delta V_{CE} = 10\%$

6.4 Modeling the number of cycles to the failure

By using the simple bimetallic model as an approximation for the thermomechanical stresses arising at the interface of a joint between aluminum and silicon when submitted to a temperature swing ΔT, we obtain a total strain

$$\varepsilon_{tot} = L\left(\alpha_{Al} - \alpha_{Si}\right)\Delta T \tag{6.1}$$

where α_{Al} and α_{Si} are the thermal expansion coefficient of aluminum and silicon, respectively, and L the typical length of the joint. Due to the large thermomechanical mismatch, the joint is operated in deep plastic regime. Thus, it can be reasonably assumed that the full strain is mainly given by the plastic strain

$$\varepsilon_{tot} = \varepsilon_{elastic} + \varepsilon_{plastic} \approx \varepsilon_{plastic} \tag{6.2}$$

Having the plastic strain, the mean number of cycles-to-failure N_f can be computed by the *Coffin-Manson* law

$$N_f \propto \varepsilon_{plastic}^{-n} \tag{6.3}$$

where the exponent n is a positive number. After insertion of *Equation 6.1* into *Equation 6.3*, we have

$$N_f = a\,(\Delta T)^{-n} \tag{6.4}$$

where a is a proportionality constant. In our case, the parameters a and n are extracted by regression of the average number of cycles to failure in the *log-log plot* of *Figure 6.2*. Due to the large dispersion of the experimental data, we consider also the pessimistic and the optimistic regression parameters, which are represented by the dashed lines in *Figure 6.2*.

6.5 Lifetime Modeling

As outlined in *Section 6.1*, we propose here a model for estimating both t_f and the mean time-to-failure of devices with a single emitter bond wire which are submitted to an arbitrary application profile. In the following, we estimate the contribution due to the topologic complexity of a real module by investigating the related reliability block diagram. In the model, we assume a degradation of the bond wires which depends linearly on the number of thermal cycles. Under this assumption, a given number of thermal cycles can be transformed into time just by multiplication with the cycle period. The derived equations are valid if t_f and *MTTF* do not depend on the cycling frequency. This assumption is correct when the thermal cycling period is much longer than the typical relaxation time of the thermomechanical stresses within the aluminum-silicon joint.

6.6 Modeling the mean time-to-failure (MTTF)

Each dot in *Figure 6.2* is the average of the different number of cycles-to-failure of devices cycled at the same ΔT. If the time-to-failure of these devices are distributed according to a *Weibull* distribution with scale parameter λ and shape parameter β, the *MTTF* is defined by

$$MTTF(\Delta T) = \frac{\Gamma(1+1/\beta)}{\lambda} \tag{6.5}$$

where Γ is the complete gamma function [106]. Unfortunately, *power cycling* does not provide sufficient experimental data for extrapolating the parameters α and β from a *Weibull* plot according to the traditional procedure. Thus, every $N_f(\Delta T)$ has to be approximated by the point estimate, as it has been made in *Figure 6.2*.

For N thermal cycles performed at a single ΔT during a given operating time (*e.g. 1 year*), the cumulated fatigue function is defined as

$$Q(N) \equiv \frac{f(N,\Delta T)}{N_f} \tag{6.6}$$

where f is a generic function, $Q(N)$ is monotonic and it is *equal to 1* when the bond wire failure occurs. The assumption of a linear fatigue damage accumulation with the number of thermal cycles yields

$$Q(N) = \frac{N}{N_f} \tag{6.7}$$

As it has been shown elsewhere [1], the junction temperature of an *IGBT* in railway traction applications is a strong function of the application profile of the device, depending for instance on the line topography, on the load, on the schedule, on the locomotive control (automatic, driver), etc.

The frequency distribution of the thermal excursions encountered by the device during a stated period of time (*e.g. 1 year*) is inferred from field measurements [1], and it is usually expressed for a given application profile as a histogram, whose envelope is $g(\Delta T)$. The incremental fatigue damage produced by the thermal cycles within the interval ΔT and $\Delta T+d(\Delta T)$ of $g(\Delta T)$ is computed by

$$dQ = \frac{g(\Delta T)d(\Delta T)}{N_f(\Delta T)} \tag{6.8}$$

Integration of *Equation 6.8* and insertion of *Equation 6.4*, yields the fatigue cumulated by a single bond wire during one-year operation

$$Q(1\,y\,operation) = \int_{\Delta T_{min}}^{\Delta T_{max}} \frac{g(\Delta T)}{N_f(\Delta T)} d(\Delta T)$$
$$= \frac{1}{a} \int_{\Delta T_{min}}^{\Delta T_{max}} \frac{g(\Delta T)}{\Delta T^{-n}} d(\Delta T) \qquad (6.9)$$

Finally, by assuming that the application profile is constant over the time, the mean time-to-failure of a single bond wire expressed in years is simply given by

$$MTTF_{1\,bond\,wire} = \frac{1}{Q(1\,y)} \qquad (6.10)$$

6.7 Modeling the time to the failure of the f-quantile (t_f)

Like in *Section 6.6*, we assume that the number of cycles-to-failure of a bond wire population which is cycled at a given temperature swing ΔT_0 is distributed according to a *Weibull* distribution. If λ and β are experimentally known, the number of cycles N_f required for the f-quantile of the population will fail is given by

$$N_f(\Delta T_0) = \frac{1}{\lambda}\left[\ln\left(\frac{1}{1-f}\right)\right]^{1/\beta} \qquad (6.11)$$

Because λ and β need to be measured with high accuracy, *thermal cycling* of single bond wires shall be preferred here rather than *power cycling* of full-featured devices. If the shape factor of the distribution does not depend on ΔT (performing the experiment at a different ΔT just shifts the distribution along the cycle-axis) and if the dependence of N_f on ΔT is the same than in *Equation 6.4*, we have for an arbitrary ΔT

$$N_f(\Delta T) = N_f(\Delta T_0)\left(\frac{\Delta T}{\Delta T_0}\right)^{-n} \qquad (6.12)$$

By proceeding like in *Section 6.6*, we have

$$Q(1\,y\,operation) = \frac{1}{N_f(\Delta T_0)\Delta T_0^{n}} \int_{\Delta T_{min}}^{\Delta T_{max}} \frac{g(\Delta T)}{\Delta T^{-n}} d(\Delta T) \qquad (6.13)$$

By assuming again that the application profile is constant over the time, the t_f of a single bond wire expressed in years is simply given by

$$t_{f,\,1\,bond\,wire} = \frac{1}{Q(1\,y)} \tag{6.14}$$

6.8 Modeling of the Complexity Factor

The influence of the complexity on the lifetime of a module is modeled here by introducing a time-dependent complexity factor $M(t)$, such that

$$MTTF_{module} = M(t)\, MTTF_{1\,bond\,wire}$$

and

$$t_{f,\,module} = M(t)\, t_{f,\,1\,bond\,wire} \tag{6.15}$$

The complexity factor takes into account the gain (or the loss) in lifetime, due to bond wires that are connected either in parallel or in series and it is computed from the reliability block diagram of *Figure 6.3*. This reliability block diagram refers to a device, which consists of *C chips*, each having N_e *emitter bond wires* and a single gate bond wire. All N_e emitter bond wires are in warm redundancy (parallel). On the contrary, since the failure of a single gate bond wire causes the failure of a chip (and then of a whole module), each gate bond wire appears as a series element. For computing $M(t)$, we distinguish three cases of practical interest.

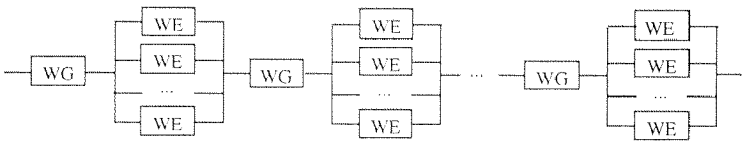


Figure 6.3 Reliability block diagram of a module with C chips (series connection). In each chip there are N_e bond wires in warm redundancy (parallel connection) and one series gate bond wire

Case 1

The time to failure of the gate bond wire is assumed to be much longer than the time to failure of an emitter bond wire. In this case, the bond reliability function of a gate bond wire can be assumed as a constant (*value 1*) during the whole life of an emitter bond wire. This assumption leads to the reduced reliability block diagram of *Figure 6.4*.

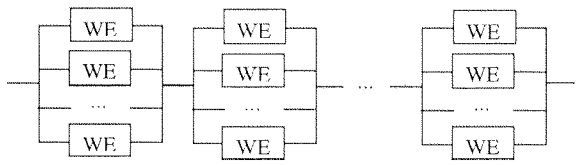


Figure 6.4 Reduced reliability diagram of a module when the reliability function of a gate bond wire can be assumed to be one

Let $R(t)$ be the reliability function of a single bond wire, and k the minimum allowed number of non-failed bond wires for each chip (k out of N_e redundancy), we have for the reliability function of a single chip

$$R_C(t) = \sum_{i=k}^{N_e} \binom{N_e}{i} R^i(t)(1-R(t))^{N_e-i} \tag{6.16}$$

Further, we impose that no chip is allowed to fail. This yields to the reliability function of the full module

$$R_M(t) = \left(R_C(t)\right)^C \tag{6.17}$$

The complexity factor $M(t)$ is defined as the ratio of the survival probability (reliability function) of the whole module with the survival probability of a single bond wire

$$M(t) = \frac{R_M(t)}{R(t)} \tag{6.18}$$

$M(t)$ in the case of a *12* (k) out of *16* (N_e) redundancy of a module consisting of *8 chips* is represented in *Figure 6.6*. In this particular case, the gain in lifetime due to redundancy does not exceed *10%* of the single bond wire lifetime. Moreover, the complexity factor $M(t)$ of a module with C chips and each N_e emitter bond wires exceeds the unity, only when

the survival probability of a single bond wire is above *0.9*. On the contrary, when the failure probability is over *0.1*, the complexity factor is less than one, and it rapidly decreases down to zero when the survival probability of a single bond wire approaches *0.25*. A similar behavior can be observed in *Figure 6.6* for a single chip. Summarizing, under the assumption of *Case 1*, the complexity factor due to the redundancy of the emitter bond wires is almost of the order of the unity.

Case 2

The mean time to failure of the gate bond wire is assumed to be much shorter than the mean time to failure of the paralleled emitter bond wires. In this case, the bond reliability function of the paralleled bond wires can be assumed as a constant (*value 1*) during the whole life of gate bond wire. This assumption leads to the reduced reliability block diagram of *Figure 6.5*.

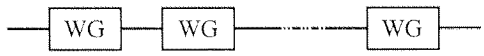


Figure 6.5 Reduced reliability diagram of a module when the reliability function of the paralleled emitter bond wires can be assumed to be one.

In this case the reliability function of a module is defined by

$$R_M = \prod_{i=1}^{C} R(t) = R^C(t) \tag{6.19}$$

The complexity factor $M(t)$, as it has been defined in *Equation 6.18* becomes

$$M(t) = \frac{R(t)^C}{R(t)} = R(t)^{C-1} \tag{6.20}$$

$M(t)$ is plotted in *Figure 6.7*, by assuming again a module with eight chips. From *Figure 6.7* it can be seen that the $M(t)$ is always less than one, and, when the failure probability of a single gate bond wire is *0.25*, the survival probability of the whole module is less than *0.15*.

Furthermore, *Equation 6.20* clearly shows that the reliability function of a module is a strong function of the total number of gate bond wires.

Case 3

The mean time to failure of a single gate bond wire is assumed to be the same as the mean time-to-failure of a single emitter bond wire. This assumption leads to the reliability block diagram of *Figure 6.3*. The related reliability function is

$$R_M = R(t)^C \left(\sum_{i=k}^{N_e} \binom{N_e}{i} R^i(t)\left(1-R(t)\right)^{N_e-i} \right)^C \tag{6.21}$$

while the complexity factor is described by

$$M(t) = R(t)^{C-1} \left(\sum_{i=k}^{N_e} \binom{N_e}{i} R^i(t)\left(1-R(t)\right)^{N_e-i} \right)^C \tag{6.22}$$

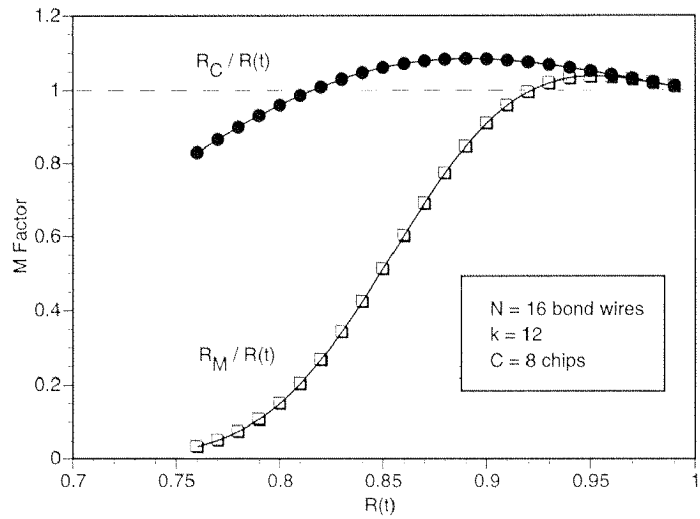


Figure 6.6 Complexity factor of a module (squares) and complexity factor of a single chip (dots) as a function of the survival probability of a single emitter bond wire (*Case 1*)

and represented in *Figure 6.7*. As expected, the $M(t)$ factor is always less than one, since the series connection of C gate bond wires is the dominating factor.

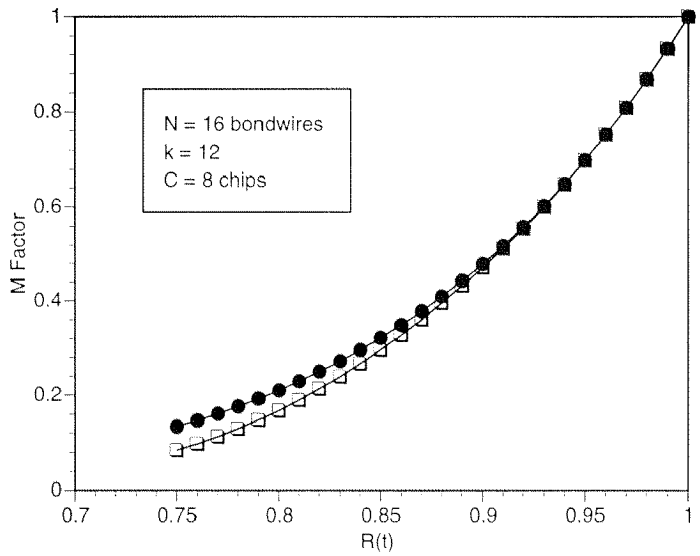


Figure 6.7 Complexity factor of a module as a function of the survival probability of a single emitter bond wire (dots: case 2, squares: case 3)

Remarks

In spite of the fact that usually gate bond wires are located at the chip center where the chip reaches its maximum temperature, there is experimental evidence that they are less prone to bond wire lift-off than emitter bond wires. This effect is commonly explained by the fact that gate bond wires are not subjected to ohmic self-heating as emitter bond wires are [17]. Although no quantitative data are available about the acceleration of bond wire lift-off through ohmic self-heating, basing on heuristic considerations, the pessimistic assumption can be made that the *MTTF* of gate bond wires is the same than the *MTTF* of emitter bond wires. Thus, the *M(t)* of real *IGBT* modules can be computed according to the approximation described in *Equation 6.22*. By using *Equation 6.17* the lifetime would be overestimated by about *30%*.

6.9 Application to Gaussian distributions

In this section, we briefly quantify the lifetime of an *IGBT* module, basing both on the experimental data of *Figure 6.2*, and on the model of *Equation 6.10*.

Table 6.1 Lifetime extrapolation for three different ΔT frequency distributions and related parameters

I [kCy]	I_0 [K]	σ [K]	ΔT_{min} [K]	ΔT_{max} [K]	N_{tot} [kCy]	M	TTF [y]	
							Min	714
20	20	10	0	70	350	0.9	Max	34
							Aver	100
							Min	91
20	30	10	0	70	350	0.9	Max	7
							Aver	18
							Min	17
20	40	10	0	70	350	0.9	Max	2
							Aver	5

For sake of simplicity we assume that the frequency distribution of the ΔT during one-year device operation ($g(\Delta T)$) is represented by a normal distribution characterized by the intensity I, the mean value I_0 and the variance σ. Since in the case of a normal distribution *Equation 6.9* turns into an elliptical integral, the function Q is evaluated by numerical integration. Three examples of such an extrapolation are shown in *Figure 6.8*. The full parameter set assumed for the extrapolation is summarized in *Table 6.1*.

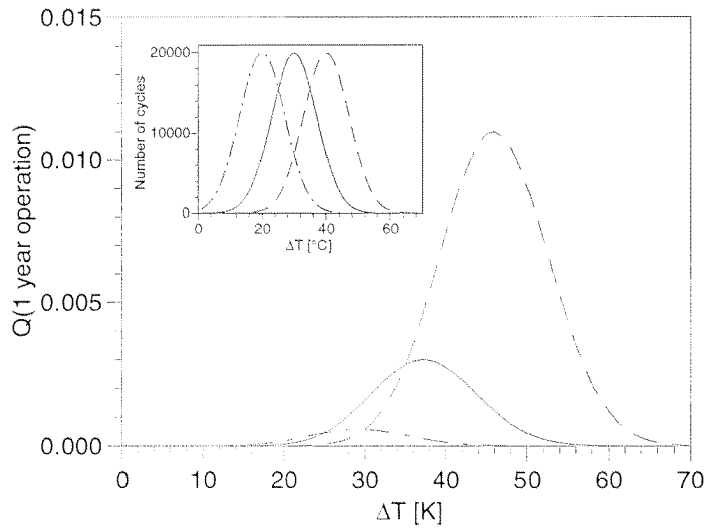


Figure 6.8 Cumulated damage during 1 year operation computed for three ΔT frequency distributions (plotted in the insert) and with the parameters from the average curve of Figure 6.2

From *Table 6.1* it can be observed that the extrapolated times to failure strongly depend on the *Coffin-Manson* parameters of *Equation 6.4*. In fact, under the same thermal conditions the extrapolated lifetime can differ up to a *factor of 20*. This difference decreases by increasing the average ΔT.

Moreover, by increasing the average ΔT from $20°C$ up to $40°C$, the extrapolated lifetime decreases approximately according to an exponential law.

6.10 Application to a realistic profile

Zehringer and Stuck [1] have investigated the specific stress conditions of *IGBT* for the Swiss Intercity railway services. Measurements have been performed with *locomotives Re460* (*3 level* voltage source inverters, DC link voltage *3.5 kV*, *6.4 MW*) and *Re465* (*2 level* voltage source inverters, DC link voltage *2.8 kV*, *7 MW*) equipped with *GTO* converters and operated on major intercity lines. The *GTO* data has been converted into the junction temperature profile of equivalent *IGBT* devices by the means of electrical and thermal models. The distribution of $g(\Delta T)$ has been extracted from the temperature profile by frequency analysis and extrapolated to *30 years* assumed operating lifetime.

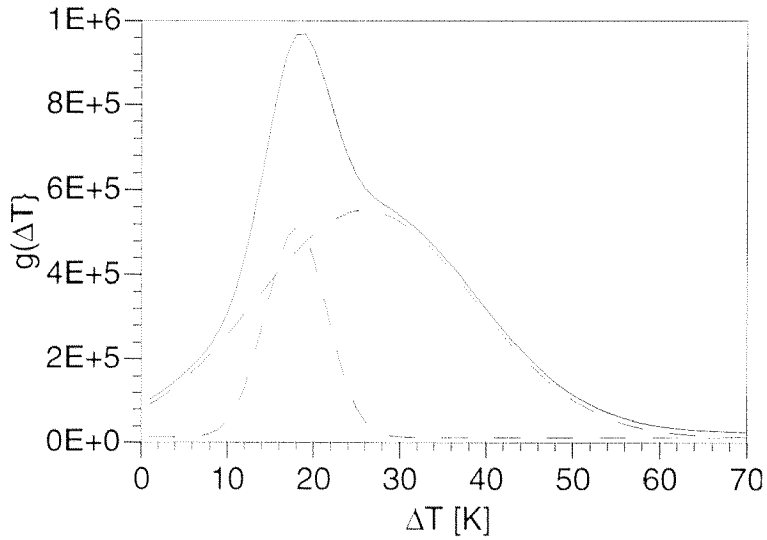


Figure 6.9 Frequency distribution of the temperature swing of an IGBT over 30 years operating lifetime of a locomotive according to *Zehringer* and *Stuck* [1]. The envelope is the sum of two Gaussian distributions (dashed line).

This study has demonstrated that $g(\Delta T)$ can be realistically described by the superposition of two normal distributions, as it is shown in *Figure 6.9*.

Figure 6.10 represents the low cycle fatigue cumulated by a bond wire during *1-year* operation as a function of the temperature swing. From this plot it can be easily seen that the reduction of thermal cycles in the *high-ΔT* tail of the distribution results into a noticeable improvement of the *MTTF*. In fact, the thermal cycles included between *60 K* and *70 K* are just *1.2%* of the total amount of cycles. Nevertheless, they account for *25%* up to *45%* of the total cumulated fatigue, depending on the considered *Coffin-Manson* parameters. *Equation 6.10* delivers a *MTTF* for a single bond wire, which is of *3 years* for the lower, of *6 years* for the average, and *16 years* for the higher estimate of the *Coffin-Manson* parameters.

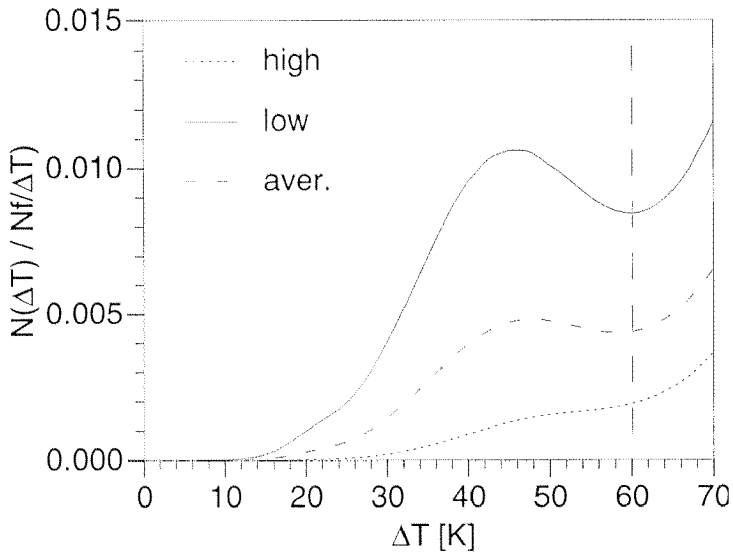


Figure 6.10 Low-stress fatigue cumulated by a single bond wire during 1-year operation in a locomotive as a function of ΔT.

Furthermore, the assumption that the modules used for high power applications include *32 chips*, having each *16 emitter bond wires per chip*, and that *4 bond wires* only per chip are allowed to fail, results into a complexity factor of *0.5*.

6.11 Final remarks

A model has been developed for extrapolating the lifetime of *IGBT* modules from experimental data. In the present case the model applies to the bond wire lift-off mechanism, only. The proposed model takes into account both the redundancy of the bond wires within an *IGBT* module and the fatigue damage due to realistic application profiles. Bond wire redundancy has been shown to play a minor role. On the contrary, the model is very sensitive both against the *Coffin-Manson* parameters, and the ΔT frequency distribution.

Finally, it should be stressed that the proposed numerical examples refer to data on aluminum bond wires, which are bonded on chips without any technological countermeasure for controlling the lift-off mechanism. Nowadays, dedicated solutions are implemented (like strain buffers or bond wire coatings) which have strongly reduced the impact of bond wire lift-off on the overall reliability characteristics of *IGBT* modules.

Appendix 1

Definitions & Reliability Fundamentals

This Appendix provides some basic concepts of the reliability theory, which are mainly used in Chapter 5 and 6. For more details refer to [106]. Table A1.1 includes the definition of the most important terms used in this thesis.

A.1 Failure Rate

In the following a heuristic definition of the failure rate λ is presented which is straightforward for the computation of the failure rate from data originated either from field or from accelerated tests. For a more complete and mathematically consequent treatment of this matter refer to [106].

When n statistically identical and independent devices are put into operation at t=0 (begin of the operation in the field or of the accelerated stress) under the same stress conditions, the number of the devices, which

are not yet failed at the time t may be expressed by the decreasing step function $\bar{v}(t)$. t_1, ..., t_n are the observed failure-free operating times of the n devices. The hook on the top of a symbol designates the empirical mean value of the expected failure free time

$$\hat{E}[\tau] = \frac{t_1 + \ldots + t_n}{n} \qquad \text{(A1.1)}$$

which converges towards the true expected failure free time $E(\tau)$ for a lot with $n \to \infty$.

The fraction of devices, which are still operating at time t expressed by the empirical reliability function

$$R(t) = \frac{\bar{v}(t)}{n} \qquad \text{(A1.2)}$$

which converges towards the true reliability function or a lot with $n \to \infty$. Thus the reliability function represents also the survival probability of a single item within the lot.

Table A1.1 Definitions

Reliability R(t)	Probability that an item will perform its required function under given conditions for a stated time interval.
Failure	Termination of the ability of an item to perform a required function.
Failure Mode	Symptom by which a failure is manifested (open, short, drift, functional failure, ...)
Failure Mechanism	Failure mechanism is the physical, chemical, or other process resulting in a failure.
Failure Cause (Root Cause)	Technology or process related cause producing the failure
Failure Rate λ(t)	Limit for $\delta t \to 0$ (if it exists) of the probability that an item will fail in the time interval $]t, t+\delta t]$, given that the item was new at t=0, and did not fail in the interval $]0, t]$, divided by δt.
Life Time	Time span between initial operation and failure of a non-repairable item.
MTTF	Expected value (mean) of an item's failure-free operating time (Mean Time To Failure). An empirical estimate for MTTF is $\hat{MTTF} = (t_1 + \ldots + t_n)/n$, where t_1 ,.. .. , t_n are the observed failure-free operating times of n statistical identical items.
Median (t_{50})	0.5 quantile of a distribution

The ratio of the devices which are failed during the time interval]t, t+δt] with the number of devices which are still operating at the time t

$$\hat{\lambda}(\tau) = \frac{\bar{v}(\tau) - \bar{v}(\tau + \delta t)}{\bar{n}(t)\,\delta t} \qquad \text{(A1.3)}$$

or

$$\hat{\lambda}(\tau) = \frac{\hat{R}(\tau) - \hat{R}(\tau + \delta\tau)}{\hat{R}(\tau)\,\delta t} \qquad \text{(A1.4)}$$

is defined as the empirical failure rate, which converges towards the true failure rate

$$\lambda(t) = \frac{-1}{R(t)}\frac{dR(t)}{dt} \qquad \text{(A1.5)}$$

when $n\,\delta\tau \to \infty$. The unit for measuring the failure rate is the FIT (Failures In Time), which corresponds to one failure for one billion of device-hours. The failure rate function fully defines the reliability function through

$$\ln(R(t)) = -\int_0^t \lambda(x)\,dx \qquad \text{(A1.6)}$$

with the initial condition $R(0) = 1$. In converse, the failure rate can be extracted from the reliability function R(t) through Equation A1.9.

A.2 Distributions

A cumulative distribution function F(t) represents the population fraction, failing by age t, i.e.

$$F(t) = 1 - R(t) \qquad \text{(A1.7)}$$

where $F(t \to -\infty) = 0$ and $F(t \to \infty) = 1$. The density f(t) of a derivable distribution F(t) is defined as

$$f(t) \equiv \frac{dF(t)}{dt} \qquad \text{(A1.8)}$$

yielding

$$\lambda(t) = \frac{f(t)}{1 - F(t)} \qquad (A1.9)$$

In order to estimate the failure rate from experimental data the empiric cumulative distribution (step function) has to be computed from the observed times to the failure t_i. Assumed that n failures with the related failure free times t_1, ..., t_n have been observed, one should proceed as follows:

a. Rank the failure free times t_1, ..., t_n such that $t(1) \leq t(2) \leq .... \leq t(n)$
b. Build the empiric function $\hat{F}(t)$: $t(i) \rightarrow \hat{F}(t_{(i)}) = i/n$

Assuming that the data are distributed according a known cumulative distribution, the unkown distribution parameters may be extracted either by fitting the data (maximum likelihood) of such a distribution function, or by a graphic techniques. Graphic techniques can be simplified by using a suitable probability chart, where a given distribution (in the present case $\hat{F}(t_{(i)})$) appears as a straight line.

An additional parameter which can be easily extracted from the plot of the empiric cumulative function is the median t_{50} or the time t_f for reaching an arbitrary quantile f of the distribution. The median represents here the time required for getting 50% of the device population failed.

In the reliability analysis different distribution functions are used, depending on the nature of the degradation mechanism which is investigated. In many cases the experimental data cannot be described just by a single distribution, but a combination of two or more functions is required. For instance this is the case when two or more failure mechanisms occur with different time behavior. This situation leads frequently to S-shaped curves when representing the empiric distribution function using probability charts.

Exponential Distribution – The exponential distribution is of theoretical importance, because it represents a reference case for all existing models. Furthermore, any other distribution can be approximated by a sum of exponentials. It is typically used if a constant failure rate is expected during the whole operating period, i.e. when the behaviour of a device

does not depend on how long it has been already operated in the past (memory-less process). It represents reasonably the bottom of the bathtub distribution, which is characterized by failures occurring randomly. This is valid in particular when the devices are mature, they have been properly screened, or they have not yet reached the end-of-life region. Early and wearout failures are not correctly described by the exponential distribution.

The failure free time, density, failure rate, and expected failure free time of the exponential and of the Weibull distributions are listed in Table A1.2.

In reality, non-repairable systems are already subject to wearout. Thus a model postulating a constant failure can only be a momentary approximation of a time-dependent failure rate. For this reason the use of MTTF = $1/\lambda$ for the extrapolation of the lifetime of a system can lead to wrong results. For example, a momentary failure rate of 10 FIT does not necessarily correspond to a lifetime of 11415 years. In addition, it should be noted that even in the case of a purely exponential distribution about 63% of the devices are already failed at t = MTTF.

Table A1.2 Relevant reliability parameters for the exponential and the Weibull distributions

	$F(t)$	$f(t)$	$\lambda(t)$	$E[\tau]$
Exponential	$1-e^{-\lambda t}$	$\lambda e^{-\lambda t}$	λ	$1/\lambda$
Weibull	$1-e^{-(\lambda t)^{\beta}}$	$\lambda\beta(\lambda t)^{\beta-1}e^{-(\lambda t)^{\beta}}$	$\lambda\beta(\lambda t)^{\beta-1}$	$\Gamma(1+1/\beta)/\lambda$

Appendix 2

Approximate solution of the elliptic integral in Chapter 5

In present Appendix, we propose an approximate solution of the elliptic integral in Equation 5.7

$$I = \int_0^t e^{-\frac{a}{x}}\, dx \qquad \text{(A2.1)}$$

which enables to estimate the error introduced by the approximation. The function to be integrated in Equation 5.7 can be approximated by

$$f(x) = (1 + \varepsilon\, x)\, e^{-\frac{a}{x}} \qquad \text{(A2.2)}$$

where ε is supposed to be a numerical factor less than one. The function $f(x)$ becomes an exact differential for

$$\varepsilon = \frac{2}{a} \tag{A2.3}$$

that is

$$\left(1+\frac{2}{a}x\right)e^{-\frac{a}{x}} = \frac{d}{dx}\left(\frac{x^2}{a}e^{-\frac{a}{x}}\right) = \frac{d}{dx}g(x) \tag{A2.4}$$

Thus the function $g(x)$ can be used for approximating the integral of Equation 5.7.

Since in present case $a \approx 350$ and $x < 20$, we have

$$\frac{2x}{a} < 10\% \tag{A2.5}$$

Thus the approximation of the integral in Equation 5.7 through the function $g(x)$ is within 10% of the analytical value.

References

[1] R. Zehringer, A. Stuck, T. Lang
Material requirements for high voltage, high power IGBT devices
Solid-State-Electronics 42(1998)2139-2151

[2] H. Zeller
High power components: from the state of the art to future trends
PCIM Nurnberg (1998)1-10

[3] D. Crook
Evolution of VLSI reliability engineering
IEEE International Reliability Physics Symposium, IRPS 28(1990)2-11

[4] T. Schütze, H. Berg, M. Hierholzer
Further improvements in the reliability of IGBT modules
Proc. 1998 IEEE Industry Applications Conference, IAS 33(1998)1022-1025

[5] P. Cova, G. Nicoletto, A. Pirondi, M. Portesine; M. Pasqualetti
Power cycling on press-pack IGBTs: measurements and thermomechanical simulation
Microelectronics-Reliability 39(1999)1165-1170

[6] M. Ciappa
Package Reliability in Microelectronics: a Review
Proc. International Workshop on Electronics and Detector Cooling
WELDEC 1(1994)133-149

[7] P. Cova, M. Ciappa, G. Franceschini, P. Malberti, F. Fantini
Thermal characterization of IGBT power modules
Microelectronics Reliability 37(1997)1731-1734

[8] C. Hager
Lifetime estimation of Al wire bonds based on computational plasticity
ETH Thesis Nr.13763

[9] S. Ramminger, P. Türkes, G. Watchutka
Crack mechanism in wire bonding joints
Microelectronics Reliability 38(1998)1301-1305

[10] H. Frost, M. Ashby
The plasticity and creep of metals and ceramics
Pergamon Press, 1982

[11] P. Scacco, M. Ciappa
Reliability Laboratory of the Swiss Federal Institute of Technology (ETH), Zurich (ETH)
unpublished results 1994

[12] A. Hamidi, N. Beck, K. Thomas, E. Herr
Reliability and lifetime evaluation of different wire bonding technologies for high power IGBT modules
Microelectronics Reliability 39(1999)1153-1158

[13] H. Schafft
Testing and fabrication of wire bonds electrical connections – A comprehensive survey
National Bureau of Standards, Tech. Note 726(1972)106-109

[14] *Reliability analysis/assessment of advanced technologies*
Technical Report RADC-TR-90-72

[15] C. Santoro
Thermal cycling and surface reconstruction in aluminum thin films
Journal of the Electrochemical Society 116(1969)361-364

[16] E. Philofsky, K. Ravi, E. Hall, J. Black
Surface reconstruction of aluminum metalization – a new potential failure mechanism
IEEE International Reliability Physics Symposium 9(1971)120-128

[17] M. Ciappa, P. Malberti
Plastic-strain of aluminum interconnections during pulsed operation of IGBT multichip modules
Quality and Reliability Engineering International 12(1996)297-303

[18] P. Malberti, M. Ciappa, R. Cattomio
A Power-cycling-induced failure mechanism of IGBT multichip modules
International Symposium for Testing and Failure Analysis 21(1995)163-168

[19] P. Lall, M. Pecht, E. Hakim
Influence of temperature on microelectronics and system reliability
CRC Press, 1997

[20] L. Ciampolini, M. Ciappa, P. Malberti, P. Regli, W. Fichtner
Modelling Thermal Effects of Large Contiguous Voids in Solder Joints
Microelectronics Journal 30(1999)1115-1123

[21] M. Ciappa
Proc. of the International Course on Failure Mechanisms and Failure Analysis of Semiconductor Devices
Swiss Federal Institute of Technology (ETH), April 2-4, 1996

[22] *Uhlig's corrosion handbook*
R. Winston Editor
Wiley, 2000

[23] D. Olsen, R. Wright, H. Berg
Effects of intermetallics on the reliability of tin coated Cu, Ag, and Ni parts
IEEE International Reliability Physics Symposium 13(1975)80-86

[24] M. Rodriguez, N. Shammas, N. Plumpton, D. Newcombe, D. Crees
Static and dynamic finite element modeling of thermal fatigue effects in IGBT modules
Microelectronics Reliability 40(2000)455-463

[25] G. Mitic, R. Beinert, P. Klofac, H. Schultz, G. Lefranc
Reliability of AlN substrates and their solder joints in IGBT power modules
Microelectronics Reliability 39(1999)1159-1164

[26] E. Herr, T. Frey, R. Schlegel, A. Stuck, R. Zehringer
Substrate-to-base solder joint delamination in high power IGBT modules
Microelectronics Reliability 37(1997)1719-1722

[27] H. Zeller
RAPSDRA project BE 95-2105, 30/36m Report

[28] T. Stockmeier, U. Schlapbach
1200A, 3300 V IGBT power module exhibiting very low internal stray inductance
Proc. of the PCIM Conference PCIM(1998)331-337

[29] S. Gekenidis, E. Ramezani, H. Zeller
Explosion tests on IGBT high voltage modules
International Symposium on Power Semiconductor Devices and ICs. ISPSD 11(1999)129-132

[30] P. Palmer, J. Joyce, B. Stark
Measurement of chip currents in IGBT modules
European Conference on Power Electronics and Applications EPE 7(1997)406-411

[31] B. Baliga
Modern power devices
Wiley, 1987

[32] F. Bauer, T. Stockmeier, H. Lendenmann, H. Dettmer, W. Fichtner
MOS controlled power switches above 2000 V: MCT versus IGBT
Proc. of the 26th International Power Conversion Conference, PCIM(1993)312-326

[33] N. Iwamuro, A. Okamoto, S. Tagami, H. Motoyama
Numerical analysis of short circuit safe operating area for p-channel and n-channel IGBTs
IEEE Transactions on Electron Devices 38(1991)303-308

[34] H. Zeller
Cosmic ray induced failures in high power semiconductor devices
Microelectronics Reliability 37(1997)1711-1718

[35] C. Findeisen, E. Herr, M. Schenkel, R. Schlegel, H. Zeller
Extrapolation of cosmic ray induced failures from test to field conditions for IGBT modules
Microelectronics Reliability 38(1998)1335-1339

[36] L. Reimer
Scanning Electron Microscopy
Springer, 1998

[37] *Electron and Optical Beam Testing of Integrated Circuits*
Special Issue of Microelectronics Engineering 1-4(1993)1-585
A. Birolini, M. Ciappa, and E. Wolfgang Editors

[38] M. Ciappa, P. Malberti, P. Furcas, M. Vanzi
A new adaptive amplifier for biased electron beam induced current applications
Microelectronics Reliability 38(1998)889-893

[39] J. Soden, C. Hawkins
Test considerations for gate oxide shorts in CMOS ICs
IEEE Design and Test of Computers 3(1986)56-54

[40] P. Malberti, M. Ciappa
Handbook on basic failure analysis techniques for IGBTs
Integrated Systems Laboratory, Technical Report No. 98/38
November 1998

[41] P. Malberti, M. Ciappa
Selective wet-etch of silicon nitride passivation layers
Proc. of the 24th International Symposium for Testing and Failure Analysis, ISTFA 24(1998)429-435

[42] P. Scacco, P. Malberti, M. Ciappa
Wet-Etch of Nitride Passivation Layers: An Effective Alternative to Plasma-Etch for Failure Analysis
Proc. of the 20th International Symposium for Testing and Failure Analysis, ISTFA 20(1994)157-161

[43] Y. Kunii
Wet etching of doped and non doped silicon oxide films using buffered hydrogen fluoride solutions
Proc. of the 3rd Symposium of the Electrochemical Society on Silicon Nitride and Silicon Dioxide Thin Films (1994)261-266

[44] V. Rathi
The dependence of the etch rate of photo-CVD silicon nitride films on the NH_4F content in buffered HF solutions
Microelectronics Journal 26(1995)563-567

[45] T. Shankoff
Controlling the interfacial oxide layer of Ti-Al contacts with the CrO_3-H_3PO_4 etch
Journal of the Electrochemical Society 125(1978)467-471

[46] M. Ciappa, P. Malberti
Proceeding of the International Course on sample preparation techniques for the failure analysis of silicon and III-V devices
Swiss Federal Institute of Technology Zurich, February and March 1997

[47] M. Wright
A new preferential etch for defects in silicon crystals
Journal of the Electrochemical Society 124(1977)757-762

[48] F. Secco d'Aragona
Dislocation etch for (100) planes in silicon
Journal of the Electrochemical Society 119(1972)948-951

[49] E. von Sirtl, A. Adler
Chromsäure-Flusssäure als spezifisches System zur Ätzengrubenentwicklung auf Silizium
Zeitschrift für Metallkunde 52(1961)529-531

[50] D. Schimmel
Defect etch for (100) silicon ingot evaluation
Journal of the Electrochemical Society 126(1979)479

[51] P. Malberti, L. Ciampolini, M. Ciappa, W. Fichtner
Quantification of Scanning Capacitance Microscopy Measurements for 2D Dopant Profiling
Microelectronics Reliability 40(2000)1395-1399

[52] P. de Wolf, R. Stephenson, T. Trenkler, T. Clarysse, T. Hantschel, W. Vandervorst
Status and review of 2D carrier profiling using scanning probe microscopy
Journal of Vacuum Science and Technology B18(2000)361-368

[53] P. Malberti, R. Bottini
Private communication

[54] J. Kolzer, C. Boit, A. Dallmann, G. Deboy, J. Otto, D. Weinmann
Quantitative emission microscopy
Journal of Applied Physics 71(1992)R23-41

[55] G. Barbottin, A. Vapaille
Instabilities in silicon devices, Volume 1
Noth Holland (1986)

[56] D. DiMaria, T. Theis, J. Kirtley, F. Pesavento, D.Dong
Electron heating in silicon dioxide and off-stoichiometric silicon dioxide films
Journal of Applied Physics 57(1985)1214-1238

[57] Panasolve 215, Elosol AG, Dufourstr. 101, CH-8034 Zurich

[58] L. Ciampolini, M. Ciappa, P. Malberti, W. Fichtner
Two-dimensional simulation of scanning capacitance microscopy measurements of arbitrary doping profiles
Proc. of the International Conference on Modeling and Simulation of Microsystems (MSM 2000), S. Diego, March 27-29, 2000

[59] L. Ciampolini, M. Ciappa, P. Malberti, W. Fichtner
Investigating the accuracy of constant-dC scanning capacitance by finite element device simulation
Proc. of the 1st Workshop on Ultimate Integration of Silicon (ULIS'2000), January 21, 2000

[60] R. Wiesendanger
Scanning probe microscopy: analytical methods
Springer Verlag, Berlin 1998

[61] L. Ciampolini, M. Ciappa, P. Malberti
Scanning capacitance microscopy: measurement and simulation
IIS Technical Report 2000/04

[62] P. Malberti, M. Ciappa, P. Scacco
A new back etch for silicon devices
Proc. of the 21st International Symposium for Testing and Failure Analysis, ISTFA 21(1995)257-261

[63] M. Abramo, R. Wasielewski
FIB for failure analysis
Semiconductor International 20(1997)133-134

[64] T. Moore, C. Hartfield
Trends in non-destructive imaging of IC packages
Proc. of the American Institute of Physics Conference 449(1998)598-604

[65] G. Quinones, E. Allen
Comparison study of lifetime measurement techniques
Proceedings of the SPIE 3509(1998)137-146

[66] B. Simmacher, G. Deboy, M. Ruff, H. Schulze, B. Kolbesen
Analysis of the carrier and temperature distributions in gate turn-off thyristors by internal laser deflection
IEEE International Symposium on Power Semiconductor Devices ISPSD (1997)177-180

[67] P. Palmer, J. Joyce
Current redistribution in multi-chip IGBT modules under various gate drive conditions
7th International Conference on Power Electronics and Variable Speed Drives 7(1998)246-251

[68] A. Hamidi, G. Coquery, R. Lallemand, P. Vales, J. Dorkel
Temperature measurements and thermal modeling of hihgh power multichip modules for reliability investigations in traction applications
Microelectronics Reliability 38(1998)1353-1359

[69] K. Nassim, L. Joannes, A. Cornet, S. Dilhaire, E. Schaub, W. Claeys
Thermomecanical deformation imaging of power devices by electronic speckle pattern interferometry
Microelectronics Reliability 38(1998)1341-1345

[70] V. Nerozzi, D. Diversi, P. Mignardi, C. Palazzini
Field reliability results in thyristors and diodes employed in choppers and inverters for Italian railway locomotives
Workshop on Power Devices at the 5th European Symposium on Reliability of Electron Devices Failure Physics and Analysis (ESREF), Glasgow 1994

[71] M. Ciappa
RAPSDRA Task 7, 24th month Report

[72] Y. Gerstenmaier, G. Wachutka
A new procedure for the calculation of the temperature development in electronic circuits
Proc. of the 8th European Conference on Power Electronics and Applications (EPE) 8(1999)1-10

[73] M. Ciappa, A. Orzati
RAPSDRA Task 7, 30th month Report

[74] C. Yun, M. Ciappa, P. Malberti, W. Fichtner
Thermal Component Model for Electro-Thermal Analysis of IGBT Module Systems
IEEE 2nd International Workshop on Chip-Package Co-Design, CPD-2000

[75] C. Yun
Static and dynamic behavior of IGBT power modules
ETH Thesis Nr.13784, 2000

[76] A. Gazzola
Determinazione dell'impedenza termica negli IGBT
MS Thesis 1995-96, University of Parma
Supervisors: M. Ciappa, F. Fantini

[77] P. Malberti, M. Ciappa
The thermal impedance measurement: an efficient failure analysis tool for laser diodes
Proc. of the 20th Int. Symposium for Testing and Failure Analysis, ISTFA 20(1994)97-105

[78] MIL-STD-750D, Method 3103
Thermal impedance measurements for Insulated Gate Bipolar Transistors
28 February 1995

[79] D. Blackburn, F. Oettinger
Transient thermal response measurements of power transistors
Proc. IEEE Power Electronics Specialists Conference, PESC (1974)140-140

[80] M. V. Fischetti
Generation of positive charge in silicon dioxide during avalanche and tunnel electron injection
Journal of Applied Physics 57(1985)2860

[81] M. Shatzkes, M. Av-Ron
Impact ionization and positive charge in SiO_2 films
Journal of Applied Physics 47(1975)3192

[82] D. Dimaria, D. Arnold, E. Cartier
Degradation and breakdown of silicon dioxide films on silicon
Applied Physics Letters 61(1992)2329

[83] E. Herr
Gate oxide integrity of BiMOS power devices
ETH Thesis Nr. 10678, 1994

[84] M. Liehr, G. Bronner, J. Lewis
Stacking-faults-induced defect creation in SiO2 on Si(100)
Applied Physics Letters 52(1988)1892

[85] K. Honda, A. Ohsawa, N. Toyokura
Breakdown in silicon oxides – correlation with Cu and Fe precipitates
Applied Physics Letters 45(1984)270
Applied Physics Letters 46(1985)582

[86] R. Degraeve, B. Kaczer, G. Groeseneken
Degradation and breakdown in thin oxide layers: mechanisms, models, and reliability prediction
Microelectronics Reliability 39(1999)1424

[87] M. Liang, C. Hu
Electron trapping in very thin thermal silicon oxides
IEDM Technical Digest (1981)396

[88] D. DiMaria, E. Cartier, D. Buchanan
Anode hole injection and trapping in silicon dioxide
Journal of Applied Physics 80(1996)304

[89] I. Chen, S. Holland, C. Hu
Electrical breakdown inthin gate and tunneling oxides
IEEE Transaction on Electron Devices ED-32(1985)413

[90] I. Chen, S. Holland, K. Young, C. Chang, C. Hu
Substrate hole current and oxide breakdown
Applied Physics Letters 49(1986)669

[91] D. DiMaria, E. Cartier, E. Arnold
Impact ionization trap creation, degradation, and breakdown of silicon dioxide films on silicon
Journal of Applied Physics 73(1993)3367

[92] Y. Nissan-Cohen, J. Shappir, D. Frohman-Bentchkowsky
Trap generation and occupation dynamics in SiO_2 under charge injection stress
Journal of Applied Physics 60(1986)2024

[93] J. McPherson, H. Mogul
Disturbed bonding states in SiO2 thin-films and their impact on time-dependent dielectric breakdown
IEEE International Reliability Physics Symposium 36(1998)47

[94] E. Vincent, S. Bruyere, C. Papadas, P. Mortini
Dielectric reliability in deep-submicron technologies from thin to ultrathin oxides
Microelectronics Reliability 37(1997)1499

[95] J. McPherson
Stress dependent activation energy
IEEE International Reliability Physics Symposium 24(1986)12

[96] I. Chen, S. Holland, C. Hu
A quantitative physical model for time-dependent breakdown
IEEE International Reliability Physics Symposium 23(1985)24

[97] E. Rosenbaum
Accelerated testing of SiO_2 reliability
IEEE Transaction on Electron Devices ED-43(1996)70

[98] R. Moazzami, J. Lee, I. Chen, C. Hu
Projecting the minimum acceptable oxide thickness foe time-dependent dielectric breakdown
IEDM Technical Digest (1988)710

[99] K. Schuegraf, C. Hu
Effects of temperature and defects on breakdown lifetime of thin SiO_2 at low voltages
IEEE International Reliability Physics Symposium 32(1994)126

[100] C. Osburn, D. Ormond
Dielectric breakdown in silicon dioxide films on silicon
Journal of the Electrochemical Society 119(1972)597

[101] R. Moazzami, J. Lee, I. Chen, C. Hu
Temperature accelration of time-dependent dielectric breakdown
IEEE Transaction on Electron Devices ED-36(1989)2402

[102] A. Berman
Time-zero dielectric reliability test by a ramp method
IEEE International Reliability Physics Symposium 19(1981)204

[103] G. Ghibaudo, G. Pananakakis, R. Kies, E. Vincent, C. Papadas
Accelerated dielectric breakdown and wear out standard testing methods and structures for reliability evaluation of thin oxides
Microelectronics Reliability 39(1999)597

[104] R. Degraeve, J. Olgier, R. Bellens, P. Roussel, G. Groeseneken, H. Maes
On the field dependence of intrinsic and extrinsic time-dependent dielectric breakdown
IEEE International Reliability Physics Symposium 34(1996)44

[105] B. Gnedenko
Lehrbuch der Wahrscheinlichkeitstheorie
Akademie-Verlag, Berlin 1970

[106] A. Birolini
Quality and reliability of technical systems
Springer Verlag, 1997

[107] H. Berg, E. Wolfgang
Advanced IGBT modules for railway traction applications: reliability testing
Microelectronic Reliability 38(1998)1319-1323

[108] P. Cova, F. Fantini
On the effect of power cycling stress on IGBT modules
Microelectronic Reliability 38(1998)1347-1352

[109] S. Dewar, G. Debled, E. Herr
A 1200 A, 3300 V IGBT power module for traction applications
PCIM, Nürnberg 1998

[110] M. Ciappa, W. Fichtner
Lifetime prediction of IGBT modules for reaction applications
IEEE International Reliability Physics Symposium 38(2000)210-216

Curriculum vitae

Mauro Ciappa is born in 1961 in Bellinzona (Switzerland), where he received his literary license in 1977 and his baccalaureate in sciences in 1980. He graduated in 1986 in experimental physics at the Physics Institute of the University of Zurich with a thesis work on *In-situ Characterization of electrolytic thin films by Rutherford Backscattering Spectrometry* (prof. Verena Meyer).

Mauro Ciappa joined in 1986 the Reliability Laboratory (RL, prof. Alessandro Birolini) of the Swiss Federal Institute of Technology (ETH), where he was Head of the Laboratory of Reliability Physics and Failure Analysis from 1988 to 1997. In this period, more than 130 consultancy cases for Swiss and foreign companies were issued under his responsibility.

During his stay at RL, Mauro Ciappa has been involved with leading responsibilities in following scientific projects: *Failure Mechanisms of Complex Integrated Circuits* (NF 2000-5.615, 1988-1991, techn. project leader), *Modelization of Failure Mechanisms and Faults in VLSI and ULSI Integrated Circuits* (NF 30'293.93, 1991-1993, techn. project leader), *Reliability Optimization of embedded EEPROM memories for ASIC applications* (KWF 2493.1, 1993-1995, techn. project leader), *Reliability of advanced high power semiconductor for railway traction*

applications RAPSDRA (BRITE BE 95-2105, 1995-1999, Task leader). He was also actively involved in additional projects like *Test and Screening Strategies of High-Density Semiconductor Memories* (KWF, 1988-1991), *Extreme Environment Technology for Spaceborne Electronic Assemblies* (European Space Agency, 1995-1996, WDP/PP/901-803), and LESIT Module 9 (1992-1995).

From 1993 to 1997, Mauro Ciappa has been lecturer for Reliability Physics and Failure Analysis Techniques at the Electrical Engineering Department of the ETH in Zurich. In 1996 and 1997 he has given invited lectures at the University of Parma (Italy) and at the University of Cagliari (Italy), respectively.

Since 1998, Mauro Ciappa is in charge for the Physical Characterization Group at the Integrated Systems Laboratory of the ETH Zurich (IIS, prof. Wolfgang Fichtner). His present activities cover physical characterization and analysis of semiconductor devices.

Mauro Ciappa published 5 invited papers. He authored or co-authored more than 40 contributed papers in international conferences and journals, and was co-editor of a monograph on electron and optical beam testing.

In 2000, Mauro Ciappa has been awarded by the Institute of Electrical and Electronic Engineers (IEEE) with the IEEE Third Millennium Medal for his contributions in the field of the Physics of Failures.