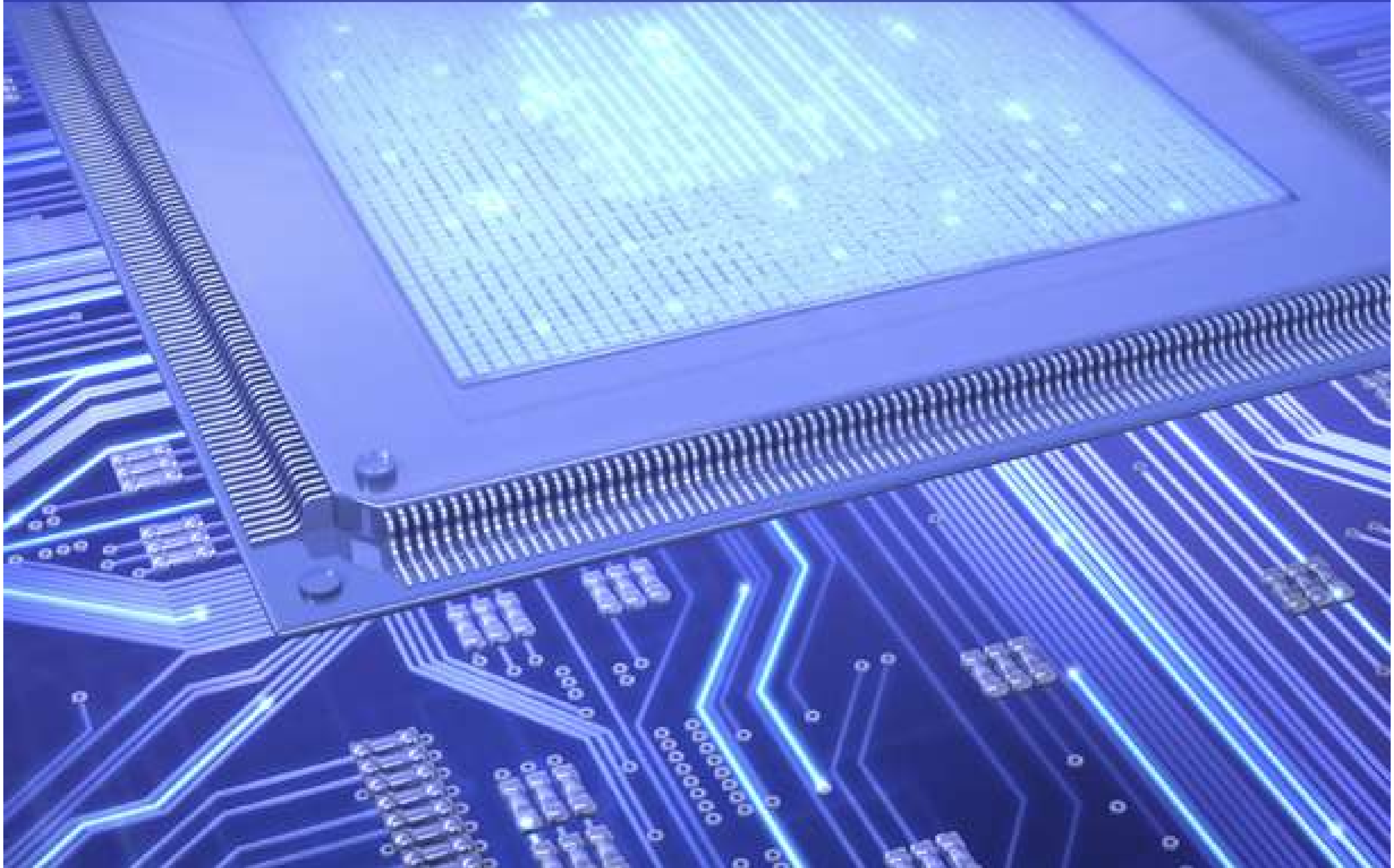
Semiconductors and Failure Modes of Electronics

Lashell Desantis
Frances Metzger



First Edition, 2012

ISBN 978-81-323-1360-1

Published by:
College Publishing House
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Semiconductor

A **semiconductor** is a material with electrical conductivity due to electron flow (as opposed to ionic conductivity) intermediate in magnitude between that of a conductor and an insulator. This means a conductivity roughly in the range of 10^3 to 10^{-8} siemens per centimeter. Semiconductor materials are the foundation of modern electronics, including radio, computers, telephones, and many other devices. Such devices include transistors, solar cells, many kinds of diodes including the light-emitting diode, the silicon controlled rectifier, and digital and analog integrated circuits. Similarly, semiconductor solar photovoltaic panels directly convert light energy into electrical energy. In a metallic conductor, current is carried by the flow of electrons. In semiconductors, current is often schematized as being carried either by the flow of electrons or by the flow of positively charged "holes" in the electron structure of the material. Actually, however, in both cases only electron movements are involved.

Common semiconducting materials are crystalline solids, but amorphous and liquid semiconductors are known. These include hydrogenated amorphous silicon and mixtures of arsenic, selenium and tellurium in a variety of proportions. Such compounds share with better known semiconductors intermediate conductivity and a rapid variation of conductivity with temperature, as well as occasional negative resistance. Such disordered materials lack the rigid crystalline structure of conventional semiconductors such as silicon and are generally used in thin film structures, which are less demanding for as concerns the electronic quality of the material and thus are relatively insensitive to impurities and radiation damage. Organic semiconductors, that is, organic materials with properties resembling conventional semiconductors, are also known.

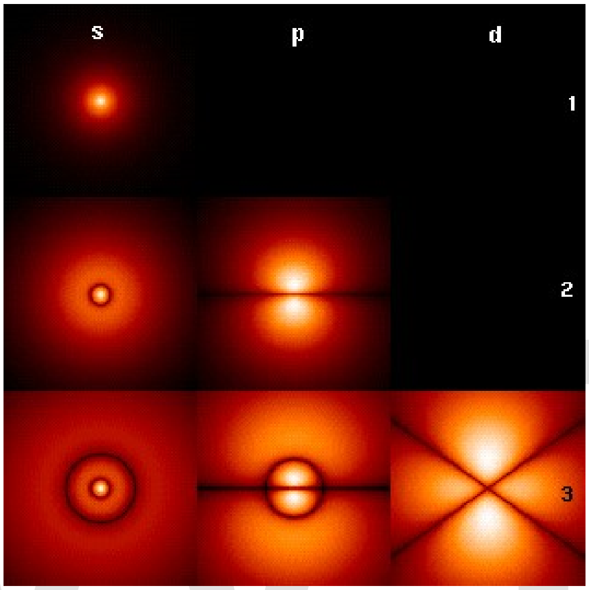
Silicon is used to create most semiconductors commercially. Dozens of other materials are used, including germanium, gallium arsenide, and silicon carbide. A pure semiconductor is often called an “intrinsic” semiconductor. The electronic properties and the conductivity of a semiconductor can be changed in a controlled manner by adding very small quantities of other elements, called “dopants”, to the intrinsic material. In crystalline silicon typically this is achieved by adding impurities of boron or phosphorus to the melt and then allowing the melt to solidify into the crystal. This process is called "doping".

Explaining semiconductor energy bands

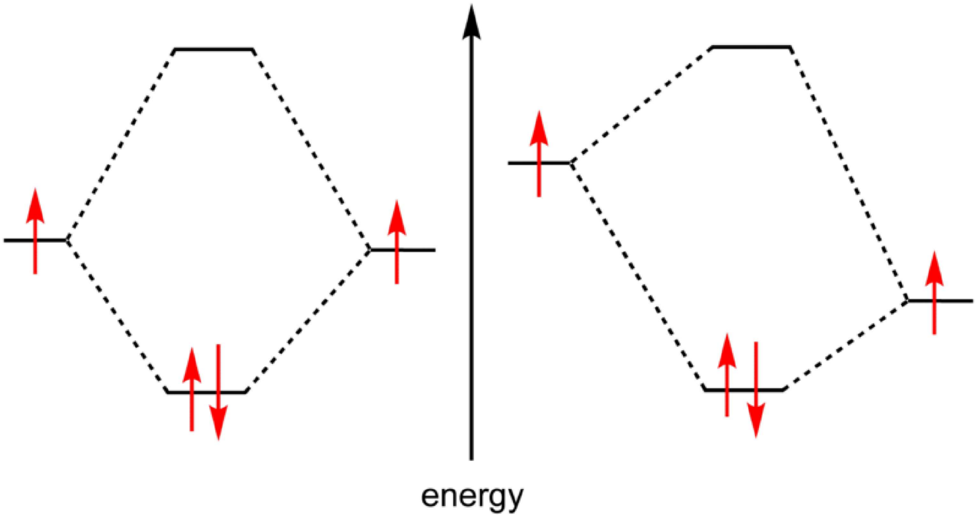
There are three popular ways to classify the electronic structure of a crystal.

Band structure

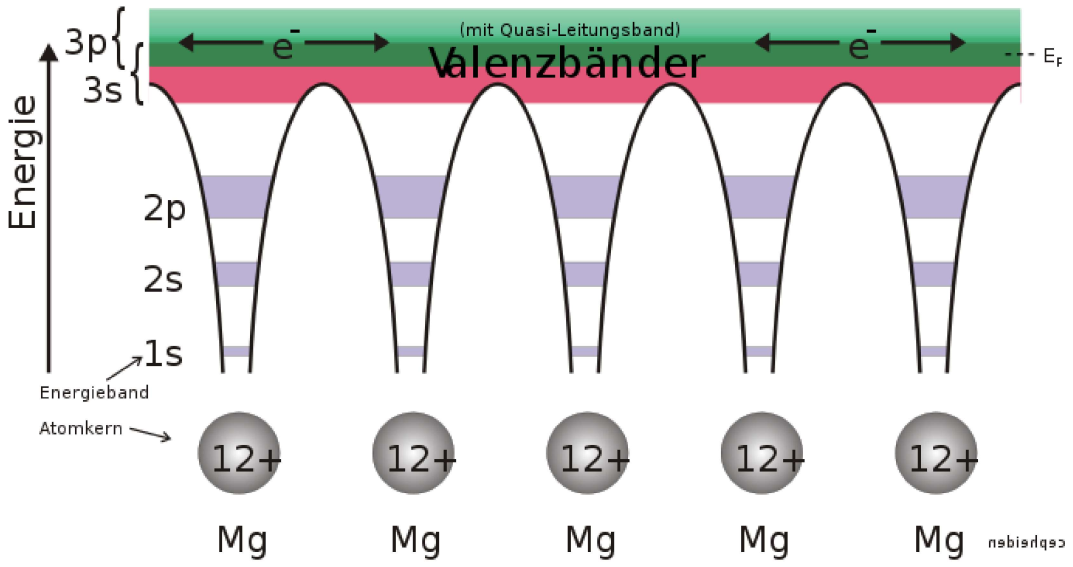
atoms – crystal – vacuum



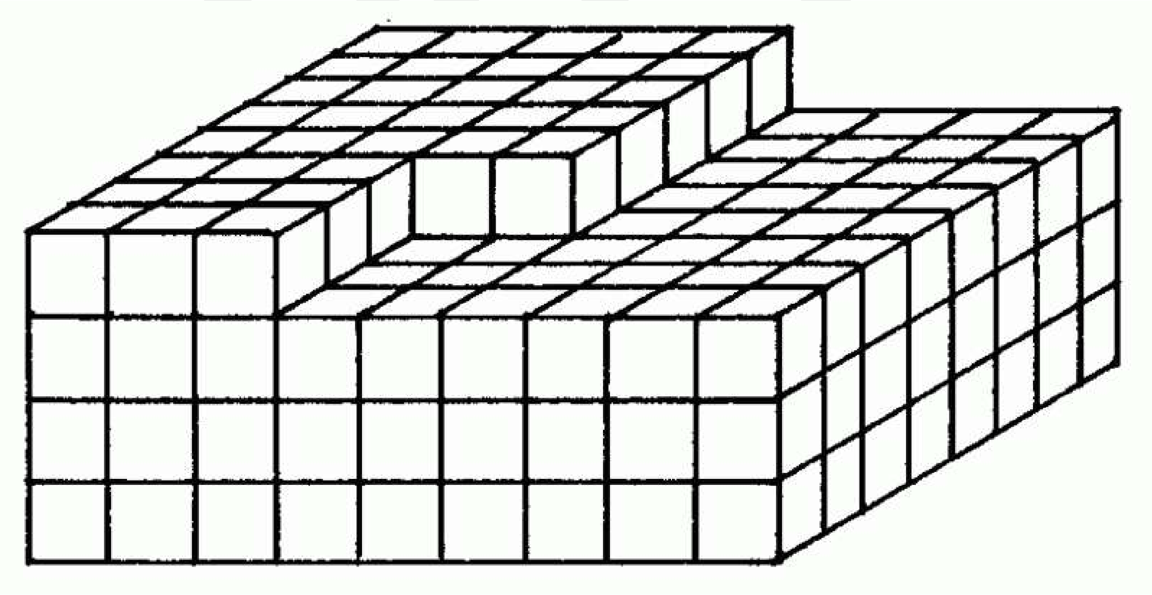
In a single H-atom an electron resides in well known orbitals. Note that the orbitals are called s,p,d in order of increasing circular current.



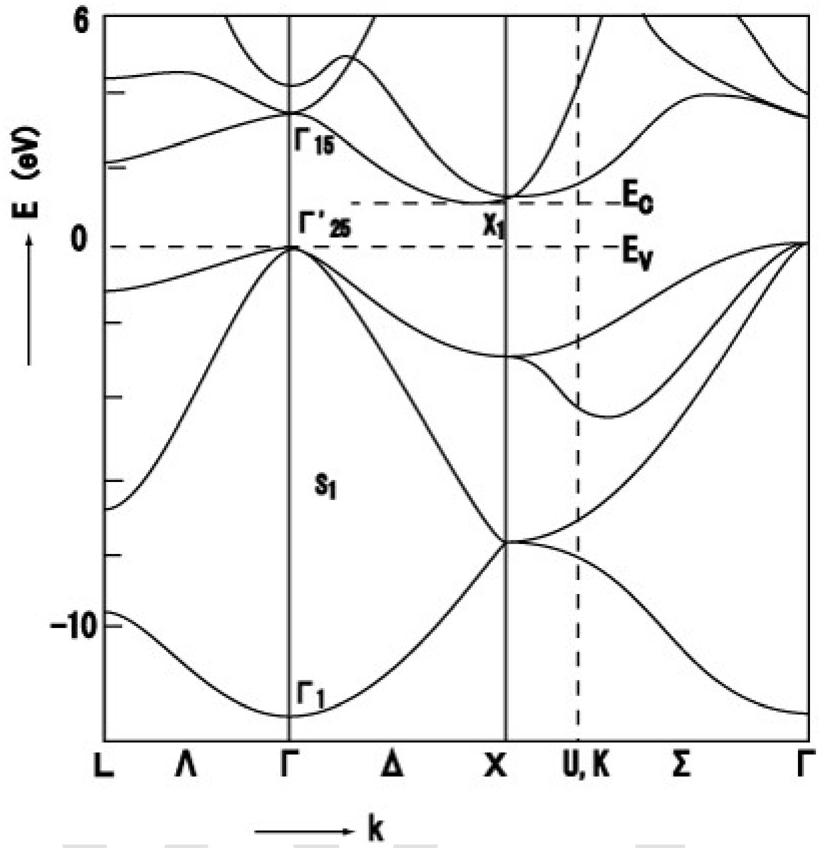
Putting two atoms together leads to delocalized orbitals across two atoms, yielding a covalent bond. Due to the Pauli exclusion principle, every state can contain only one electron.



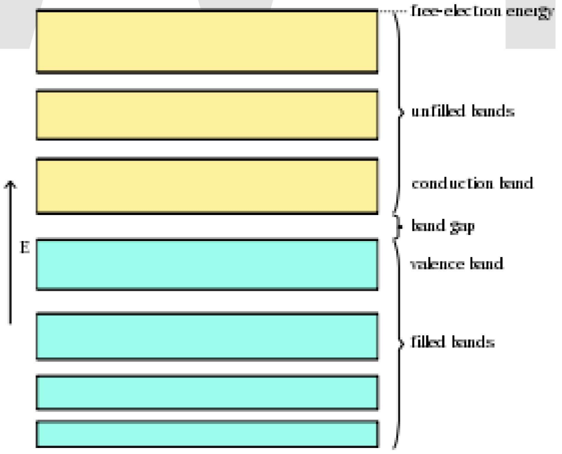
This can be continued with more atoms. Note: This picture shows a metal, not an actual semiconductor.



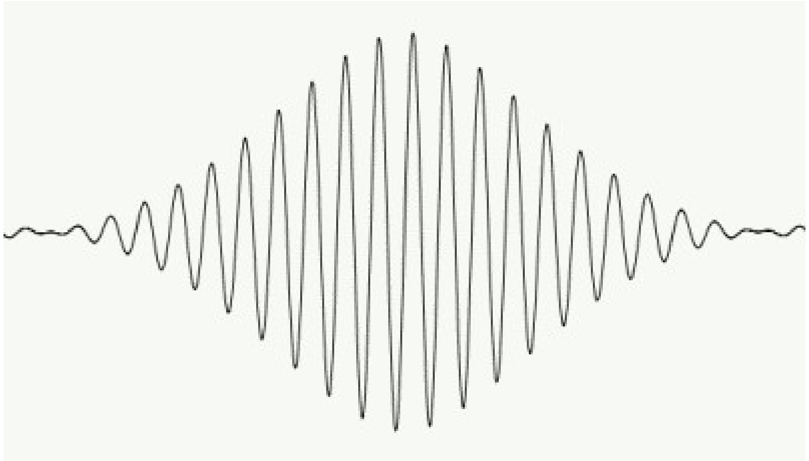
Continuing to add creates a crystal, which may then be cut into a tape and fused together at the ends to allow circular currents.



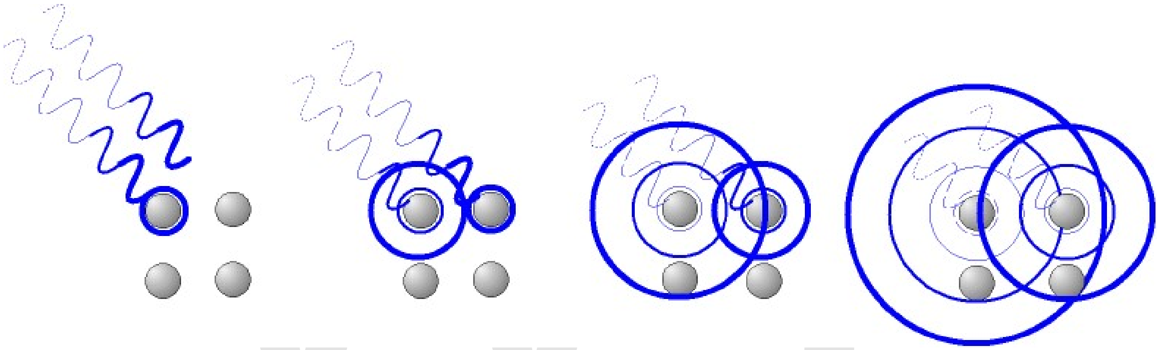
For this regular solid the band structure can be calculated or measured.



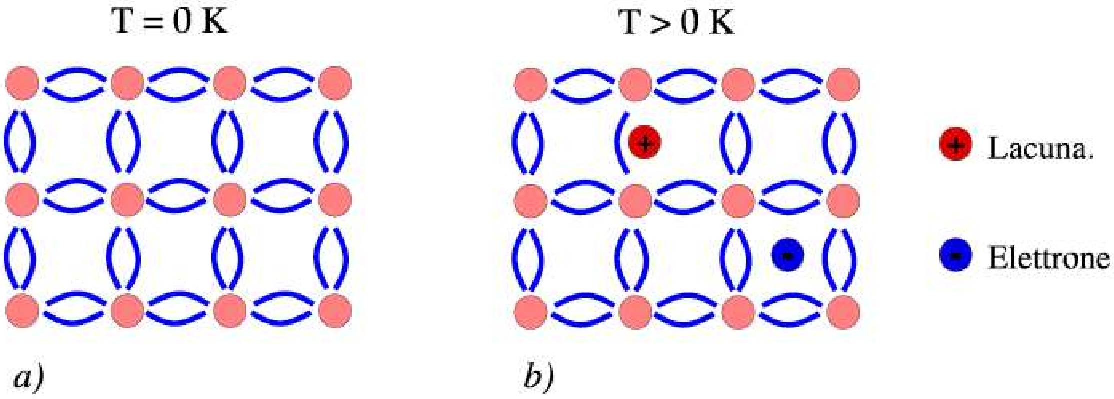
Integrating over the k axis gives the bands of a semiconductor showing a full valence band and an empty conduction band. Generally stopping at the vacuum level is undesirable, because some people want to calculate: photoemission, inverse photoemission



After the band structure is determined states can be combined to generate wave packets. As this is analogous to wave packages in free space, the results are similar.



An alternative description, which does not really appreciate the strong Coulomb interaction, shoots free electrons into the crystal and looks at the scattering.



A third alternative description uses strongly localized unpaired electrons in chemical bonds, which looks almost like a Mott insulator.

Energy bands and electrical conduction

In classic crystalline semiconductors, the electrons can have energies only within certain bands (i.e. ranges of levels of energy). Energetically, these bands are located between the energy of the ground state, corresponding to electrons tightly bound to the atomic nuclei of the material, and the free electron energy. The latter is the energy required for an electron to escape entirely from the material. The energy bands each correspond to a large number of discrete quantum states of the electrons, and most of the states with low energy (closer to the nucleus) are full, up to a particular band called the *valence band*. Semiconductors and insulators are distinguished from metals because the valence band in them is nearly filled with electrons under usual operating conditions, while very few (semiconductor) or virtually none (insulator) of them are available in the *conduction band*, the band immediately above the valence band.

The ease with which electrons in a semiconductor can be excited from the valence band to the conduction band depends on the band gap between the bands. The size of this energy bandgap serves as an arbitrary dividing line (roughly 4 eV) between semiconductors and insulators.

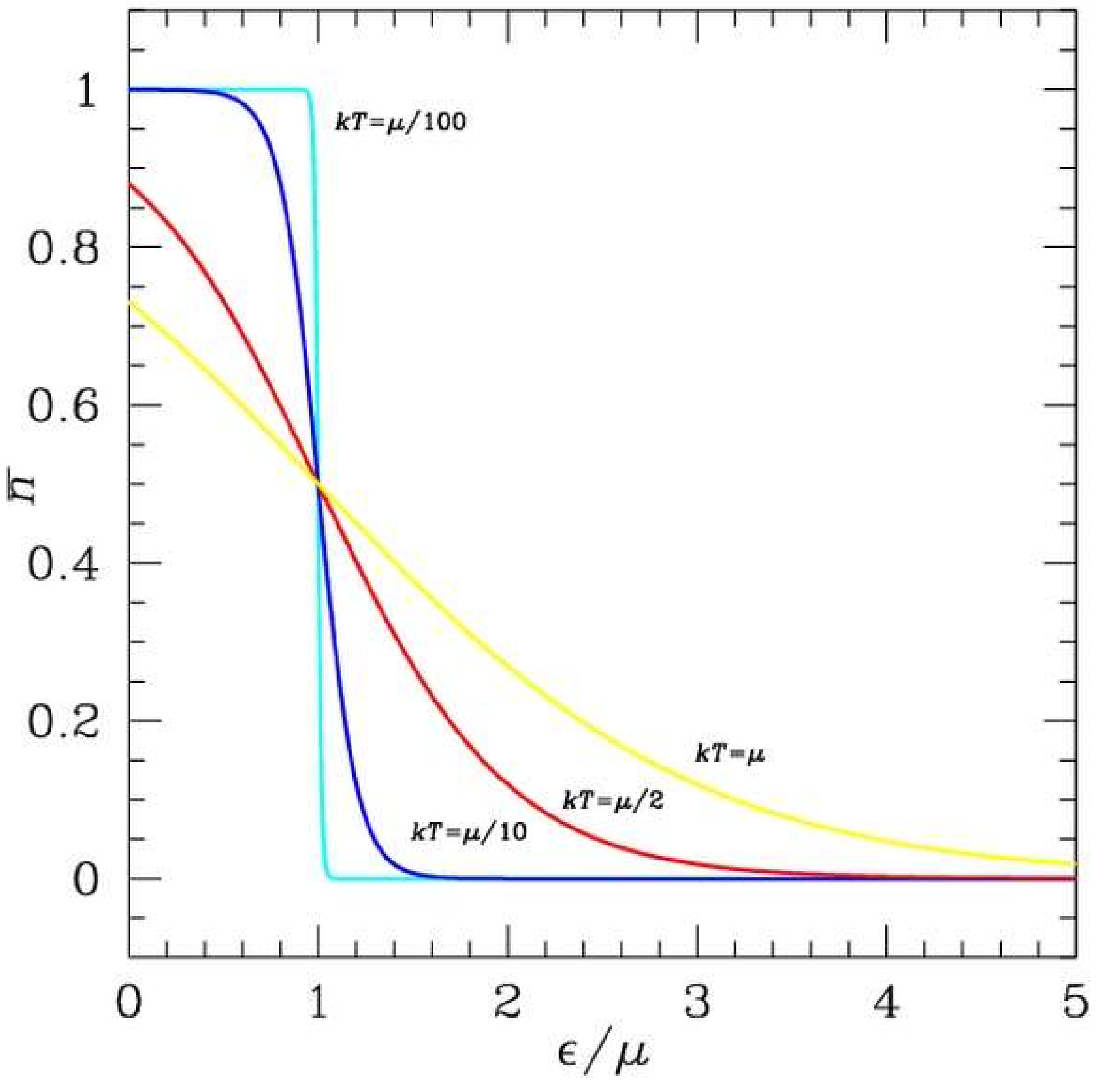
With covalent bonds, an electron moves by hopping to a neighboring bond. The Pauli exclusion principle requires the electron to be lifted into the higher anti-bonding state of that bond. For delocalized states, for example in one dimension – that is in a nanowire, for every energy there is a state with electrons flowing in one direction and another state with the electrons flowing in the other. For a net current to flow, more states for one direction than for the other direction must be occupied. For this to occur, energy is required, as in the semiconductor the next higher states lie above the band gap. Often this is stated as: full bands do not contribute to the electrical conductivity. However, as the temperature of a semiconductor rises above absolute zero, there is more energy in the semiconductor to spend on lattice vibration and — more importantly for us — on lifting some electrons into an energy states of the conduction band. The current-carrying electrons in the conduction band are known as "free electrons", although they are often simply called "electrons" if context allows this usage to be clear.

Electrons excited to the conduction band also leave behind electron holes, or unoccupied states in the valence band. Both the conduction band electrons and the valence band holes contribute to electrical conductivity. The holes themselves don't actually move, but a neighboring electron can move to fill the hole, leaving a hole at the place it has just come from, and in this way the holes appear to move, and the holes behave as if they were actual positively charged particles.

One covalent bond between neighboring atoms in the solid is ten times stronger than the binding of the single electron to the atom, so freeing the electron does not imply destruction of the crystal structure.

Holes: electron absence as a charge carrier

The concept of holes can also be applied to metals, where the Fermi level lies *within* the conduction band. With most metals the Hall effect indicates electrons are the charge carriers. However, some metals have a mostly filled conduction band. In these, the Hall effect reveals positive charge carriers, which are not the ion-cores, but holes. In the case of a metal, only a small amount of energy is needed for the electrons to find other unoccupied states to move into, and hence for current to flow. Sometimes even in this case it may be said that a hole was left behind, to explain why the electron does not fall back to lower energies: It cannot find a hole. In the end in both materials electron-phonon scattering and defects are the dominant causes for resistance.



Fermi-Dirac distribution. States with energy ε below the Fermi energy, here μ, have higher probability *n* to be occupied, and those above are less likely to be occupied. Smearing of the distribution increases with temperature.

The energy distribution of the electrons determines which of the states are filled and which are empty. This distribution is described by Fermi-Dirac statistics. The distribution is characterized by the temperature of the electrons, and the *Fermi energy* or *Fermi level*. Under absolute zero conditions the Fermi energy can be thought of as the energy up to which available electron states are occupied. At higher temperatures, the Fermi energy is the energy at which the probability of a state being occupied has fallen to 0.5.

The dependence of the electron energy distribution on temperature also explains why the conductivity of a semiconductor has a strong temperature dependency, as a semiconductor operating at lower temperatures will have fewer available free electrons and holes able to do the work.

Energy–momentum dispersion

In the preceding description an important fact is ignored for the sake of simplicity: the *dispersion* of the energy. The reason that the energies of the states are broadened into a band is that the energy depends on the value of the wave vector, or *k-vector*, of the electron. The k-vector, in quantum mechanics, is the representation of the momentum of a particle.

The dispersion relationship determines the effective mass, m^*, of electrons or holes in the semiconductor, according to the formula:

$$m^* = \hbar^2 \cdot \left[\frac{d^2E(k)}{dk^2}\right]^{-1}.$$

The effective mass is important as it affects many of the electrical properties of the semiconductor, such as the electron or hole mobility, which in turn influences the *diffusivity* of the charge carriers and the electrical conductivity of the semiconductor.

Typically the effective mass of electrons and holes are different. This affects the relative performance of *p-channel* and *n-channel* IGFETs.

The top of the valence band and the bottom of the conduction band might not occur at that same value of *k*. Materials with this situation, such as silicon and germanium, are known as *indirect bandgap* materials. Materials in which the band extrema are aligned in *k*, for example gallium arsenide, are called *direct bandgap* semiconductors. Direct gap semiconductors are particularly important in optoelectronics because they are much more efficient as light emitters than indirect gap materials.

Carrier generation and recombination

When ionizing radiation strikes a semiconductor, it may excite an electron out of its energy level and consequently leave a hole. This process is known as *electron–hole pair* *generation*. Electron-hole pairs are constantly generated from thermal energy as well, in the absence of any external energy source.

Electron-hole pairs are also apt to recombine. Conservation of energy demands that these recombination events, in which an electron loses an amount of energy larger than the band gap, be accompanied by the emission of thermal energy (in the form of phonons) or radiation (in the form of photons).

In some states, the generation and recombination of electron–hole pairs are in equipoise. The number of electron-hole pairs in the steady state at a given temperature is determined by quantum statistical mechanics. The precise quantum mechanical mechanisms of generation and recombination are governed by conservation of energy and conservation of momentum.

As the probability that electrons and holes meet together is proportional to the product of their amounts, the product is in steady state nearly constant at a given temperature, providing that there is no significant electric field (which might "flush" carriers of both types, or move them from neighbour regions containing more of them to meet together) or externally driven pair generation. The product is a function of the temperature, as the probability of getting enough thermal energy to produce a pair increases with temperature, being approximately $\exp(-E_G/kT)$, where k is Boltzmann's constant, T is absolute temperature and E_G is band gap.

The probability of meeting is increased by carrier traps—impurities or dislocations which can trap an electron or hole and hold it until a pair is completed. Such carrier traps are sometimes purposely added to reduce the time needed to reach the steady state.

Semi-insulators

Some materials are classified as **semi-insulators**. These have electrical conductivity nearer to that of electrical insulators. Semi-insulators find niche applications in micro-electronics, such as substrates for HEMT. An example of a common semi-insulator is gallium arsenide.

Doping

The property of semiconductors that makes them most useful for constructing electronic devices is that their conductivity may easily be modified by introducing impurities into their crystal lattice. The process of adding controlled impurities to a semiconductor is known as *doping*. The amount of impurity, or dopant, added to an *intrinsic* (pure) semiconductor varies its level of conductivity. Doped semiconductors are often referred to as *extrinsic*. By adding impurity to pure semiconductors, the electrical conductivity may be varied not only by the number of impurity atoms but also, by the type of impurity atom and the changes may be thousand folds and million folds. For example, 1 cm^3 of a metal or semiconductor specimen has a number of atoms on the order of 10^{22}. Since every atom in metal donates at least one free electron for conduction in metal, 1 cm^3 of metal contains free electrons on the order of 10^{22}. At the temperature close to 20 °C , 1 cm^3 of pure germanium contains about 4.2×10^{22} atoms and 2.5×10^{13} free electrons and 2.5×10^{13} holes (empty spaces in crystal lattice having positive charge) The addition of 0.001% of arsenic (an impurity) donates an extra 10^{17} free electrons in the same volume and the electrical conductivity increases about 10,000 times."

Dopants

The materials chosen as suitable dopants depend on the atomic properties of both the dopant and the material to be doped. In general, dopants that produce the desired controlled changes are classified as either electron acceptors or donors. A donor atom that activates (that is, becomes incorporated into the crystal lattice) donates weakly bound valence electrons to the material, creating excess negative charge carriers. These weakly bound electrons can move about in the crystal lattice relatively freely and can facilitate conduction in the presence of an electric field. (The donor atoms introduce some states under, but very close to the conduction band edge. Electrons at these states can be easily excited to the conduction band, becoming free electrons, at room temperature.) Conversely, an activated acceptor produces a hole. Semiconductors doped with *donor* impurities are called *n-type*, while those doped with *acceptor* impurities are known as *p-type*. The n and p type designations indicate which charge carrier acts as the material's majority carrier. The opposite carrier is called the minority carrier, which exists due to thermal excitation at a much lower concentration compared to the majority carrier.

For example, the pure semiconductor silicon has four valence electrons. In silicon, the most common dopants are IUPAC group 13 (commonly known as *group III*) and group 15 (commonly known as *group V*) elements. Group 13 elements all contain three valence electrons, causing them to function as acceptors when used to dope silicon. Group 15 elements have five valence electrons, which allows them to act as a donor. Therefore, a silicon crystal doped with boron creates a p-type semiconductor whereas one doped with phosphorus results in an n-type material.

Carrier concentration

The concentration of dopant introduced to an intrinsic semiconductor determines its concentration and indirectly affects many of its electrical properties. The most important factor that doping directly affects is the material's carrier concentration. In an intrinsic semiconductor under thermal equilibrium, the concentration of electrons and holes is equivalent. That is,

$$n = p = n_i.$$

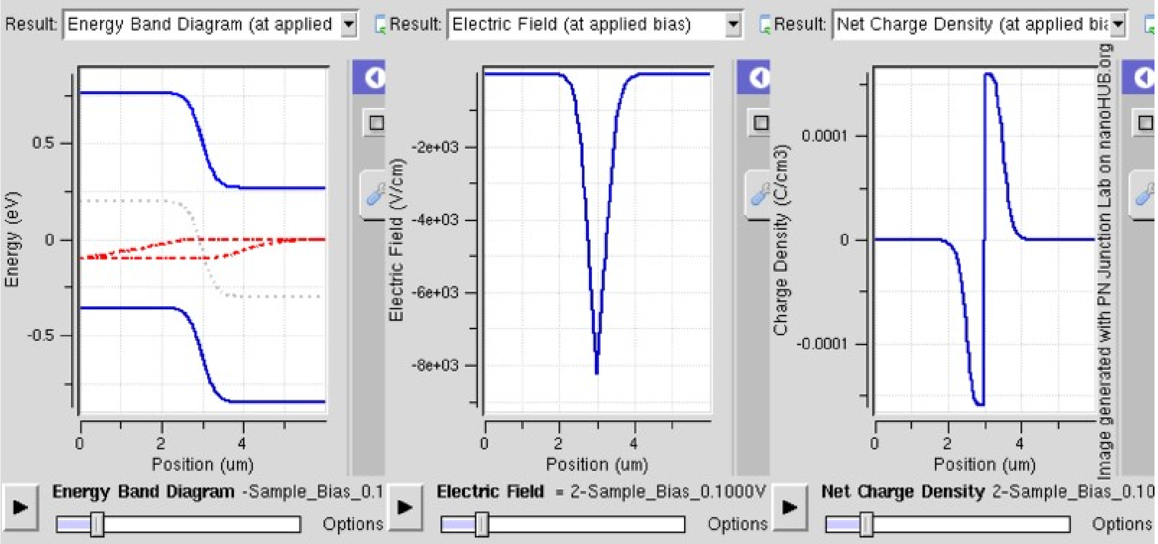
If we have a non-intrinsic semiconductor in thermal equilibrium the relation becomes:

$$n_0 \cdot p_0 = n_i^2$$

where n_0 is the concentration of conducting electrons, p_0 is the electron hole concentration, and n_i is the material's intrinsic carrier concentration. Intrinsic carrier concentration varies between materials and is dependent on temperature. Silicon's n_i, for example, is roughly 1.08×10^{10} cm^{-3} at 300 kelvins (room temperature).

In general, an increase in doping concentration affords an increase in conductivity due to the higher concentration of carriers available for conduction. Degenerately (very highly) doped semiconductors have conductivity levels comparable to metals and are often used in modern integrated circuits as a replacement for metal. Often superscript plus and minus symbols are used to denote relative doping concentration in semiconductors. For example, n^+ denotes an n-type semiconductor with a high, often degenerate, doping concentration. Similarly, p^- would indicate a very lightly doped p-type material. It is useful to note that even degenerate levels of doping imply low concentrations of impurities with respect to the base semiconductor. In crystalline intrinsic silicon, there are approximately 5×10^{22} atoms/cm³. Doping concentration for silicon semiconductors may range anywhere from 10^{13} cm^{-3} to 10^{18} cm^{-3}. Doping concentration above about 10^{18} cm^{-3} is considered degenerate at room temperature. Degenerately doped silicon contains a proportion of impurity to silicon on the order of parts per thousand. This proportion may be reduced to parts per billion in very lightly doped silicon. Typical concentration values fall somewhere in this range and are tailored to produce the desired properties in the device that the semiconductor is intended for.

Effect on band structure



Result: Energy Band Diagram (at applied

Result: Electric Field (at applied bias)

Result: Net Charge Density (at applied bia

Energy Band Diagram -Sample_Bias_0.1

Electric Field = 2-Sample_Bias_0.1000V

Net Charge Density 2-Sample_Bias_0.10

Band diagram of PN junction operation in forward bias mode showing reducing depletion width. Both p and n junctions are doped at a 1e15/cm3 doping level, leading to built-in potential of ~0.59V. Reducing depletion width can be inferred from the shrinking charge profile, as fewer dopants are exposed with increasing forward bias .

Doping a semiconductor crystal introduces allowed energy states within the band gap but very close to the energy band that corresponds to the dopant type. In other words, donor impurities create states near the conduction band while acceptors create states near the valence band. The gap between these energy states and the nearest energy band is usually referred to as dopant-site bonding energy or E_B and is relatively small. For example, the E_B for boron in silicon bulk is 0.045 eV, compared with silicon's band gap of about 1.12 eV. Because E_B is so small, it takes little energy to ionize the dopant atoms and create free carriers in the conduction or valence bands. Usually the thermal energy available at room temperature is sufficient to ionize most of the dopant.

Dopants also have the important effect of shifting the material's Fermi level towards the energy band that corresponds with the dopant with the greatest concentration. Since the Fermi level must remain constant in a system in thermodynamic equilibrium, stacking layers of materials with different properties leads to many useful electrical properties. For example, the p-n junction's properties are due to the energy band bending that happens as a result of lining up the Fermi levels in contacting regions of p-type and n-type material.

This effect is shown in a *band diagram*. The band diagram typically indicates the variation in the valence band and conduction band edges versus some spatial dimension, often denoted x. The Fermi energy is also usually indicated in the diagram. Sometimes the *intrinsic Fermi energy*, E_i, which is the Fermi level in the absence of doping, is shown. These diagrams are useful in explaining the operation of many kinds of semiconductor devices.

Preparation of semiconductor materials

Semiconductors with predictable, reliable electronic properties are necessary for mass production. The level of chemical purity needed is extremely high because the presence of impurities even in very small proportions can have large effects on the properties of the material. A high degree of crystalline perfection is also required, since faults in crystal structure (such as dislocations, twins, and stacking faults) interfere with the semiconducting properties of the material. Crystalline faults are a major cause of defective semiconductor devices. The larger the crystal, the more difficult it is to achieve the necessary perfection. Current mass production processes use crystal ingots between 100 mm and 300 mm (4–12 inches) in diameter which are grown as cylinders and sliced into wafers.

Because of the required level of chemical purity and the perfection of the crystal structure which are needed to make semiconductor devices, special methods have been developed to produce the initial semiconductor material. A technique for achieving high purity includes growing the crystal using the Czochralski process. An additional step that can be used to further increase purity is known as zone refining. In zone refining, part of a solid crystal is melted. The impurities tend to concentrate in the melted region, while the desired material recrystalizes leaving the solid material more pure and with fewer crystalline faults.

In manufacturing semiconductor devices involving heterojunctions between different semiconductor materials, the lattice constant, which is the length of the repeating element of the crystal structure, is important for determining the compatibility of materials.

Chapter 2

Electron Mobility

In solid-state physics, the **electron mobility** characterizes how quickly an electron can move through a metal or semiconductor, when pulled by an electric field. In semiconductors, there is an analogous quantity for holes, called **hole mobility**. The term **carrier mobility** refers in general to both electron and hole mobility in semiconductors.

Electron and hole mobility are special cases of electrical mobility of charged particles in a fluid under an applied electric field.

When an electric field *E* is applied across a piece of material, the electrons respond by moving with an average velocity called the drift velocity, v_d. Then the electron mobility μ is defined as

$$v_d = \mu E.$$

Electron mobility is almost always specified in units of $cm^2/(V \cdot s)$. This is different from the SI unit of mobility, $m^2/(V \cdot s)$. They are related by $1m^2/(V \cdot s) = 10^4 cm^2/(V \cdot s)$.

Conductivity is proportional to the product of mobility and carrier concentration. For example, the same conductivity could come from a small number of electrons with high mobility for each, or a large number of electrons with a small mobility for each. For metals, it would not typically matter which of these is the case. (Most metal electrical behavior in depends on conductivity alone.) Therefore mobility is relatively unimportant in metal physics. On the other hand, for semiconductors, the behavior of transistors and other devices can be very different depending on whether there are many electrons with low mobility or few electrons with high mobility. Therefore mobility is a very important parameter for semiconductor materials. Almost always, higher mobility leads to better device performance, other things equal.

Semiconductor mobility depends on the impurity concentrations (including donor and acceptor concentrations), defect concentration, temperature, and electron and hole concentrations. It also depends on the electric field, particularly at high fields when velocity saturation occurs. It can be determined by the Hall effect, or inferred from transistor behavior.

Introduction

Drift velocity in an electric field

Without any applied electric field, in a solid, electrons (or, in the case of semiconductors, both electrons and holes) move around randomly. Therefore, on average there will be no overall motion of charge carriers in any particular direction over time.

However, when an electric field is applied, each electron is accelerated by the electric field. If the electron were in a vacuum, it would be accelerated to faster and faster velocities (called ballistic transport). However, in a solid, the electron repeatedly scatters off crystal defects, phonons, impurities, etc. Therefore, it does not accelerate faster and faster; instead it moves with a finite average velocity, called the drift velocity. This net electron motion is usually much slower than the normally occurring random motion.

In a semiconductor the two charge carriers, electrons and holes, will typically have different drift velocities for the same electric field.

Quasi-ballistic transport is possible in solids if the electrons are accelerated across a very small distance (as small as the mean free path), or for a very short time (as short as the mean free time). In these cases, drift velocity and mobility are not meaningful.

Definition and units

The electron mobility is defined by the equation:

$$v_d = \mu E.$$

where:

> E is the magnitude of the electric field applied to a material,
> v_d is the magnitude of the electron drift velocity (in other words, the electron drift speed) caused by the electric field, and
> μ is the electron mobility.

The hole mobility is defined by the same equation. Both electron and hole mobilities are positive by definition.

Usually, the electron drift velocity in a material is directly proportional to the electric field, which means that the electron mobility is a constant (independent of electric field). When this is not true (for example, in very large electric fields), the mobility depends on the electric field.

The SI unit of velocity is m/s, and the SI unit of electric field is V/m. Therefore the SI unit of mobility is (m/s)/(V/m) = $m^2/(V·s)$. However, mobility is much more commonly expressed in $cm^2/(V·s) = 10^{-4}\ m^2/(V·s)$.

Mobility is usually a strong function of material impurities and temperature, and is determined empirically, mobility values are typically presented in table or chart form. Mobility is also different for electrons and holes in a given semiconductor.

Relation to conductivity

There is a simple relation between mobility and electrical conductivity. Let *n* be the number density of electrons, and let μ_e be their mobility. In the electric field **E**, each of these electrons will move with the velocity vector $-\mu_e \mathbf{E}$, for a total current density of $n e \mu_e \mathbf{E}$(where *e* is the elementary charge). Therefore, the electrical conductivity σ satisfies:

$$\sigma = n e \mu_e.$$

This formula is valid when the conductivity is due entirely to electrons. In a p-type semiconductor, the conductivity is due to holes instead, but the formula is essentially the same: If *p* is the density of holes and μ_h is the hole mobility, then the conductivity is

$$\sigma = p e \mu_h.$$

If a semiconductor has both electrons and holes, the total conductivity is

$$\sigma = e(n\mu_e + p\mu_h).$$

Examples

Typical electron mobility for Si at room temperature (300 K) is 1400 cm^2/ (V·s) and the hole mobility is around 450 cm^2/ (V·s).

Very high mobility has been found in several low-dimensional systems, such as two-dimensional electron gases (2DEG) (3,000,000 cm^2/(V·s) at low temperature), carbon nanotubes (100,000 cm^2/(V·s) at room temperature) and more recently, graphene (200,000 cm^2/ V·s at low temperature). Organic semiconductors (polymer, oligomer) developed thus far have carrier mobilities below 10 cm^2/(V·s), and usually much lower.

Electric field dependence and velocity saturation

At low fields, the drift velocity v_d is proportional to the electric field *E*, so mobility *μ* is constant. This value of *μ* is called the *low-field mobility*.

As the electric field is increased, however, the carrier velocity increases sub-linearly and asymptotically towards an maximum possible value, called the *saturation velocity* v_{sat}. For example, the value of v_{sat} is on the order of 1×10^7 cm/s for both electrons and holes in Si. It is on the order of 6×10^6 cm/s for Ge. This velocity is a characteristic of the material and a strong function of doping or impurity levels and temperature. It is one of the key material and semiconductor device properties that determine a device such as a transistor's ultimate limit of speed of response and frequency.

This velocity saturation phenomenon results from a process called *optical phonon scattering*. At high fields, carriers are accelerated enough to gain sufficient kinetic energy between collisions to emit an optical phonon, and they do so very quickly, before being accelerated once again. The velocity that the electron reaches before emitting a phonon is:

$$\frac{m^* v_{emit}^2}{2} \approx \hbar\omega_{phonon(opt.)}$$

where $\omega_{phonon(opt.)}$ is the optical phonon angular frequency and m*is the carrier effective mass in the direction of the electric field. The value of $E_{\text{phonon (opt.)}}$ is 0.063 eV for Si and 0.034 eV for GaAs and Ge. The saturation velocity is only one-half of v_{emit}, because the electron starts at zero velocity and accelerates up to v_{emit} in each cycle. (This is a somewhat oversimplified description.)

Velocity saturation is not the only possible high-field behavior. Another is the Gunn effect, where a sufficiently high electric field can causes intervalley electron transfer, which reduces drift velocity. (This is unusual; increasing the electric field almost always *increases* the drift velocity, or else leaves it unchanged.) The result is negative differential resistance.

In the regime of velocity saturation (or other high-field effects), mobility is a strong function of electric field. This means that mobility is a somewhat less useful concept, compared to simply discussing drift velocity directly.

Relation between scattering and mobility

Recall that by definition, mobility is dependent on the drift velocity. The main factor determining drift velocity (other than effective mass) is scattering time, i.e., how long the carrier is ballistically accelerated by the electric field until it scatters (collides) with something that changes its direction and/or energy. The most important sources of scattering in typical semiconductor materials, discussed below, are ionized impurity scattering and acoustic phonon scattering (also called lattice scattering). In some cases other sources of scattering may be important, such as neutral impurity scattering, optical phonon scattering, surface scattering, and defect scattering.

Ionized impurity scattering

Semiconductors are doped with donors and/or acceptors, which are typically ionized, and are thus charged. The Coulombic forces will deflect an electron or hole approaching the ionized impurity. This is known as *ionized impurity scattering*. The amount of deflection depends on the speed of the carrier and its proximity to the ion. The more heavily a material is doped, the higher the probability that a carrier will collide with an ion in a given time, and the smaller the mean free time between collisions, and the smaller the mobility.

Lattice (phonon) scattering

At any temperature above absolute zero, the vibrating atoms create pressure (acoustic) waves in the crystal, which are termed phonons. Like electrons, phonons can be considered to be particles. A phonon can interact (collide) with an electron (or hole) and scatter it. At higher temperature, there are more phonons, therefore increased phonon scattering which tends to reduce mobility.

Relation between mobility and scattering time

A simple model gives the approximate relation between scattering time (average time between scattering events) and mobility. It is assumed that after each scattering event, the carrier's motion is randomized, so it has zero average velocity. After that, it accelerates uniformly in the electric field, until it scatters again. The resulting average drift velocity is:

$$\mu = \frac{q}{m^*}\overline{\tau}$$

where *q* is the elementary charge, m* is the carrier effective mass, and τ is the average scattering time.

If the effective mass is anisotropic (direction-dependent), m* is the effective mass in the direction of the electric field.

Matthiessen's rule

Normally, more than one source of scattering is present, for example both impurities and lattice phonons. It is normally a very good approximation to combine their influences using "Matthiessen's Rule" (developed from work by Augustus Matthiessen in 1864):

$$\frac{1}{\mu} = \frac{1}{\mu_{\mathrm{impurities}}} + \frac{1}{\mu_{\mathrm{lattice}}}.$$

where μ is the actual mobility, $\mu_{\mathrm{impurities}}$ is the mobility that the material would have if there was impurity scattering but no other source of scattering, and μ_{lattice} is the mobility that the material would have if there was lattice phonon scattering but no other source of scattering. Other terms may be added for other scattering sources, for example

$$\frac{1}{\mu} = \frac{1}{\mu_{\mathrm{impurities}}} + \frac{1}{\mu_{\mathrm{lattice}}} + \frac{1}{\mu_{\mathrm{defects}}} + \cdots.$$

Matthiessen's rule can also be stated in terms of the scattering time:

$$\frac{1}{\tau} = \frac{1}{\tau_{\text{impurities}}} + \frac{1}{\tau_{\text{lattice}}} + \frac{1}{\tau_{\text{defects}}} + \cdots .$$

where τ is the true average scattering time and $\tau_{\text{impurities}}$ is the scattering time if there was impurity scattering but no other source of scattering, etc.

Matthiessen's rule is an approximation and is not universally valid. For example, lattice scattering alters the average electron velocity (in the electric-field direction), which in turn alters the tendency to scatter off impurities. There are more complicated formulas that attempt to take these effects into account.

Temperature dependence of mobility

Typical temperature dependence of mobility
(depends on the temperature range and sample)

	Si	Ge	GaAs
Electrons	$\propto T^{-2.4}$	$\propto T^{-1.7}$	$\propto T^{-1.0}$
Holes	$\propto T^{-2.2}$	$\propto T^{-2.3}$	$\propto T^{-2.1}$

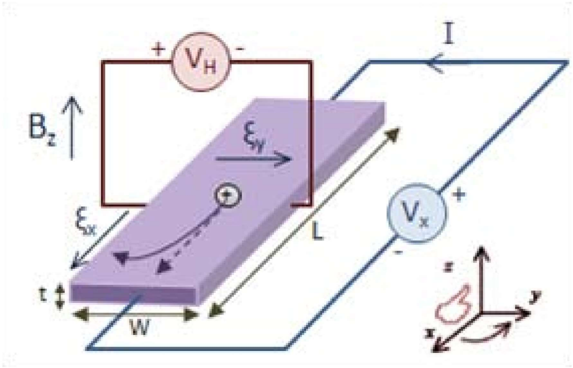
With increasing temperature, phonon concentration increases and causes increased scattering. Thus lattice scattering lowers the carrier mobility more and more at higher temperature. Theoretical calculations reveal that the mobility in non-polar semiconductors, such as silicon and germanium, is dominated by acoustic phonon interaction. The resulting mobility is expected to be proportional to $T^{-3/2}$, while the mobility due to optical phonon scattering only is expected to be proportional to $T^{-1/2}$. Experimentally, values of the temperature dependence of the mobility in Si, Ge and GaAs are listed in table.

The effect of ionized impurity scattering, however, *decreases* with increasing temperature because the average thermal speeds of the carriers are increase. Thus, the carriers spend less time near an ionized impurity as they pass and the scattering effect of the ions is thus reduced.

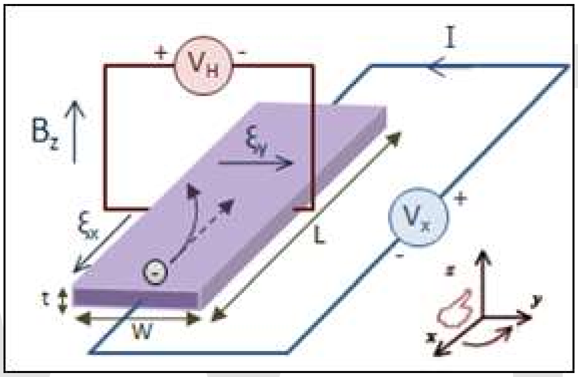
These two effects operate simultaneously on the carriers through Matthiessen's rule. At lower temperatures, ionized impurity scattering dominates, while at higher temperatures, phonon scattering dominates, and the actual mobility reaches a maximum at an intermediate temperature.

Measurement of semiconductor mobility

Hall mobility



Hall Effect measurement setup for holes



Hall Effect measurement setup for electrons

Carrier mobility is most commonly measured using the Hall effect. The result of the measurement is called the "Hall mobility" (meaning "mobility inferred from a Hall-effect measurement").

Consider a semiconductor sample with a rectangular cross section as shown in the figures, a current is flowing in the *x*-direction and a magnetic field is applied in the *z*-direction. The resulted Lorenz's force will accelerates the electrons (*n*-type materials) or holes (*p*-type materials) in the (−*y*) direction, according to the right hand rule and set up an electric field ξ_y. As a result there is a voltage across the sample, which can be measured with a high-impedance voltmeter. This voltage, V_H, is called the Hall voltage. V_H is positive for *n*-type material and negative for *p*-type material.

Mathematically, the Lorentz force acting on a charge *Q* is given by

For electrons:

$$\overrightarrow{F}_{Hn} = -q(\overrightarrow{v}_n \times \overrightarrow{B}_z)$$

For holes:

$$\overrightarrow{F}_{Hp} = +q(\overrightarrow{v}_p \times \overrightarrow{B}_z)$$

In steady state this force is balanced by the force set up by the Hall voltage, so that there is no net force on the carriers in the *y* direction. For electron,

$$\overrightarrow{F}_y = (-q)\overrightarrow{\xi}_y + (-q)[\overrightarrow{v}_n \times \overrightarrow{B}_z] = 0$$
$$\Rightarrow -q\xi_y + qv_x B_z = 0$$
$$\xi_y = v_x B_z$$

For electrons, the field points in the +*y* direction, and for holes, it points in the −*y* direction.

The electron current *I* is given by $I = -\,qnv_xtW$. Sub v_x into the expression for ξ_y,

$$\xi_y = -\frac{IB}{nqtW} = +\frac{R_{Hn}IB}{tW}$$

where R_{Hn} is the Hall coefficient for electron, and is defined as

$$R_{Hn} = -\frac{1}{nq}$$

Since $\xi_y = \frac{V_H}{W}$

$$R_{Hn} = -\frac{1}{nq} = \frac{V_{Hn}t}{IB}$$

Similarly, for holes

$$R_{Hp} = \frac{1}{pq} = \frac{V_{Hp}t}{IB}$$

From Hall coefficient, we can obtain the carrier mobility as follows:

$$\mu_n = (-nq)\mu_n(-\frac{1}{nq}) = -\sigma_n R_{Hn}$$
$$= -\frac{\sigma_n V_{Hn}t}{IB}$$

Similarly,

$$\mu_p = \frac{\sigma_p V_{Hp} t}{IB}$$

Here the value of *V_{Hp} (Hall voltage), t (sample thickness), I (current) and B (magnetic field)* can be measured directly, and the conductivities σ_n or σ_p are either known or can be obtained from measuring the resistivity.

Field-effect mobility

The mobility can also be measured using a field-effect transistor (FET). The result of the measurement is called the "field-effect mobility" (meaning "mobility inferred from a field-effect measurement").

The measurement can work in two ways: From saturation-mode measurements, or linear-region measurements.

Using saturation mode

In this technique, for each fixed gate voltage V_{GS}, the drain-source voltage V_{DS} is increased until the current I_D saturates. Next, the square root of this saturated current is plotted against the gate voltage, and the slope m_{sat} is measured. Then the mobility is:

$$\mu = m_{sat}^2 \frac{2L}{W} \frac{1}{C_i}$$

where *L* and *W* are the length and width of the channel and C_i is the gate insulator capacitance per unit area. This equation comes from the approximate equation for a MOSFET in saturation mode:

$$I_D = \frac{\mu C_i}{2} \frac{W}{L} (V_{GS} - V_{th})^2.$$

where V_{th} is the threshold voltage. This approximation ignores the Early effect (channel length modulation), among other things. In practice, this technique may underestimate the true mobility.

Using the linear region

In this technique, the transistor is operated in the linear region (or "ohmic mode"), where V_{DS} is small and $I_D \propto V_{GS}$with slope m_{lin}. Then the mobility is:

$$\mu = m_{lin} \frac{L}{W} \frac{1}{V_{DS}} \frac{1}{C_i}.$$

This equation comes from the approximate equation for a MOSFET in the linear region:

$$I_D = \mu C_i \frac{W}{L}\left((V_{GS} - V_{th})V_{DS} - \frac{V_{DS}^2}{2}\right)$$

In practice, this technique may overestimate the true mobility, because if V_{DS} is not small enough and V_G is not large enough, the MOSFET may not stay in the linear region.

Doping concentration dependence in heavily-doped silicon

The charge carriers in semiconductors are electrons and holes. Their numbers are controlled by the concentrations of impurity elements, i.e. doping concentration. Thus doping concentration has great influence on carrier mobility.

While there is considerable scatter in the experimental data, for noncompensated material (no counter doping) for heavily doped substrates (i.e. $10^{18} cm^{-3}$ and up), the mobility in silicon is often characterized by the empirical relationship:

$$\mu = \mu_o + \frac{\mu_1}{1 + (\frac{N}{N_{\mathrm{ref}}})^{\alpha}}$$

where N is the doping concentration (either N_D or N_A), and N_{ref} and α are fitting parameters. At room temperature, the above equation becomes: Majority carriers:

$$\mu_n(N_D) = 65 + \frac{1265}{1 + (\frac{N_D}{8.5\times 10^{16}})^{0.72}}$$
$$\mu_p(N_A) = 48 + \frac{447}{1 + (\frac{N_A}{6.3\times 10^{16}})^{0.76}}$$

Minority carriers:

$$\mu_n(N_A) = 232 + \frac{1180}{1 + (\frac{N_A}{8\times 10^{16}})^{0.9}}$$
$$\mu_p(N_D) = 130 + \frac{370}{1 + (\frac{N_D}{8\times 10^{17}})^{1.25}}$$

These equations apply only to silicon, and only under low field.

Chapter 3

Semiconductor Device

Semiconductor devices are electronic components that exploit the electronic properties of semiconductor materials, principally silicon, germanium, and gallium arsenide, as well as organic semiconductors. Semiconductor devices have replaced thermionic devices (vacuum tubes) in most applications. They use electronic conduction in the solid state as opposed to the gaseous state or thermionic emission in a high vacuum.

Semiconductor devices are manufactured both as single discrete devices and as *integrated circuits* (ICs), which consist of a number—from a few (as low as two) to billions—of devices manufactured and interconnected on a single semiconductor substrate.

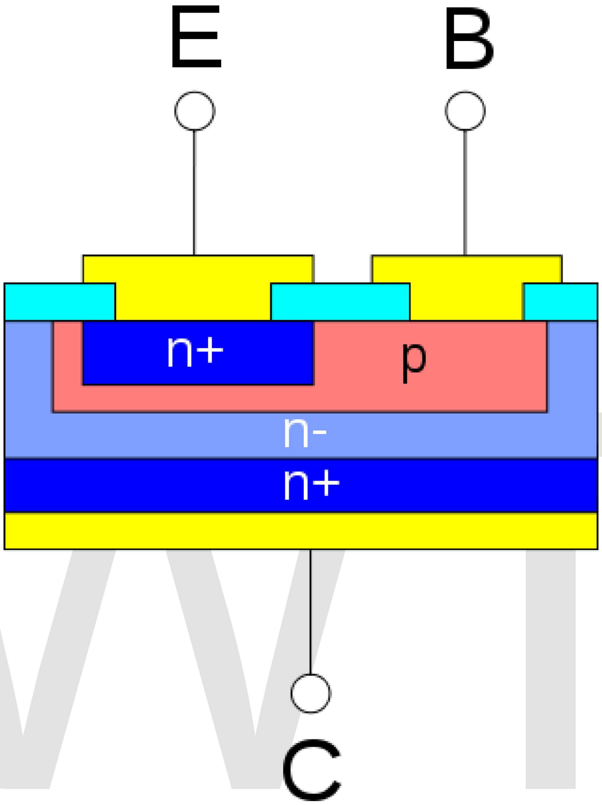
The main reason why semiconductor materials are so useful is that the behavior of a semiconductor can be easily manipulated by the addition of impurities, known as doping. Semiconductor conductivity can be controlled by introduction of an electric field, by exposure to light, and even pressure and heat; thus, semiconductors can make excellent sensors. Current conduction in a semiconductor occurs via mobile or "free" *electrons* and *holes*, collectively known as *charge carriers*. Doping a semiconductor such as silicon with a small amount of impurity atoms, such as phosphorus or boron, greatly increases the number of free electrons or holes within the semiconductor. When a doped semiconductor contains excess holes it is called "p-type", and when it contains excess free electrons it is known as "n-type", where *p* (positive for holes) or *n* (negative for electrons) is the sign of the charge of the majority mobile charge carriers. The semiconductor material used in devices is doped under highly controlled conditions in a fabrication facility, or *fab*, to precisely control the location and concentration of p- and n-type dopants. The junctions which form where n-type and p-type semiconductors join together are called p-n junctions.

Diode

The diode is a device made from a single p-n junction. At the junction of a p-type and an n-type semiconductor there forms a region called the depletion zone which blocks current conduction from the n-type region to the p-type region, but allows current to conduct from the p-type region to the n-type region. Thus when the device is *forward biased*, with the p-side at higher electric potential, the diode conducts current easily; but the current is very small when the diode is *reverse biased*.

Exposing a semiconductor to light can generate electron–hole pairs, which increases the number of free carriers and its conductivity. Diodes optimized to take advantage of this phenomenon are known as *photodiodes*. Compound semiconductor diodes can also be used to generate light, as in light-emitting diodes and laser diodes.

Transistor



An NPN bipolar junction transistor structure

Bipolar junction transistors are formed from two p-n junctions, in either n-p-n or p-n-p configuration. The middle, or *base*, region between the junctions is typically very narrow. The other regions, and their associated terminals, are known as the *emitter* and the *collector*. A small current injected through the junction between the base and the emitter changes the properties of the base-collector junction so that it can conduct current even though it is reverse biased. This creates a much larger current between the collector and emitter, controlled by the base-emitter current.

Another type of transistor, the field effect transistor operates on the principle that semiconductor conductivity can be increased or decreased by the presence of an electric field. An electric field can increase the number of free electrons and holes in a semiconductor, thereby changing its conductivity. The field may be applied by a reverse-biased p-n junction, forming a *junction field effect transistor*, or JFET; or by an electrode isolated from the bulk material by an oxide layer, forming a *metal-oxide-semiconductor field effect transistor*, or MOSFET.

The MOSFET is the most used semiconductor device today. The *gate* electrode is charged to produce an electric field that controls the conductivity of a "channel" between two terminals, called the *source* and *drain*. Depending on the type of carrier in the channel, the device may be an *n-channel* (for electrons) or a *p-channel* (for holes) MOSFET. Although the MOSFET is named in part for its "metal" gate, in modern devices polysilicon is typically used instead. MOSFET is an IC which is semiconductor device.

Semiconductor device materials

By far, silicon (Si) is the most widely used material in semiconductor devices. Its combination of low raw material cost, relatively simple processing, and a useful temperature range make it currently the best compromise among the various competing materials. Silicon used in semiconductor device manufacturing is currently fabricated into boules that are large enough in diameter to allow the production of 300 mm (12 in.) wafers.

Germanium (Ge) was a widely used early semiconductor material but its thermal sensitivity makes it less useful than silicon. Today, germanium is often alloyed with silicon for use in very-high-speed SiGe devices; IBM is a major producer of such devices.

Gallium arsenide (GaAs) is also widely used in high-speed devices but so far, it has been difficult to form large-diameter boules of this material, limiting the wafer diameter to sizes significantly smaller than silicon wafers thus making mass production of GaAs devices significantly more expensive than silicon.

Other less common materials are also in use or under investigation.

Silicon carbide (SiC) has found some application as the raw material for blue light-emitting diodes (LEDs) and is being investigated for use in semiconductor devices that could withstand very high operating temperatures and environments with the presence of significant levels of ionizing radiation. IMPATT diodes have also been fabricated from SiC.

Various indium compounds (indium arsenide, indium antimonide, and indium phosphide) are also being used in LEDs and solid state laser diodes. Selenium sulfide is being studied in the manufacture of photovoltaic solar cells.

The most common use for organic semiconductors is Organic light-emitting diodes.

List of common semiconductor devices

Two-terminal devices:

- DIAC

- Diode (rectifier diode)
- Gunn diode
- IMPATT diode
- Laser diode
- Light-emitting diode (LED)
- Photocell
- PIN diode
- Schottky diode
- Solar cell
- Tunnel diode
- VCSEL
- VECSEL
- Zener diode

Three-terminal devices:

- Bipolar transistor
- Darlington transistor
- Field effect transistor
- GTO (Gate Turn-Off)
- IGBT (Insulated Gate Bipolar Transistor)
- SCR (Silicon Controlled Rectifier)
- SGCT (Switched Gate Commuted Thyristor)
- Thyristor
- TRIAC
- Unijunction transistor

Four-terminal devices:

- Hall effect sensor (magnetic field sensor)

Multi-terminal devices:

- Integrated Circuit (ICs)
- Charge-coupled device (CCD)
- Microprocessor
- Random Access Memory (RAM)
- Read-only memory (ROM)

Semiconductor device applications

All transistor types can be used as the building blocks of logic gates, which are fundamental in the design of digital circuits. In digital circuits like microprocessors, transistors act as on-off switches; in the MOSFET, for instance, the voltage applied to the gate determines whether the switch is on or off.

Transistors used for analog circuits do not act as on-off switches; rather, they respond to a continuous range of inputs with a continuous range of outputs. Common analog circuits include amplifiers and oscillators.

Circuits that interface or translate between digital circuits and analog circuits are known as mixed-signal circuits.

Power semiconductor devices are discrete devices or integrated circuits intended for high current or high voltage applications. Power integrated circuits combine IC technology with power semiconductor technology, these are sometimes referred to as "smart" power devices. Several companies specialize in manufacturing power semiconductors.

Component identifiers

The type designators of semiconductor devices are often manufacturer specific. Nevertheless, there have been attempts at creating standards for type codes, and a subset of devices follow those. For discrete devices, for example, there are three standards: JEDEC JESD370B in USA, Pro Electron in Europe and JIS in Japan.

History of semiconductor device development

Cat's-whisker detector

Semiconductors had been used in the electronics field for some time before the invention of the transistor. Around the turn of the 20th century they were quite common as detectors in radios, used in a device called a "cat's whisker". These detectors were somewhat troublesome, however, requiring the operator to move a small tungsten filament (the whisker) around the surface of a galena (lead sulfide) or carborundum (silicon carbide) crystal until it suddenly started working. Then, over a period of a few hours or days, the cat's whisker would slowly stop working and the process would have to be repeated. At the time their operation was completely mysterious. After the introduction of the more reliable and amplified vacuum tube based radios, the cat's whisker systems quickly disappeared. The "cat's whisker" is a primitive example of a special type of diode still popular today, called a Schottky diode.

Metal rectifier

Another early type of semiconductor device is the metal rectifier in which the semiconductor is copper oxide or selenium. Westinghouse Electric (1886) was a major manufacturer of these rectifiers.

World War II

During World War II, radar research quickly pushed radar receivers to operate at ever higher frequencies and the traditional tube based radio receivers no longer worked well.

The introduction of the cavity magnetron from Britain to the United States in 1940 during the Tizard Mission resulted in a pressing need for a practical high-frequency amplifier.

On a whim, Russell Ohl of Bell Laboratories decided to try a cat's whisker. By this point they had not been in use for a number of years, and no one at the labs had one. After hunting one down at a used radio store in Manhattan, he found that it worked much better than tube-based systems.

Ohl investigated why the cat's whisker functioned so well. He spent most of 1939 trying to grow more pure versions of the crystals. He soon found that with higher quality crystals their finicky behaviour went away, but so did their ability to operate as a radio detector. One day he found one of his purest crystals nevertheless worked well, and interestingly, it had a clearly visible crack near the middle. However as he moved about the room trying to test it, the detector would mysteriously work, and then stop again. After some study he found that the behaviour was controlled by the light in the room–more light caused more conductance in the crystal. He invited several other people to see this crystal, and Walter Brattain immediately realized there was some sort of junction at the crack.

Further research cleared up the remaining mystery. The crystal had cracked because either side contained very slightly different amounts of the impurities Ohl could not remove–about 0.2%. One side of the crystal had impurities that added extra electrons (the carriers of electrical current) and made it a "conductor". The other had impurities that wanted to bind to these electrons, making it (what he called) an "insulator". Because the two parts of the crystal were in contact with each other, the electrons could be pushed out of the conductive side which had extra electrons (soon to be known as the *emitter*) and replaced by new ones being provided (from a battery, for instance) where they would flow into the insulating portion and be collected by the whisker filament (named the *collector*). However, when the voltage was reversed the electrons being pushed into the collector would quickly fill up the "holes" (the electron-needy impurities), and conduction would stop almost instantly. This junction of the two crystals (or parts of one crystal) created a solid-state diode, and the concept soon became known as semiconduction. The mechanism of action when the diode is off has to do with the separation of charge carriers around the junction. This is called a "depletion region".

Development of the diode

Armed with the knowledge of how these new diodes worked, a vigorous effort began to learn how to build them on demand. Teams at Purdue University, Bell Labs, MIT, and the University of Chicago all joined forces to build better crystals. Within a year germanium production had been perfected to the point where military-grade diodes were being used in most radar sets.

Development of the transistor

After the war, William Shockley decided to attempt the building of a triode-like semiconductor device. He secured funding and lab space, and went to work on the problem with Brattain and John Bardeen.

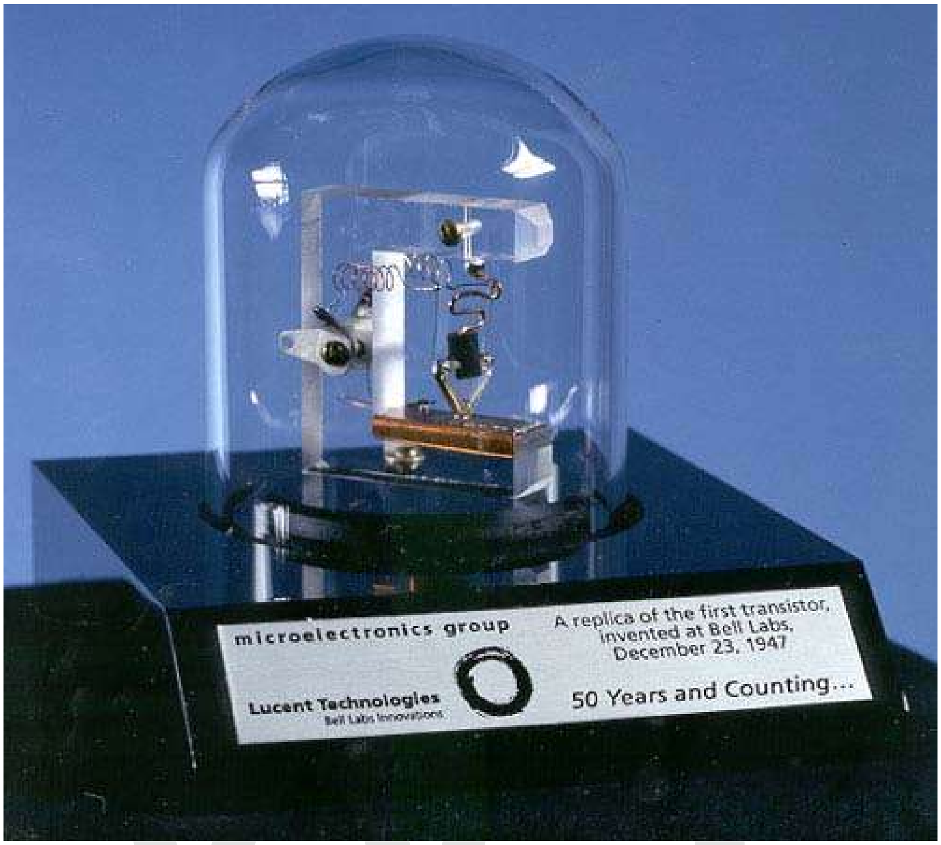
The key to the development of the transistor was the further understanding of the process of the electron mobility in a semiconductor. It was realized that if there was some way to control the flow of the electrons from the emitter to the collector of this newly discovered diode, one could build an amplifier. For instance, if you placed contacts on either side of a single type of crystal the current would not flow through it. However if a third contact could then "inject" electrons or holes into the material, the current would flow.

Actually doing this appeared to be very difficult. If the crystal were of any reasonable size, the number of electrons (or holes) required to be injected would have to be very large -– making it less than useful as an amplifier because it would require a large injection current to start with. That said, the whole idea of the crystal diode was that the crystal itself could provide the electrons over a very small distance, the depletion region. The key appeared to be to place the input and output contacts very close together on the surface of the crystal on either side of this region.

Brattain started working on building such a device, and tantalizing hints of amplification continued to appear as the team worked on the problem. Sometimes the system would work but then stop working unexpectedly. In one instance a non-working system started working when placed in water. Ohl and Brattain eventually developed a new branch of quantum mechanics known as surface physics to account for the behaviour. The electrons in any one piece of the crystal would migrate about due to nearby charges. Electrons in the emitters, or the "holes" in the collectors, would cluster at the surface of the crystal where they could find their opposite charge "floating around" in the air (or water). Yet they could be pushed away from the surface with the application of a small amount of charge from any other location on the crystal. Instead of needing a large supply of injected electrons, a very small number in the right place on the crystal would accomplish the same thing.

Their understanding solved the problem of needing a very small control area to some degree. Instead of needing two separate semiconductors connected by a common, but tiny, region, a single larger surface would serve. The emitter and collector leads would both be placed very close together on the top, with the control lead placed on the base of the crystal. When current was applied to the "base" lead, the electrons or holes would be pushed out, across the block of semiconductor, and collect on the far surface. As long as the emitter and collector were very close together, this should allow enough electrons or holes between them to allow conduction to start.

The first transistor



A stylized replica of the first transistor

The Bell team made many attempts to build such a system with various tools, but generally failed. Setups where the contacts were close enough were invariably as fragile as the original cat's whisker detectors had been, and would work briefly, if at all. Eventually they had a practical breakthrough. A piece of gold foil was glued to the edge of a plastic wedge, and then the foil was sliced with a razor at the tip of the triangle. The result was two very closely spaced contacts of gold. When the plastic was pushed down onto the surface of a crystal and voltage applied to the other side (on the base of the crystal), current started to flow from one contact to the other as the base voltage pushed the electrons away from the base towards the other side near the contacts. The point-contact transistor had been invented.

While the device was constructed a week earlier, Brattain's notes describe the first demonstration to higher-ups at Bell Labs on the afternoon of 23 December 1947, often given as the birthdate of the transistor. The "PNP point-contact germanium transistor" operated as a speech amplifier with a power gain of 18 in that trial. Known generally as a point-contact transistor today, John Bardeen, Walter Houser Brattain, and William Bradford Shockley were awarded the Nobel Prize in physics for their work in 1956.

Origin of the term "transistor"

Bell Telephone Laboratories needed a generic name for their new invention: "Semiconductor Triode", "Solid Triode", "Surface States Triode" [sic], "Crystal Triode" and "Iotatron" were all considered, but "transistor", coined by John R. Pierce, won an internal ballot. The rationale for the name is described in the following extract from the company's Technical Memoranda (May 28, 1948) [26] calling for votes:

Transistor. This is an abbreviated combination of the words "transconductance" or "transfer", and "varistor". The device logically belongs in the varistor family, and has the transconductance or transfer impedance of a device having gain, so that this combination is descriptive.

Improvements in transistor design

Shockley was upset about the device being credited to Brattain and Bardeen, who he felt had built it "behind his back" to take the glory. Matters became worse when Bell Labs lawyers found that some of Shockley's own writings on the transistor were close enough to those of an earlier 1925 patent by Julius Edgar Lilienfeld that they thought it best that his name be left off the patent application.

Shockley was incensed, and decided to demonstrate who was the real brains of the operation. Only a few months later he invented an entirely new type of transistor with a layer or 'sandwich' structure. This new form was considerably more robust than the fragile point-contact system, and would go on to be used for the vast majority of all transistors into the 1960s. It would evolve into the bipolar junction transistor.

With the fragility problems solved, a remaining problem was purity. Making germanium of the required purity was proving to be a serious problem, and limited the number of transistors that actually worked from a given batch of material. Germanium's sensitivity to temperature also limited its usefulness. Scientists theorized that silicon would be easier to fabricate, but few bothered to investigate this possibility. Gordon K. Teal was the first to develop a working silicon transistor, and his company, the nascent Texas Instruments, profited from its technological edge. Germanium disappeared from most transistors by the late 1960s.

Within a few years, transistor-based products, most notably radios, were appearing on the market. A major improvement in manufacturing yield came when a chemist advised the companies fabricating semiconductors to use distilled water rather than tap water: calcium ions were the cause of the poor yields. "Zone melting", a technique using a moving band of molten material through the crystal, further increased the purity of the available crystals.

Chapter 4

Electronic Band Structure

In solid-state physics, the **electronic band structure** (or simply **band structure**) of a solid describes those ranges of energy an electron is "forbidden" or "allowed" to have. Band structure derives from the diffraction of the quantum mechanical electron waves in a periodic crystal lattice with a specific crystal system and Bravais lattice. The band structure of a material determines several characteristics, in particular the material's electronic and optical properties.

Why bands occur in materials

The electrons of a single isolated atom occupy atomic orbitals, which form a discrete set of energy levels. If several atoms are brought together into a molecule, their atomic orbitals split, as in a coupled oscillation. This produces a number of molecular orbitals proportional to the number of atoms. When a large number of atoms (of order $\times 10^{20}$ or more) are brought together to form a solid, the number of orbitals becomes exceedingly large. Consequently, the difference in energy between them becomes very small. Thus, in solids the levels form continuous *bands* of energy rather than the discrete energy levels of the atoms in isolation. However, some intervals of energy contain no orbitals, no matter how many atoms are aggregated, forming *band gaps*.

Within an energy band, energy levels form a near continuum. First, the separation between energy levels in a solid is comparable with the energy that electrons constantly exchange with phonons (atomic vibrations). Second, it is comparable with the energy uncertainty due to the Heisenberg uncertainty principle, for reasonably long intervals of time. As a result, the separation between energy levels is of no consequence.

Several approaches to finding band structure are discussed below.

Basic concepts

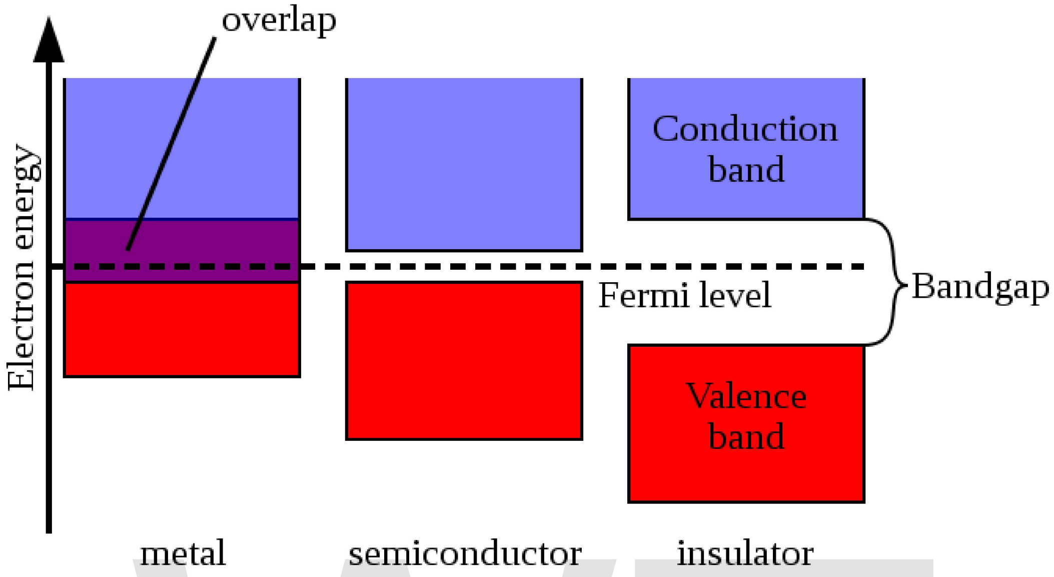


Figure 1: Simplified diagram of the electronic band structure of metals, semiconductors, and insulators.

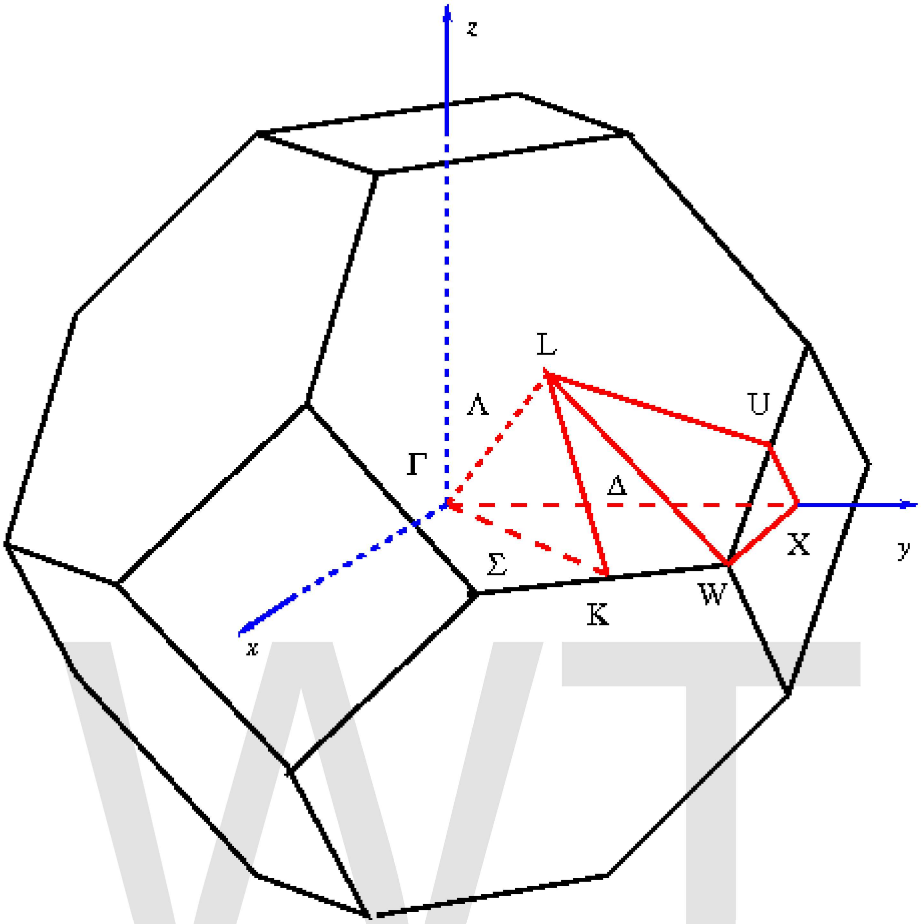


Figure 2: First Brillouin zone of FCC lattice showing symmetry labels

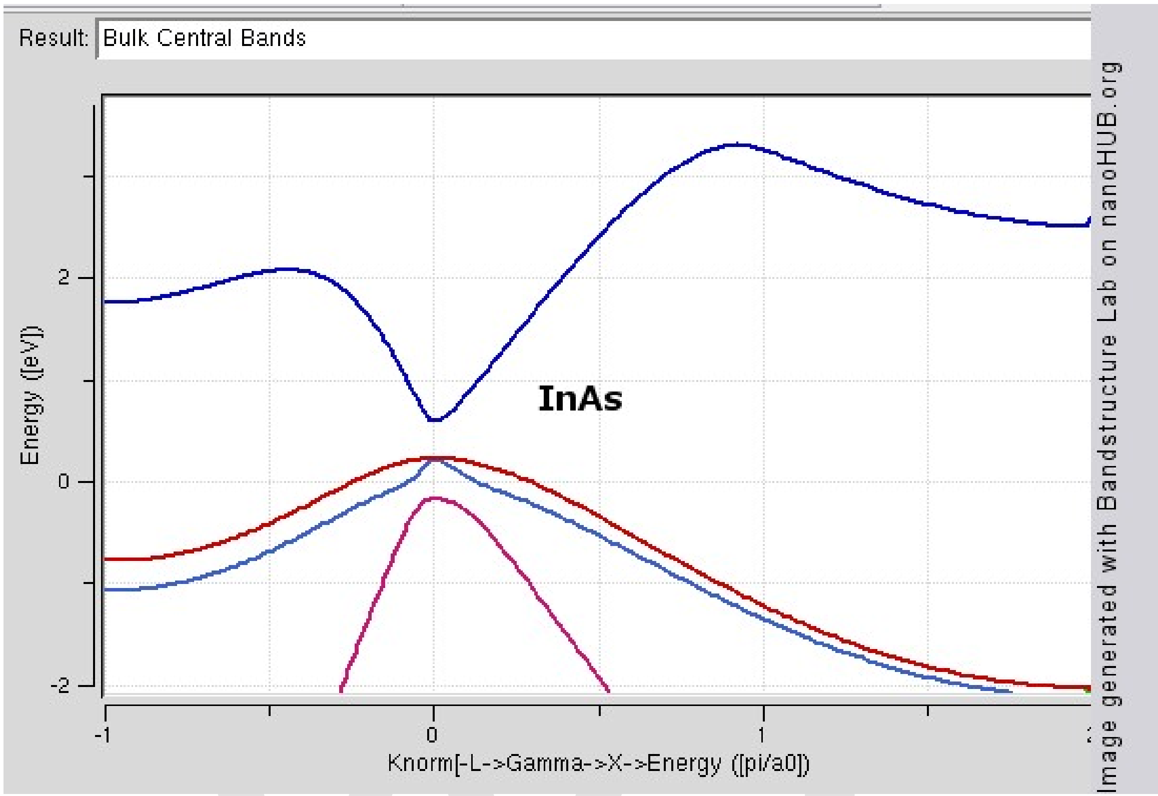


Figure 3: Bulk band structure for Si,Ge,GaAs and InAs generated with tight binding model. Note that Si and Ge are indirect while GaAs and InAs are direct band gap materials.

Any solid has a large number of bands. In theory, a solid can have infinitely many bands (just as an atom has infinitely many energy levels). However, all but a few of these bands lie at energies so high that any electron that attains those energies will escape from the solid. These bands are usually disregarded.

Bands have different widths, based upon the properties of the atomic orbitals from which they arise. Also, allowed bands may overlap, producing (for practical purposes) a single large band.

Figure 1 shows a simplified picture of the bands in a solid that allows the three major types of materials to be identified: metals, semiconductors and insulators.

Metals contain a band that is partly empty and partly filled regardless of temperature. Therefore they have very high conductivity.

The lowermost, almost fully occupied band in an *insulator* or *semiconductor*, is called the *valence band* by analogy with the valence electrons of individual atoms. The uppermost, almost unoccupied band is called the *conduction band* because only when electrons are excited to the conduction band can current flow in these materials. The difference between insulators and semiconductors is only that the forbidden band gap between the valence band and conduction band is larger in an insulator, so that fewer electrons are found there and the electrical conductivity is lower. Because one of the main mechanisms for electrons to be excited to the conduction band is due to thermal energy, the conductivity of semiconductors is strongly dependent on the temperature of the material.

This band gap is one of the most useful aspects of the band structure, as it strongly influences the electrical and optical properties of the material. Electrons can transfer from one band to the other by means of carrier generation and recombination processes. The band gap and defect states created in the band gap by doping can be used to create semiconductor devices such as solar cells, diodes, transistors, laser diodes, and others.

Symmetry

A more complete view of the band structure takes into account the periodic nature of a crystal lattice using the symmetry operations that form a space group. The Schrödinger equation is solved for the crystal, which has Bloch waves as solutions:

$$\psi_{n\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}),$$

where **k** is called the wavevector, and is related to the direction of motion of the electron in the crystal, and *n* is the band index, which simply numbers the energy bands. The wavevector **k** takes on values within the Brillouin zone (BZ) corresponding to the crystal lattice, and particular directions/points in the BZ are assigned conventional names like Γ, Δ, Λ, Σ, *etc.* These directions are shown for the face-centered cubic lattice geometry in Figure 2.

The available energies for the electron also depend upon **k**, as shown in Figure 3 for silicon in the more complex energy band diagram at the right. In this diagram the topmost energy of the valence band is labeled E_v and the bottom energy in the conduction band is labeled E_c. The top of the valence band is not directly below the bottom of the conduction band (E_v is for an electron traveling in direction Γ, E_c in direction X), so silicon is called an **indirect gap** material. For an electron to be excited from the valence band to the conduction band, it needs something to give it energy $E_c – E_v$ *and* a change in direction/momentum. In other semiconductors (for example GaAs) both are at Γ, and these materials are called **direct gap** materials (no momentum change required). Direct gap materials benefit the operation of semiconductor laser diodes.

Anderson's rule is used to align band diagrams between two different semiconductors in contact.

Band structures in different types of solids

Although electronic band structures are usually associated with crystalline materials, quasi-crystalline and amorphous solids may also exhibit band structures. However, the periodic nature and symmetrical properties of crystalline materials makes it much easier to examine the band structures of these materials theoretically. In addition, the well-defined symmetry axes of crystalline materials makes it possible to determine the dispersion relationship between the momentum (a 3-dimension vector quantity) and energy of a material. As a result, virtually all of the existing theoretical work on the electronic band structure of solids has focused on crystalline materials.

Density of states

While the density of energy states in a band could be very large for some materials, it may not be uniform. It approaches zero at the band boundaries, and is generally highest near the middle of a band. The density of states for the free electron model in three dimensions is given by,

$$D(\epsilon) = \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\epsilon^{1/2}$$

Filling of bands

Although the number of states in all of the bands is effectively infinite, in an uncharged material the number of electrons is equal only to the number of protons in the atoms of the material. Therefore not all of the states are occupied by electrons ("filled") at any time. The likelihood of any particular state being filled at any temperature is given by Fermi-Dirac statistics. The probability is given by the following expression:

$$f(E) = \frac{1}{1 + e^{\frac{E-\mu}{k_B T}}}$$

where:

- k_B is Boltzmann's constant,
- T is the temperature,
- μ is the chemical potential (in semiconductor physics, this quantity is more often called the "Fermi level" and denoted E_F).

The Fermi level naturally is the level at which the electrons and protons are balanced.

At *T=0*, the distribution is a simple step function:

$$f(E) = \begin{cases} 1 & \text{if } 0 < E \leq E_F \\ 0 & \text{if } E_F < E \end{cases}$$

At nonzero temperatures, the step "smooths out", so that an appreciable number of states below the Fermi level are empty, and some states above the Fermi level are filled.

Theory of band structures in crystals

The ansatz is the special case of electron waves in a periodic crystal lattice using Bloch waves as treated generally in the dynamical theory of diffraction. Every crystal is a periodic structure which can be characterized by a Bravais lattice, and for each Bravais lattice we can determine the reciprocal lattice, which encapsulates the periodicity in a set of three reciprocal lattice vectors ($\mathbf{b_1},\mathbf{b_2},\mathbf{b_3}$). Now, any periodic potential V(**r**) which shares the same periodicity as the direct lattice can be expanded out as a Fourier series whose only non-vanishing components are those associated with the reciprocal lattice vectors. So the expansion can be written as:

$$V(\mathbf{r}) = \sum_{\mathbf{K}} V_{\mathbf{K}} e^{i\mathbf{K}\cdot\mathbf{r}}$$

where $\mathbf{K} = m_1\mathbf{b_1} + m_2\mathbf{b_2} + m_3\mathbf{b_3}$ for any set of integers (m_1,m_2,m_3).

From this theory, an attempt can be made to predict the band structure of a particular material, however most ab initio methods for electronic structure calculations fail to predict the observed band gap.

Nearly free electron approximation

In the nearly free electron approximation, interactions between electrons are completely ignored. This approximation allows use of Bloch's Theorem which states that electrons in a periodic potential have wavefunctions and energies which are periodic in wavevector up to a constant phase shift between neighboring reciprocal lattice vectors. The consequences of periodicity are described mathematically by the Bloch wavefunction:

$$\Psi_{n,\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_n(\mathbf{r})$$

where the function $u_n(\mathbf{r})$is periodic over the crystal lattice, that is,

$$u_n(\mathbf{r}) = u_n(\mathbf{r} - \mathbf{R}).$$

Here index *n* refers to the *n-th* energy band, wavevector **k** is related to the direction of motion of the electron, **r** is position in the crystal, and **R** is location of an atomic site.

The NFE model works particularly well in materials like metals where distances between neighbouring atoms are small. In such materials the overlap of atomic orbitals and potentials on neighbouring atoms is relatively large. In that case the wave function of the electron can be approximated by a (modified) plane wave. The band structure of a metal like Aluminum even gets close to the Empty Lattice Approximation.

Tight-binding model

The opposite extreme to the nearly-free electron approximation assumes the electrons in the crystal behave much like an assembly of constituent atoms. This tight-binding model assumes the solution to the time-independent single electron Schrödinger equation Ψ is well approximated by a linear combination of atomic orbitals $\psi_n(\mathbf{r})$.

$$\Psi(\mathbf{r}) = \sum_{n,\mathbf{R}} b_{n,\mathbf{R}} \psi_n(\mathbf{r} - \mathbf{R}),$$

where the coefficients $b_{n,\mathbf{R}}$are selected to give the best approximate solution of this form. Index *n* refers to an atomic energy level and **R** refers to an atomic site. A more accurate approach using this idea employs Wannier functions, defined by:

$$a_n(\mathbf{r} - \mathbf{R}) = \frac{V_C}{(2\pi)^3} \int_{BZ} d\mathbf{k} e^{-i\mathbf{k}\cdot(\mathbf{R}-\mathbf{r})} u_{n\mathbf{k}};$$

in which $u_{n\mathbf{k}}$is the periodic part of the Bloch wave and the integral is over the Brillouin zone. Here index *n* refers to the *n*-th energy band in the crystal. The Wannier functions are localized near atomic sites, like atomic orbitals, but being defined in terms of Bloch functions they are accurately related to solutions based upon the crystal potential. Wannier functions on different atomic sites **R** are orthogonal. The Wannier functions can be used to form the Schrödinger solution for the *n*-th energy band as:

$$\Psi_{n,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}} e^{-i\mathbf{k}\cdot(\mathbf{R}-\mathbf{r})} a_n(\mathbf{r} - \mathbf{R}).$$

The TB model works well in materials with limited overlap between atomic orbitals and potentials on neighbouring atoms. Band structures of materials like Si, GaAs, SiO_2 and diamond for instance are well described by TB-Hamiltonians on the basis of atomic sp^3 orbitals. In transition metals a mixed TB-NFE model is used to describe the broad NFE conduction band and the narrow embedded TB d-bands. The radial functions of the atomic orbital part of the Wannier functions are most easily calculated by the use of pseudopotential methods. NFE, TB or combined NFE-TB band structure calculations, sometimes extended with wave function approximations based on pseudopotential methods, are often used as an economic starting point for further calculations.

KKR model

The simplest form of this approximation centers non-overlapping spheres (referred to as *muffin tins*) on the atomic positions. Within these regions, the potential experienced by an electron is approximated to be spherically symmetric about the given nucleus. In the remaining interstitial region, the screened potential is approximated as a constant.

Continuity of the potential between the atom-centered spheres and interstitial region is enforced.

A variational implementation was suggested by Korringa and by Kohn and Rostocker, and is often referred to as the *KKR model*.

Order-N spectral methods

To quote RP Martin: "The concept of localization can be imbedded directly into the methods of electronic structure to create algorithms that take advantage of locality … For large systems, this fact can be used to make "order-N" or *O(N)* methods where the computational time scales linearly in the size of the system".

Density-functional theory

In recent physics literature, a large majority of the electronic structures and band plots are calculated using density-functional theory (DFT), which is not a model but rather a theory, i.e., a microscopic first-principles theory of condensed matter physics that tries to cope with the electron-electron many-body problem via the introduction of an exchange-correlation term in the functional of the electronic density. DFT-calculated bands are in many cases found to be in agreement with experimentally measured bands, for example by angle-resolved photoemission spectroscopy (ARPES). In particular, the band shape is typically well reproduced by DFT. But there are also systematic errors in DFT bands when compared to experiment results. In particular, DFT seems to systematically underestimate by about 30-40% the band gap in insulators and semiconductors.

It must be said that DFT is, in principle an exact theory to reproduce and predict ground state properties (e.g., the total energy, the atomic structure, etc.). However, DFT is not a theory to address excited state properties, such as the band plot of a solid that represents the excitation energies of electrons injected or removed from the system. What in literature is quoted as a DFT band plot is a representation of the DFT Kohn-Sham energies, i.e., the energies of a fictive non-interacting system, the Kohn-Sham system, which has no physical interpretation at all. The Kohn-Sham electronic structure must not be confused with the real, quasiparticle electronic structure of a system, and there is no Koopman's theorem holding for Kohn-Sham energies, as there is for Hartree-Fock energies, which can be truly considered as an approximation for quasiparticle energies. Hence, in principle, DFT is not a band theory, i.e., not a theory suitable for calculating bands and band-plots.

Green's function methods and the *ab initio* GW approximation

To calculate the bands including electron-electron interaction many-body effects, one can resort to so-called Green's function methods. Indeed, knowledge of the Green's function of a system provides both ground (the total energy) and also excited state observables of the system. The poles of the Green's function are the quasiparticle energies, the bands of a solid. The Green's function can be calculated by solving the Dyson equation once the self-energy of the system is known. For real systems like solids, the self-energy is a very complex quantity and usually approximations are needed to solve the problem. One such approximation is the GW approximation, so called from the mathematical form the self-energy takes as the product $\Sigma = GW$ of the Green's function G and the dynamically screened interaction W. This approach is more pertinent when addressing the calculation of band plots (and also quantities beyond, such as the spectral function) and can also be formulated in a completely *ab initio* way. The GW approximation seems to provide band gaps of insulators and semiconductors in agreement with experiment, and hence to correct the systematic DFT underestimation.

Mott insulators

Although the nearly-free electron approximation is able to describe many properties of electron band structures, one consequence of this theory is that it predicts the same number of electrons in each unit cell. If the number of electrons is odd, we would then expect that there is an unpaired electron in each unit cell, and thus that the valence band is not fully occupied, making the material a conductor. However, materials such as CoO that have an odd number of electrons per unit cell are insulators, in direct conflict with this result. This kind of material is known as a Mott insulator, and requires inclusion of detailed electron-electron interactions (treated only as an averaged effect on the crystal potential in band theory) to explain the discrepancy. The Hubbard model is an approximate theory that can include these interactions. It can be treated non-perturbatively within the so-called Dynamical Mean Field Theory, which bridges the gap between the nearly-free electron approximation and the atomic limit.

Augmented plane waves

John Clarke Slater and members of his Solid State and Molecular Theory Group in the Physics Department at MIT, comprised one of the main research centers for the calculation of band structures. John Wood played a very strong role in large scale computations using the augmented plane wave (APW) method.

Others

Calculating band structures is an important topic in theoretical solid state physics. In addition to the models mentioned above, other models include the following:

- k·p perturbation theory is a technique that allows a band structure to be approximately described in terms of just a few parameters. The technique is commonly used for semiconductors, and the parameters in the model are often determined by experiment.
- The Kronig-Penney Model, a one-dimensional rectangular well model useful for illustration of band formation. While simple, it predicts many important phenomena, but is not quantitative.

- Bands may also be viewed as the large-scale limit of molecular orbital theory. A solid creates a large number of closely spaced molecular orbitals, which appear as a band.
- Hubbard model

The band structure has been generalised to wavevectors that are complex numbers, resulting in what is called a *complex band structure*, which is of interest at surfaces and interfaces.

Each model describes some types of solids very well, and others poorly. The nearly-free electron model works well for metals, but poorly for non-metals. The tight binding model is extremely accurate for ionic insulators, such as metal halide salts (e.g. NaCl).

Chapter 5

Thermal Copper Pillar Bump

The **thermal copper pillar bump**, also known as the "thermal bump", is a thermoelectric device made from thin-film thermoelectric material embedded in flip chip interconnects (in particular copper pillar solder bumps) for use in electronics and optoelectronic packaging, including: flip chip packaging of CPU and GPU integrated circuits (chips), laser diodes, and semiconductor optical amplifiers (SOA). Unlike conventional solder bumps that provide an electrical path and a mechanical for connection to the package, thermal bumps act as solid-state heat pumps and add thermal management functionality locally on the surface of a chip or to another electrical component. The diameter of a thermal bump is 238 µm (micrometres) and 60 µm high.

The thermal bump uses the thermoelectric effect, which is the direct conversion of temperature differences to electric voltage and vice versa. Simply put, a thermoelectric device creates a voltage when there is a different temperature on each side, or when a voltage is applied to it, it creates a temperature difference. This effect can be used to generate electricity, to measure temperature, to cool objects, or to heat them.

For each bump, thermoelectric cooling (TEC) occurs when a current is passed through the bump. The thermal bump pulls heat from one side of the device and transfers it to the other as current is passed through the material. This is known as the Peltier effect. The direction of heating and cooling is determined by the direction of current flow and the sign of the majority electrical carrier in the thermoelectric material. Thermoelectric power generation (TEG) on the other hand occurs when the thermal bump is subjected to a temperature gradient (i.e., the top is hotter than the bottom). In this instance, the device generates current, converting heat into electrical power. This is termed the Seebeck effect.

The thermal bump was developed by Nextreme Thermal Solutions as a method for integrating active thermal management functionality at the chip level in the same manner that transistors, resistors and capacitors are integrated in conventional circuit designs today. Nextreme chose the copper pillar bump as an integration strategy due to its widespread acceptance by Intel, Amkor and other industry leaders as the method for connecting microprocessors and other advanced electronics devices to various surfaces during a process referred to as “flip-chip” packaging. The thermal bump can be integrated as a part of the standard flip-chip process (Figure 1) or integrated as discrete devices.

The efficiency of a thermoelectric device is measured by the heat moved (or pumped) divided by the amount of electrical power supplied to move this heat. This ratio is termed the coefficient of performance or COP and is a measured characteristic of a thermoelectric device. The COP is inversely related to the temperature difference that the device produces. As you move a cooling device further away from the heat source, parasitic losses between the cooler and the heat source necessitate additional cooling power: the further the distance between source and cooler, the more cooling is required. For this reason, the cooling of electronic devices is most efficient when it occurs closest to the source of the heat generation.

Use of the thermal bump does not displace system level cooling, which is still needed to move heat out of the system; rather it introduces a fundamentally new methodology for achieving temperature uniformity at the chip and board level. In this manner, overall thermal management of the system becomes more efficient. In addition, while conventional cooling solutions scale with the size of the system (bigger fans for bigger systems, etc.), the thermal bump can scale at the chip level by using more thermal bumps in the overall design.

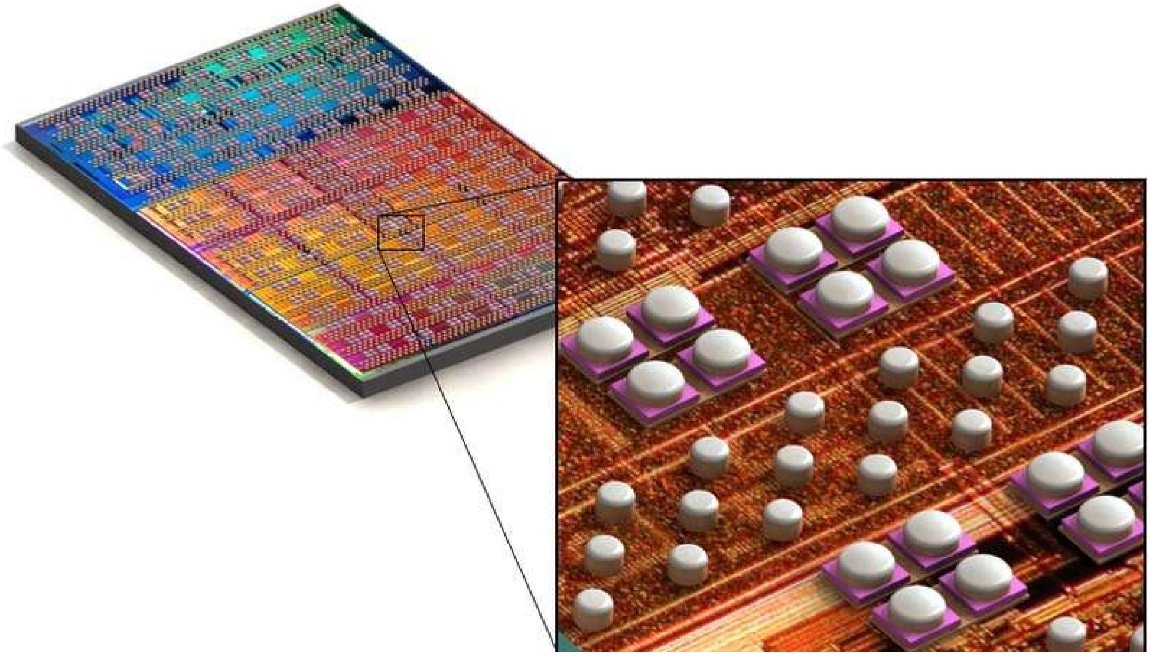


Figure 1. Thermal and electrical bumps integrated on a single substrate.

A brief history of solder and flip chip/chip scale packaging

Solder bumping technology (the process of joining a chip to a substrate without shorting using solder) was first conceived and implemented by IBM in the early ‘60s. Three versions of this type of solder joining were developed. The first was to embed copper balls in the solder bumps to provide a positive stand-off. The second solution, developed by Delco Electronics (General Motors) in the late ‘60s, was similar to embedding copper balls except that the design employed a rigid silver bump. The bump provided a positive stand-off and was attached to the substrate by means of solder that was screen-printed onto the substrate. The third solution was to use a screened glass dam near the electrode tips to act as a ‘‘stop-off’’ to prevent the ball solder from flowing down the electrode. By then the Ball Limiting Metallurgy (BLM) with a high-lead (Pb) solder system and a copper ball had proven to work well. Therefore, the ball was simply removed and the solder evaporation process extended to form pure solder bumps that were approximately 125μm high. This system became known as the controlled collapse chip connection (C3 or C4).

Until the mid-90’s, this type of flip-chip assembly was practiced almost exclusively by IBM and Delco. Around this time, Delco sought to commercialize its technology and formed Flip Chip Technologies with Kulicke & Soffa as a partner. At the same time, MCNC (which had developed a plated version of IBM’s C4 process) received funding from DARPA to commercialize its technology. These two organizations, along with APTOS (Advanced Plating Technologies on Silicon), formed the nascent out-sourcing market.

During this same time, companies began to look at reducing or streamlining their packaging, from the earlier multi-chip-on-ceramic packages that IBM had originally developed C4 to support, to what were referred to as Chip Scale Packages (CSP). There were a number of companies developing products in this area. These products could usually be put into one of two camps: either they were scaled down versions of the multi-chip on ceramic package (of which the Tessera package would be one example); or they were the streamlined versions developed by Unitive Electronics, et al. (where the package wiring had been transferred to the chip, and after bumping, they were ready to be placed).

One of the issues with the CSP type of package (which was intended to be soldered directly to an FR4 or flex circuit) was that for high-density interconnects, the soft solder bump provided less of a stand-off as the solder bump diameter and pitch were decreased. Different solutions were employed including one developed by Focus Interconnect Technology (former APTOS engineers), which used a high aspect ratio plated copper post to provide a larger fixed standoff than was possible for a soft solder collapse joint.

Today, flip chip is a well established technology and collapsed soft solder connections are used in the vast majority of assemblies. Interestingly, the copper post stand-off developed for the CSP market has found a home in high-density interconnects for advanced micro-processors and is used today by IBM for its CPU packaging.

Copper pillar solder bumping

Recent trends in high-density interconnects have led to the use of copper pillar solder bumps (CPB) for CPU and GPU packaging. CPBs are an attractive replacement for traditional solder bumps because they provide a fixed stand-off independent of pitch. This is extremely important as most of the high-end products are underfilled and a smaller standoff may create difficulties in getting the underfill adhesive to flow under the die.

Figure 2 shows an example of a CPB fabricated by Intel and incorporated into their Presler line of microprocessors among others. The cross section shows copper and a copper pillar (approximately 60 um high) electrically connected through an opening (or via) in the chip passivation layer at the top of the picture. At the bottom is another copper trace on the package substrate with solder between the two copper layers.

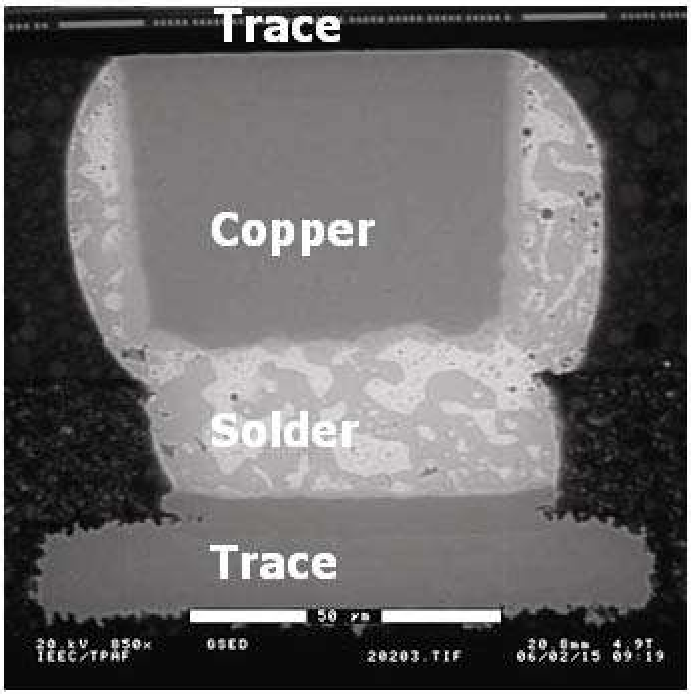


Figure 2. Intel Presler copper pillar solder bump

Thin-film thermoelectric technology

Thin films are thin material layers ranging from fractions of a nanometer to several micrometers in thickness. Thin-film thermoelectric materials are grown by conventional semiconductor deposition methods and fabricated using conventional semiconductor micro-fabrication techniques.

Thin-film thermoelectrics have been demonstrated to provide high heat pumping capacity that far exceeds the capacities provided by traditional bulk pellet TE products. The benefit of thin-films as compared to bulk materials for thermoelectric manufacturing is expressed in Equation 1. Here the Qmax (maximum heat pumped by a module) is shown to be inversely proportional to the thickness of the film, L.

$$Q_{max} = \frac{S^2T^2}{2 \cdot R_{Total}} = \frac{S^2T^2A}{2p_cL} \qquad \text{Eq. 1}$$

As such, TE coolers manufactured with thin-films can easily have 10x – 20x higher Qmax values for a given active area A. This makes thin-film TECs ideally suited for applications involving high heat-flux flows. In addition to the increased heat pumping capability, the use of thin films allows for truly novel implementation of TE devices. Instead of a bulk module that is 1-3 mm in thickness, a thin-film TEC can be fabricated less than 100 um in thickness.

In its simplest form, the P or N leg of a TE couple (the basic building block of all thin-film TE devices) is a layer of thin-film TE material with a solder layer above and below, providing electrical and thermal functionality.

Thermal copper pillar bump

The thermal bump is compatible with the existing flip-chip manufacturing infrastructure, extending the use of conventional solder bumped interconnects to provide active, integrated cooling of a flip-chipped component using the widely accepted copper pillar bumping process. The result is higher performance and efficiency within the existing semiconductor manufacturing paradigm. The thermal bump also enables power generating capabilities within copper pillar bumps for energy recycling applications.

Thermal bumps have been shown to achieve a temperature differential of 60 °C between the top and bottom headers; demonstrated power pumping capabilities exceeding 150 W/cm2; and when subjected to heat, have demonstrated the capability to generate up to 10 mW of power per bump.

Thermal copper pillar bump structure

Figure 3 shows an SEM cross-section of a TE leg. Here it is demonstrated that the thermal bump is structurally identical to a CPB with an extra layer, the TE layer, incorporated into the stack-up. The addition of the TE layer transforms a standard copper pillar bump into a thermal bump. This element, when properly configured electrically and thermally, provides active thermoelectric heat transfer from one side of the bump to the other side. The direction of heat transfer is dictated by the doping type of the thermoelectric material (either a P-type or N-type semiconductor) and the direction of electrical current passing through the material. This type of thermoelectric heat transfer is known as the Peltier effect. Conversely, if heat is allowed to pass from one side of the thermoelectric material to the other, a current will be generated in the material in a phenomenon known as the Seebeck effect. The Seebeck effect is essentially the reverse of the Peltier effect. In this mode, electrical power is generated from the flow of heat in the TE element. The structure shown in Figure 3 is capable of operating in both the Peltier and Seebeck modes, though not simultaneously.

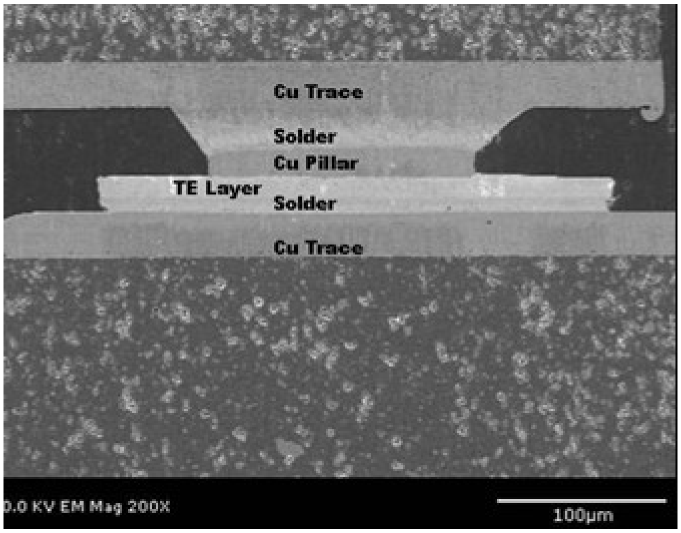


Figure 3. Cross sectional view of Nextreme's thermal copper pillar bump.

Figure 4 shows a schematic of a typical CPB and a thermal bump for comparison purposes. These structures are similar, with both having copper pillars and solder connections. The primary distinction between the two is the introduction of either a P- or N-type thermoelectric layer between two solder layers. The solders used with CPBs and thermal bumps can be any one of a number of commonly used solders including, but not limited to, Sn, SnPb eutectic, SnAg or AuSn.

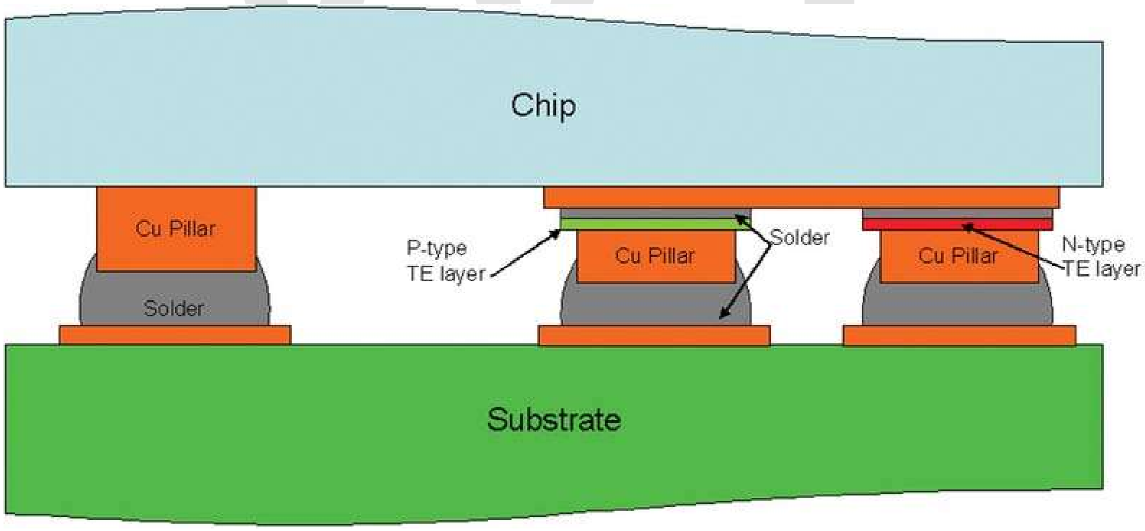


Figure 4. Schematic showing traditional CPB next to a P-type and N-type pillar bump. The P- and N-type bumps together make up a P/N couple that, when connected in series electrically, provides for either Peltier cooling or Seebeck power generation.

Figure 5 shows a device equipped with a thermal bump. The thermal flow is shown by the arrows labeled "heat." Metal traces, which can be several micrometres high, can be stacked or interdigitated to provide highly conductive pathways for collecting heat from the underlying circuit and funneling that heat to the thermal bump.

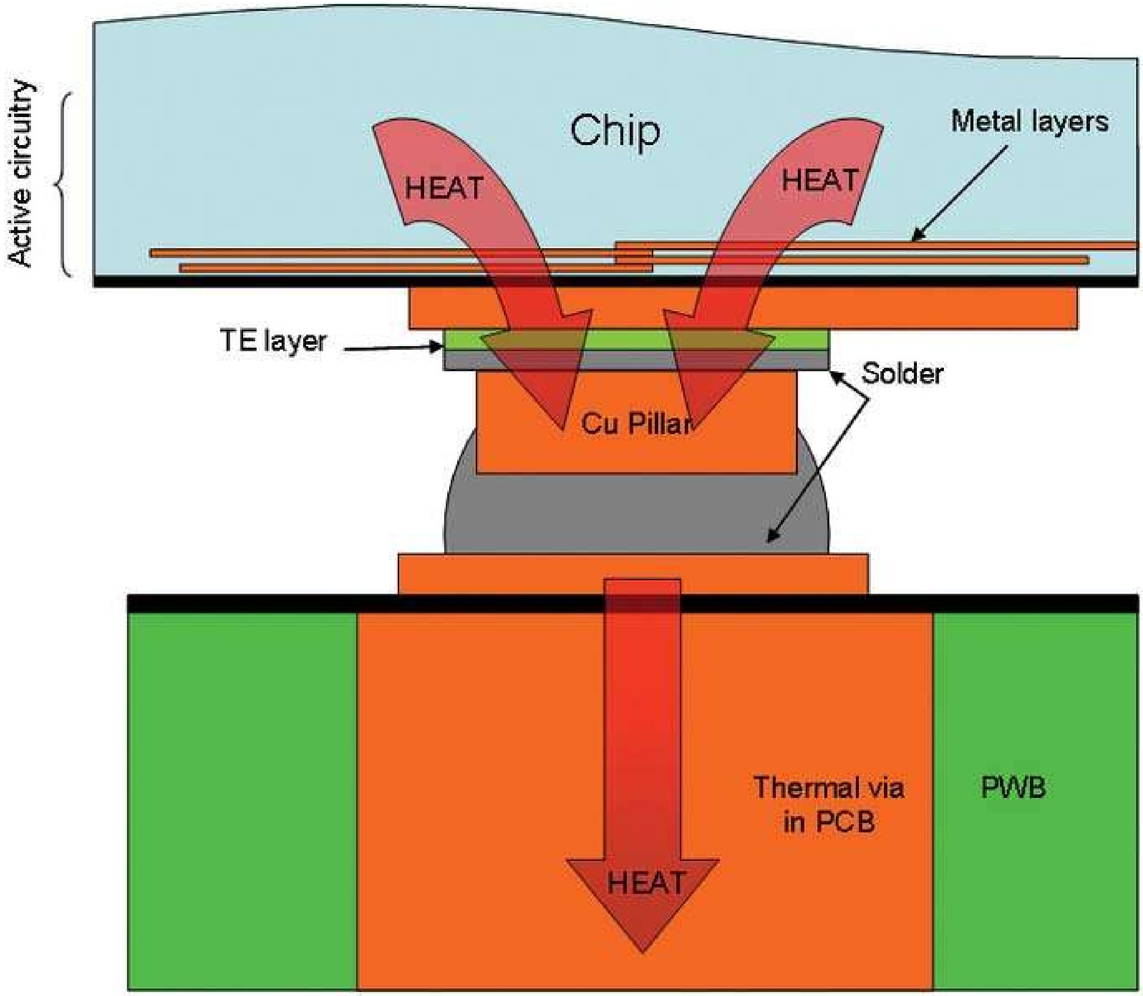


Figure 5. Close-up schematic showing flow of heat through a thermal bump. Also shown are the multi-layer metal traces often used in complex integrated circuits. These metal layers can be beneficial for gathering heat from larger areas and funneling it into the thermal bump, reducing the thermal constriction resistance in the circuit. A thermal via is shown in the printed wire board for improved heat rejection path.

The metal traces shown in the figure for conducting electrical current into the thermal bump may or may not be directly connected to the circuitry of the chip. In the case where there are electrical connections to the chip circuitry, on-board temperature sensors and driver circuitry can be used to control the thermal bump in a closed loop fashion to maintain optimal performance. Second, the heat that is pumped by the thermal bump and the additional heat created by the thermal bump in the course of pumping that heat will need to be rejected into the substrate or board. Since the performance of the thermal bump can be improved by providing a good thermal path for the rejected heat, it is beneficial to provide high thermally conductive pathways on the backside of the thermal bump. The substrate could be a highly conductive ceramic substrate like AlN or a metal (e.g., Cu, CuW, CuMo, etc.) with a dielectric. In this case, the high thermal conductance of the substrate will act as a natural pathway for the rejected heat. The substrate might also be a multilayer substrate like a printed wiring board (PWB) designed to provide a high-density interconnect. In this case, the thermal conductivity of the PWB may be relatively poor, so adding thermal vias (e.g. metal plugs) can provide excellent pathways for the rejected heat.

Applications

Thermal bumps can be used in a number of different ways to provide chip cooling and power generation.

General cooling

Thermal bumps can be evenly distributed across the surface of a chip to provide a uniform cooling effect. In this case, the thermal bumps may be interspersed with standard bumps that are used for signal, power and ground. This allows the thermal bumps to be placed directly under the active circuitry of the chip for maximum effectiveness. The number and density of thermal bumps are based on the heat load from the chip. Each P/N couple can provide a specific heat pumping (Q) at a specific temperature differential (ΔT) at a given electrical current. Temperature sensors on the chip ("on board" sensors) can provide direct measurement of the thermal bump performance and provide feedback to the driver circuit.

Precision temperature control

Since thermal bumps can either cool or heat the chip depending on the current direction, they can be used to provide precision control of temperature for chips that must operate within specific temperature ranges irrespective of ambient conditions. For example, this is a common problem for many optoelectronic components.

Hotspot cooling

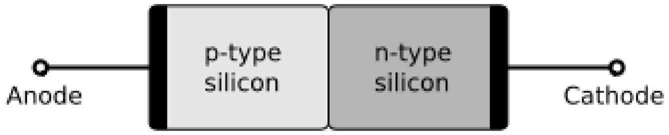
In microprocessors, graphics processors and other high-end chips, hotspots can occur as power densities vary significantly across a chip. These hotspots can severely limit the performance of the devices. Because of the small size of the thermal bumps and the relatively high density at which they can be placed on the active surface of the chip, these structures are ideally suited for cooling hotspots. In such a case, the distribution of the thermal bumps may not need to be even. Rather, the thermal bumps would be concentrated in the area of the hotspot while areas of lower heat density would have fewer thermal bumps per unit area. In this way, cooling from the thermal bumps is applied only where needed, thereby reducing the added power necessary to drive the cooling and reducing the general thermal overhead on the system.

Power generation

In addition to chip cooling, thermal bumps can also be applied to high heat-flux interconnects to provide a constant, steady source of power for energy scavenging applications. Such a source of power, typically in the mW range, can trickle charge batteries for wireless sensor networks and other battery operated systems.

Chapter 6

p-n Junction



A silicon p–n junction with no applied voltage

A **p–n junction** is formed by joining P-type and N-type semiconductors together in very close contact. The term *junction* refers to the boundary interface where the two regions of the semiconductor meet. If they were constructed of two separate pieces this would introduce a grain boundary, so p–n junctions are created in a single crystal of semiconductor by doping, for example by ion implantation, diffusion of dopants, or by epitaxy (growing a layer of crystal doped with one type of dopant on top of a layer of crystal doped with another type of dopant).

P-N junctions are elementary "building blocks" of almost all semiconductor electronic devices such as diodes, transistors, solar cells, LEDs, and integrated circuits; they are the active sites where the electronic action of the device takes place. For example, a common type of transistor, the bipolar junction transistor, consists of two p–n junctions in series, in the form n–p–n or p–n–p.

The discovery of the p–n junction is usually attributed to American physicist Russell Ohl of Bell Laboratories.

Schottky junction is a special case of a p-n junction, where metal serves the role of the n-type semiconductor.

Manufacture

Normally, p–n junctions are manufactured from a single crystal with different dopant concentrations diffused across it. Creating a semiconductor from two separate pieces of material would introduce a grain boundary between the semiconductors which severely inhibits its utility by scattering the electrons and holes.. However, in the case of solar cells, polycrystalline silicon is often used to reduce expense, despite the lower efficiency.

Properties of a p-n junction

The p–n junction possesses some interesting properties which have useful applications in modern electronics. A p-doped semiconductor is relatively conductive. The same is true of an n-doped semiconductor, but the junction between them can become depleted of charge carriers, and hence nonconductive, depending on the relative voltages of the two semiconductor regions. By manipulating this non-conductive layer, p–n junctions are commonly used as diodes: circuit elements that allow a flow of electricity in one direction but not in the other (opposite) direction. This property is explained in terms of *forward bias* and *reverse bias*, where the term *bias* refers to an application of electric voltage to the p–n junction.

Equilibrium (zero bias)

In a p–n junction, without an external applied voltage, an equilibrium condition is reached in which a potential difference is formed across the junction. This potential difference is called built-in potential V_{bi}.

After joining p-type and n-type semiconductors, electrons near the p–n interface tend to diffuse into the p region. As electrons diffuse, they leave positively charged ions (donors) in the n region. Similarly, holes near the p–n interface begin to diffuse into the n-type region leaving fixed ions (acceptors) with negative charge. The regions nearby the p–n interfaces lose their neutrality and become charged, forming the space charge region or depletion layer.

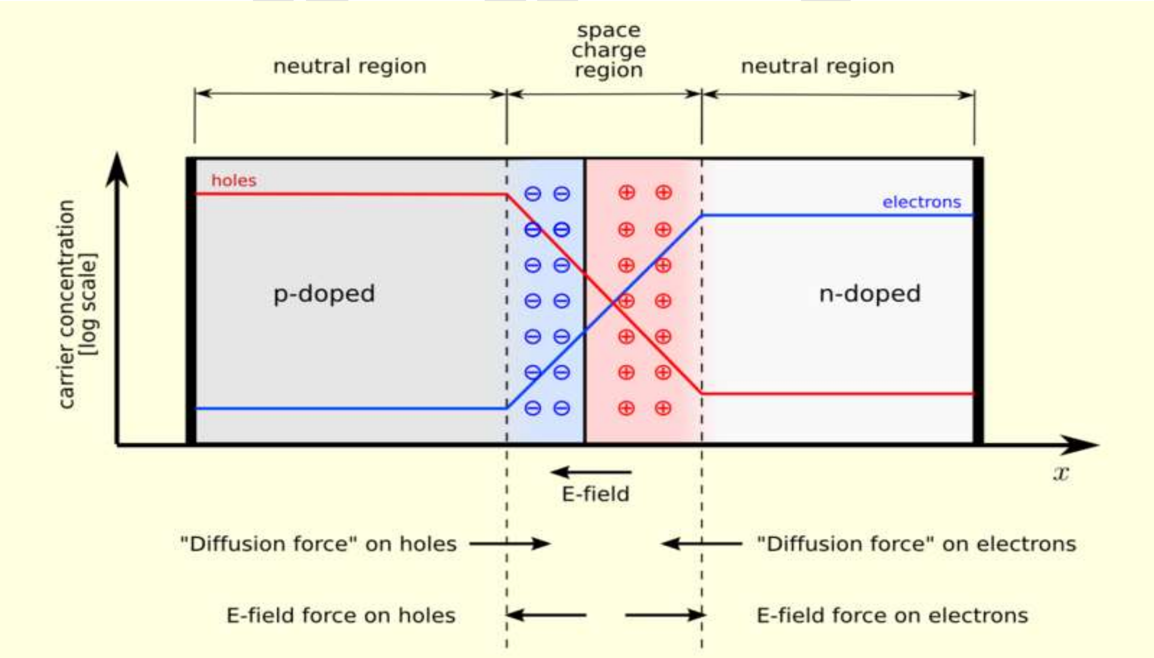

space charge region
neutral region
neutral region
holes
electrons
p-doped
n-doped
carrier concentration [log scale]
x
E-field
"Diffusion force" on holes
"Diffusion force" on electrons
E-field force on holes
E-field force on electrons

Figure A. A p–n junction in thermal equilibrium with zero bias voltage applied. Electrons and holes concentration are reported respectively with blue and red lines. Gray regions are charge neutral. Light red zone is positively charged. Light blue zone is negatively charged. The electric field is shown on the bottom, the electrostatic force on electrons and holes and the direction in which the diffusion tends to move electrons and holes.

The electric field created by the space charge region opposes the diffusion process for both electrons and holes. There are two concurrent phenomena: the diffusion process that tends to generate more space charge, and the electric field generated by the space charge that tends to counteract the diffusion. The carrier concentration profile at equilibrium is shown in figure A with blue and red lines. Also shown are the two counterbalancing phenomena that establish equilibrium.

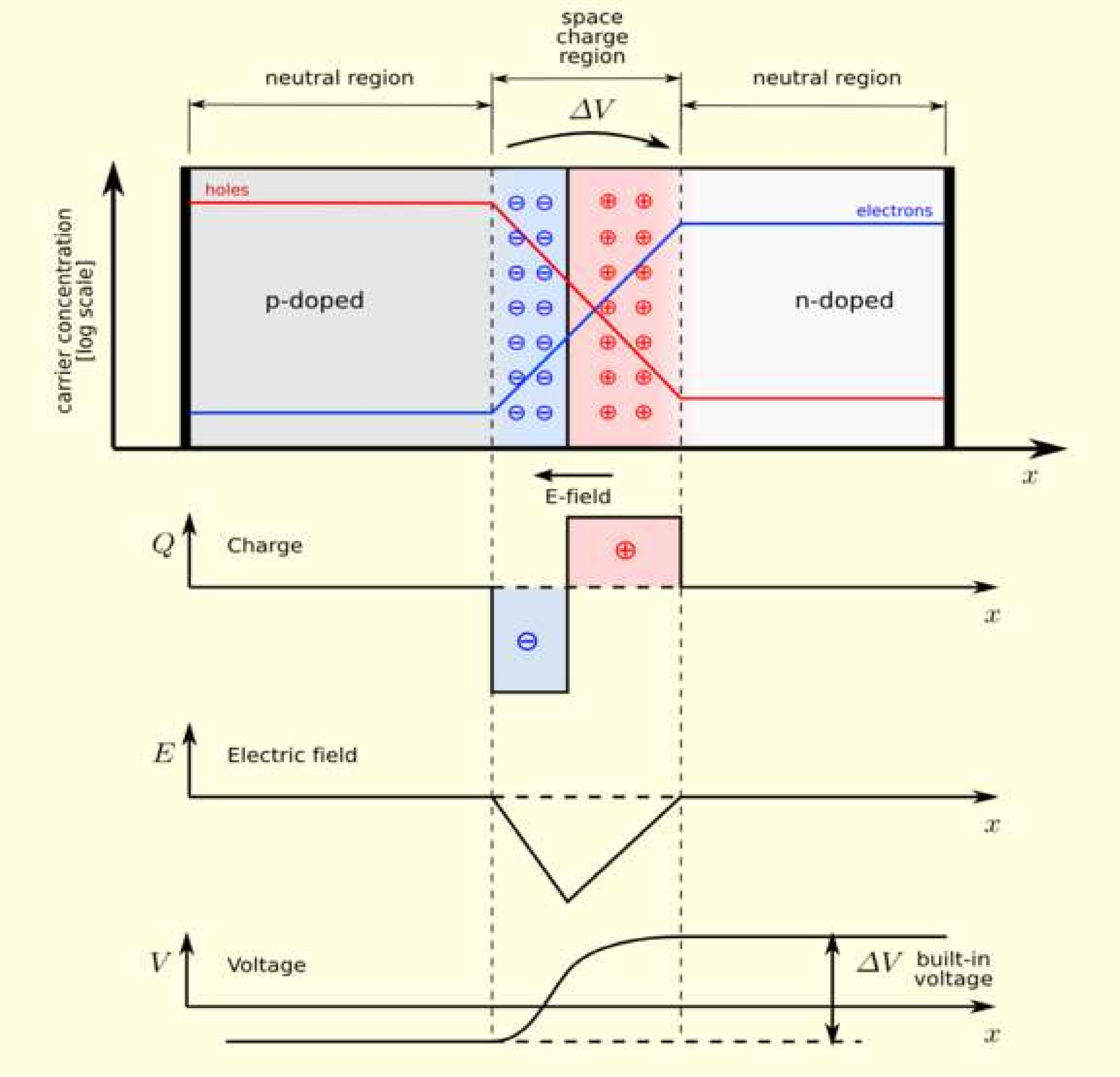
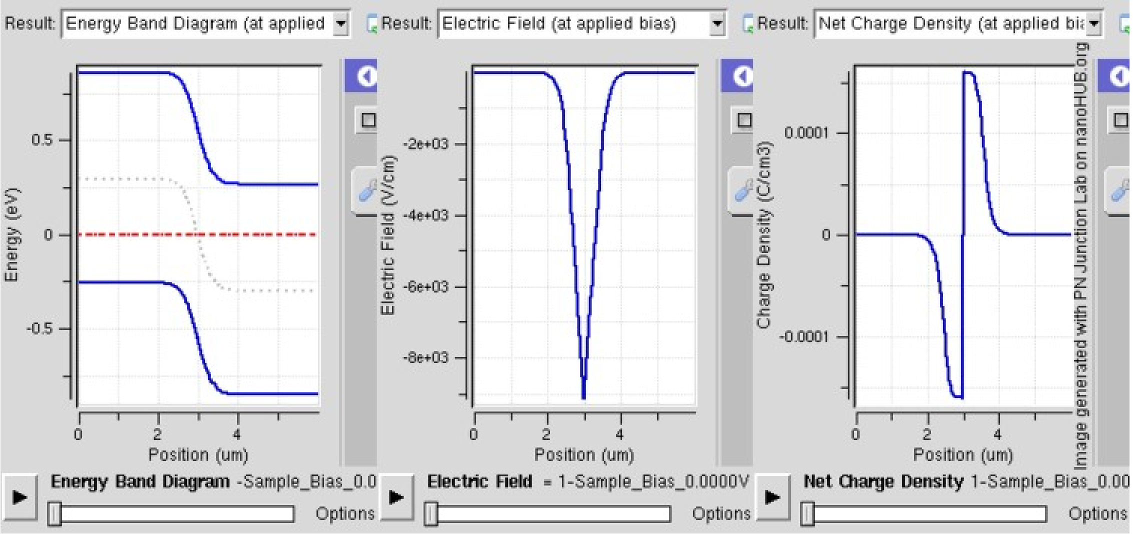


Figure B. A p–n junction in thermal equilibrium with zero bias voltage applied. Under the junction, plots for the charge density, the electric field and the voltage are reported.

The space charge region is a zone with a net charge provided by the fixed ions (donors or acceptors) that have been left *uncovered* by majority carrier diffusion. When equilibrium is reached, the charge density is approximated by the displayed step function. In fact, the region is completely depleted of majority carriers (leaving a charge density equal to the net doping level), and the edge between the space charge region and the neutral region is quite sharp. The space charge region has the same magnitude of charge on both sides of the p–n interfaces, thus it extends farther on the less doped side (the n side in figures A and B).

Forward bias

In forward bias, the p-type is connected with the positive terminal and the n-type is connected with the negative terminal.



Result: Energy Band Diagram (at applied

Result: Electric Field (at applied bias)

Result: Net Charge Density (at applied bi

PN junction operation in forward bias mode showing reducing depletion width. Both p and n junctions are doped at a 1e15/cm3 doping level, leading to built-in potential of ~0.59V. Reducing depletion width can be inferred from the shrinking charge profile, as fewer dopants are exposed with increasing forward bias.

With a battery connected this way, the holes in the P-type region and the electrons in the N-type region are pushed towards the junction. This reduces the width of the depletion zone. The positive charge applied to the P-type material repels the holes, while the negative charge applied to the N-type material repels the electrons. As electrons and holes are pushed towards the junction, the distance between them decreases. This lowers the barrier in potential. With increasing forward-bias voltage, the depletion zone eventually becomes thin enough that the zone's electric field can't counteract charge carrier motion across the p–n junction, consequently reducing electrical resistance. The electrons which cross the p–n junction into the P-type material (or holes which cross into the N-type material) will diffuse in the near-neutral region. Therefore, the amount of minority diffusion in the near-neutral zones determines the amount of current that may flow through the diode.

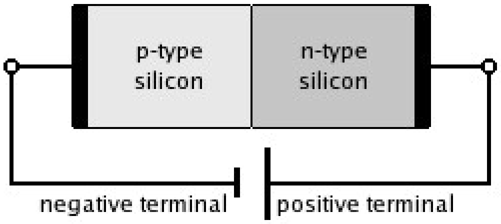
Only majority carriers (electrons in N-type material or holes in P-type) can flow through a semiconductor for a macroscopic length. With this in mind, consider the flow of electrons across the junction. The forward bias causes a force on the electrons pushing them from the N side toward the P side. With forward bias, the depletion region is narrow enough that electrons can cross the junction and *inject* into the P-type material. However, they do not continue to flow through the P-type material indefinitely, because it is energetically favorable for them to recombine with holes. The average length an electron travels through the P-type material before recombining is called the *diffusion length*, and it is typically on the order of microns.

Although the electrons penetrate only a short distance into the P-type material, the electric current continues uninterrupted, because holes (the majority carriers) begin to flow in the opposite direction. The total current (the sum of the electron and hole currents) is constant in space, because any variation would cause charge buildup over time (this is Kirchhoff's current law). The flow of holes from the P-type region into the N-type region is exactly analogous to the flow of electrons from N to P (electrons and holes swap roles and the signs of all currents and voltages are reversed).

Therefore, the macroscopic picture of the current flow through the diode involves electrons flowing through the N-type region toward the junction, holes flowing through the P-type region in the opposite direction toward the junction, and the two species of carriers constantly recombining in the vicinity of the junction. The electrons and holes travel in opposite directions, but they also have opposite charges, so the overall current is in the same direction on both sides of the diode, as required.

The Shockley diode equation models the forward-bias operational characteristics of a p–n junction outside the avalanche (reverse-biased conducting) region.

Reverse bias



A silicon p–n junction in reverse bias

Reverse biased usually refers to how a diode is used in a circuit. If a diode is reverse biased, the voltage at the cathode is higher than that at the anode. Therefore, no current will flow until the diode breaks down. Connecting the *P-type* region to the *negative* terminal of the battery and the *N-type* region to the *positive* terminal, corresponds to reverse bias. The connections are illustrated in the following diagram:

Because the p-type material is now connected to the negative terminal of the power supply, the 'holes' in the P-type material are pulled away from the junction, causing the width of the depletion zone to increase. Similarly, because the N-type region is connected to the positive terminal, the electrons will also be pulled away from the junction. Therefore the depletion region widens, and does so increasingly with increasing reverse-bias voltage. This increases the voltage barrier causing a high resistance to the flow of charge carriers thus allowing minimal electric current to cross the p–n junction. The increase in resistance of the p-n junction results in the junction behaving as an insulator.

The strength of the depletion zone electric field increases as the reverse-bias voltage increases. Once the electric field intensity increases beyond a critical level, the p–n junction depletion zone breaks-down and current begins to flow, usually by either the Zener or avalanche breakdown processes. Both of these breakdown processes are non-destructive and are reversible, so long as the amount of current flowing does not reach levels that cause the semiconductor material to overheat and cause thermal damage.

This effect is used to one's advantage in zener diode regulator circuits. Zener diodes have a certain - low - breakdown voltage. A standard value for breakdown voltage is for instance 5.6V. This means that the voltage at the cathode can never be more than 5.6V higher than the voltage at the anode, because the diode will break down - and therefore conduct - if the voltage gets any higher. This effectively regulates the voltage over the diode.

Another application where reverse biased diodes are used is in Varicap diodes. The width of the depletion zone of any diode changes with voltage applied. This varies the capacitance of the diode.

Electrostatics

For a p-n junction Poisson's equation becomes

$$\Delta\varphi = -\frac{\rho}{\varepsilon} = \frac{q}{\varepsilon}\left(\underbrace{n_0 - p_0}_{\substack{\text{equilibrium concentration}\\ \text{difference of free charges } (\approx 0)}} + \underbrace{N_A - N_D}_{\substack{\text{concentration difference}\\ \text{of acceptor and donor atoms}}}\right)$$

where $\varphi$is the electric potential, ρ is the charge density, $\varepsilon$is permittivity and q is the magnitude of the electron charge.

Since the total charge on either side of the depletion region must cancel out it is

$$\underbrace{d_p}_{\substack{\text{width of}\\ \text{electric field}\\ \text{within p-side}}} \quad N_A = \quad \underbrace{d_n}_{\substack{\text{width of}\\ \text{electric field}\\ \text{within n-side}}} \quad N_D$$

From the above equations and by deploying basic calculus it can be shown that the total width of the depletion region is

$$d = d_p + d_n = \sqrt{\frac{2\varepsilon}{q}\frac{N_A + N_D}{N_A N_D}\left(\underbrace{V_{bi}}_{\text{built-in voltage}} - \underbrace{V}_{\substack{\text{external applied}\\ \text{voltage}}} \right)}$$

Furthermore, by implementing the Einstein relation and assuming the semiconductor is nondegenerate (i.e. the product $p_0 n_0$ is independent of the Fermi energy) it follows that

$$V_{bi} = \frac{kT}{q}\ln\left(\frac{N_A N_D}{p_0 n_0}\right)$$

where T is the temperature of the semiconductor and k is Boltzmann constant.

Summary

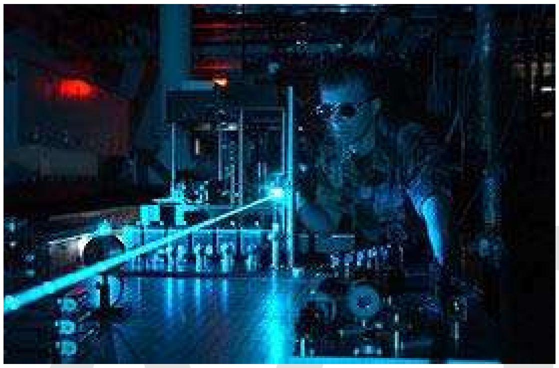
The forward-bias and the reverse-bias properties of the p–n junction imply that it can be used as a diode. A p–n junction diode allows electric charges to flow in one direction, but not in the opposite direction; negative charges (electrons) can easily flow through the junction from n to p but not from p to n and the reverse is true for holes. When the p–n junction is forward biased, electric charge flows freely due to reduced resistance of the p–n junction. When the p–n junction is reverse biased, however, the junction barrier (and therefore resistance) becomes greater and charge flow is minimal.

Non-rectifying junctions

In the above diagrams, contact between the metal wires and the semiconductor material also creates metal-semiconductor junctions called Schottky diodes. In a simplified ideal situation a semiconductor diode would never function, since it would be composed of several diodes connected back-to-front in series. But in practice, surface impurities within the part of the semiconductor which touches the metal terminals will greatly reduce the width of those depletion layers to such an extent that the metal-semiconductor junctions do not act as diodes. These "nonrectifying junctions" behave as ohmic contacts regardless of applied voltage polarity.

Chapter 7

Photoelectrochemical Processes

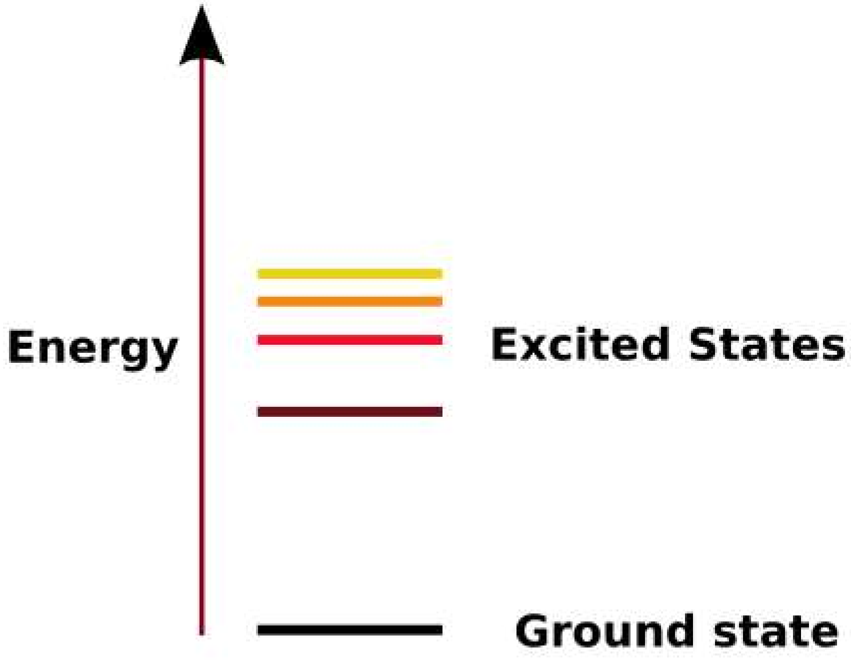


Photons emitted in a coherent beam from a laser

Interaction: Electromagnetic, Optical, Chemical

Photoelectrochemical processes usually involve transforming light into other forms of energy. These processes apply to photochemistry, optically pumped lasers, sensitized solar cells, luminescence, and the effect of reversible change of color upon exposure to light. To the right photons are emitted in a coherent beam from a laser.

Electron excitation



After absorbing energy, an electron may jump from the ground state to a higher energy excited state.

Electron excitation is the movement of an electron to a higher energy state. This can either be done by photoexcitation (PE), where the original electron absorbs the photon and gains all the photon's energy or by electrical excitation (EE), where the original electron absorbs the energy of another, energetic electron. Within a semiconductor crystal lattice, thermal excitation is a process where lattice vibrations provide enough energy to move electrons to a higher energy band. When an excited electron falls back to a lower energy state again, it is called electron relaxation. This can be done by radiation of a photon or giving the energy to a third spectator particle as well.

In physics there is a specific technical definition for energy level which is often associated with an atom being excited to an excited state. The excited state, in general, is in relation to the ground state, where the excited state is at a higher energy level than the ground state.

Photoexcitation

Photoexcitation is the mechanism of electron excitation by photon absorption, when the energy of the photon is too low to cause photoionization. The absorption of photon takes place in accordance to the Planck's Quantum Theory.

Photoexcitation plays role in photoisomerization. Photoexcitation is exploited in dye-sensitized solar cells, photochemistry, luminescence, optically pumped lasers, and in some photochromic applications.

Photoisomerization

In chemistry, **photoisomerization** is molecular behavior in which structural change between isomers is caused by photoexcitation. Both reversible and irreversible photoisomerization reactions exist. However, the word "photoisomerization" usually indicates a reversible process. Photoisomerizable molecules are already put to practical use, for instance, in pigments for rewritable CDs, DVDs, and 3D optical data storage solutions. In addition, recent interest in photoisomerizable molecules has been aimed at molecular devices, such as molecular switches, molecular motors, and molecular electronics.

Photoisomerization behavior can be roughly categorized into several classes: *trans* (or *E*) and *cis* (or *Z*) conversion, and open ring and closed ring transition. Instances of the former include stilbene and azobenzene. This class of compounds has a double bond, and rotation or inversion around the double bond affords isomerization between the two states. Examples of the latter include fulgide and diarylethene. These types of compounds undergo bond cleavage and bond creation upon irradiation with particular wavelengths of light. Sill another type is the Di-pi-methane rearrangement.

Photoionization

Photoionization is the physical process in which an incident photon ejects one or more electrons from an atom, ion or molecule. This is essentially the same process that occurs with the photoelectric effect with metals. In the case of a gas, the term photoionization is more common.

The ejected electrons, known as photoelectrons, carry information about their pre-ionized states. For example, a single electron can have a kinetic energy equal to the energy of the incident photon minus the electron binding energy of the state it left. Photons with energies less than the electron binding energy may be absorbed or scattered but will not photoionize the atom or ion.

For example, to ionize hydrogen, photons need an energy greater than 13.6 electronvolts, which corresponds to a wavelength of 91.2 nm. For photons with greater energy than this, the energy of the emitted photoelectron is given by:

$$\frac{mv^2}{2} = h\nu - 13.6eV$$

where *h* is Planck's constant and *v* is the frequency of the photon.

This formula defines the photoelectric effect.

Not every photon which encounters an atom or ion will photoionize it. The probability of photoionization is related to the photoionization cross-section, which depends on the energy of the photon and the target being considered. For photon energies below the ionization threshold, the photoionization cross-section is near zero. But with the development of pulsed lasers it has become possible to create extremely intense, coherent light where multi-photon ionization may occur. At even higher intensities (around 10^{15} - 10^{16} W/cm^2 of infrared or visible light), non-perturbative phenomena such as *barrier suppression ionization* and *rescattering ionization* are observed.

Multi-photon ionization

Several photons of energy below the ionization threshold may actually combine their energies to ionize an atom. This probability decreases rapidly with the number of photons required, but the development of very intense, pulsed lasers still makes it possible. In the perturbative regime (below about 10^{14} W/cm^2 at optical frequencies), the probability of absorbing *N* photons depends on the laser-light intensity *I* as I^N .

Above-threshold ionization (ATI) is an extension of multi-photon ionization where even more photons are absorbed than actually would be necessary to ionize the atom. The excess energy gives the released electron higher kinetic energy than the usual case of just-above threshold ionization. More precisely, The released electron will have an integer number of photon-energies more kinetic energy than in the normal (lowest possible number of photons) ionization.

Photo-Dember

In semiconductor physics the Photo-Dember effect (named after its discoverer H. Dember) consists in the formation of a charge dipole in the vicinity of a semiconductor surface after ultra-fast photo-generation of charge carriers. The dipole forms owing to the difference of mobilities (or diffusion constants) for holes and electrons which combined with the break of symmetry provided by the surface lead to an effective charge separation in the direction perpendicular to the surface.

Grotthuss–Draper law

The **Grotthuss–Draper law** (also called Principle of Photochemical Activation) states that only that light which is absorbed by a system can bring about a photochemical change. Materials such as dyes and phosphors must be able to absorb "light" at optical frequencies. A basis for Fluorescence and phosphorescence is found in this law. It was first proposed in 1817 by Theodor Grotthuss and John W. Draper. This is considered to be one of the two basic laws of photochemistry. The second law is the Stark–Einstein law, which says that primary chemical or physical reactions occur with each photon absorbed.

Stark–Einstein law

The **Stark–Einstein law** is named after the German-born physicists Johannes Stark and Albert Einstein, who independently formulated the law between 1908 and 1913. It is known also as the **photochemical equivalence law** or **photoequivalence law**. In essence it says that every photon that is absorbed will cause a (primary) chemical or physical reaction.

The photon is a quantum of radiation, or one unit of radiation. Therefore, this is a single unit of EM radiation that is equal to Planck's constant (h) times the frequency of light. This quantity is symbolized by

The photochemical equivalence law is also restated as follows: for every mole of a substance that reacts, an equivalent mole of quanta of light are absorbed. The formula is:

$$\Delta E_{mol} = N_A h\nu$$

where N_A is Avogadro's number.

The photochemical equivalence law applies to the part of a light-induced reaction that is referred to as the primary process (i.e. absorption or fluorescence).

In most photochemical reactions the primary process is usually followed by so-called secondary photochemical processes that are normal interactions between reactants not requiring absorption of light. As a result such reactions do not appear to obey the one quantum–one molecule reactant relationship.

The law is further restricted to conventional photochemical processes using light sources with moderate intensities; high-intensity light sources such as those used in flash photolysis and in laser experiments are known to cause so-called biphotonic processes; i.e., the absorption by a molecule of a substance of two photons of light.

Absorption (electromagnetic radiation)

In physics, **absorption** of electromagnetic radiation is the way by which the energy of a photon is taken up by matter, typically the electrons of an atom. Thus, the electromagnetic energy is transformed to other forms of energy, for example, to heat. The absorption of light during wave propagation is often called attenuation. Usually, the absorption of waves does not depend on their intensity (linear absorption), although in certain conditions (usually, in optics), the medium changes its transparency dependently on the intensity of waves going through, and the Saturable absorption (or nonlinear absorption) occurs.

Photosensitization

Photosensitization is a process of transferring the energy of absorbed light. After absorption, the energy is transferred to the (chosen) reactants. This is part of the work of photochemistry in general. In particular this process is commonly employed where reactions require light sources of certain wavelengths that are not readily available.

For example, mercury absorbs radiation at 1849 and 2537 angstroms, and the source is often high-intensity mercury lamps. It is a commonly used sensitizer. When mercury vapor is mixed with ethylene, and the compound is irradiated with a mercury lamp, this results in the photodecomposition of ethylene to acetylene. This occurs on absorption of light to yield excited state mercury atoms, which are able to transfer this energy to the ethylene molecules, and are in turn deactivated to their initial energy state.

Cadmium; some of the noble gases, for example (usually) xenon; zinc; benzophenone; and a large number of organic dyes, are also used as sensitizers.

Photosensitisers are a key component of photodynamic therapy used to treat cancers.

Sensitizer

A **sensitizer** in chemoluminescence is a chemical compound, capable of light emission after it has received energy from a molecule, which became excited previously in the chemical reaction. A good example is this:

When an alkaline solution of sodium hypochlorite and a concentrated solution of hydrogen peroxide are mixed, a reaction occurs:

$$ClO^-(aq) + H_2O_2(aq) \rightarrow O_2^*(g) + H^+(aq) + Cl^-(aq) + OH^-(aq)$$

$O_2^*$is excited oxygen - meaning, one or more electrons in the O_2 molecule have been promoted to higher-energy molecular orbitals. Hence, oxygen produced by this chemical reaction somehow 'absorbed' the energy released by the reaction and became excited. This energy state is unstable, therefore it will return to the ground state by lowering its energy. It can do that in more than one way:

- it can react further, without any light emission
- it can lose energy without emission, for example, giving off heat to the surroundings or transferring energy to another molecule
- it can emit light

The intensity, duration and color of emitted light depend on quantum and kinetical factors. However, excited molecules are frequently less capable of light emission in terms of brightness and duration when compared to sensitizers. This is because sensitizers can store energy (that is, be excited) for longer periods of time than other excited molecules. The energy is stored through means of quantum vibration, so sensitizers are usually compounds which either include systems of aromatic rings or many conjugated double and triple bonds in their structure. Hence, if an excited molecule transfers its energy to a sensitizer thus exciting it, longer and easier to quantify light emission is often observed.

The color (that is, the wavelength), brightness and duration of emission depend upon the sensitizer used. Usually, for a certain chemical reaction, many different sensitizers can be used.

List of some common sensitizers

- Violanthrone
- Isoviolanthrone
- Fluoresceine
- Rubrene
- 9,10-diphenylanthracene
- Tetracene
- 13,13'-dibenzantronile
- Levulinic Acid

Fluorescence spectroscopy

Fluorescence spectroscopy aka fluorometry or spectrofluorometry, is a type of electromagnetic spectroscopy which analyzes fluorescence from a sample. It involves using a beam of light, usually ultraviolet light, that excites the electrons in molecules of certain compounds and causes them to emit light of a lower energy, typically, but not necessarily, visible light. A complementary technique is absorption spectroscopy.

Devices that measure fluorescence are called fluorometers or fluorimeters.

Absorption spectroscopy

Absorption spectroscopy refers to spectroscopic techniques that measure the absorption of radiation, as a function of frequency or wavelength, due to its interaction with a sample. The sample absorbs energy, i.e., photons, from the radiating field. The intensity of the absorption varies as a function of frequency, and this variation is the absorption spectrum. Absorption spectroscopy is performed across the electromagnetic spectrum.

Chapter 8

Failure Modes of Electronics

Electronic devices have a wide range of failure modes. These can be distinguished by their development in time (sudden failure or gradual degradation), by environmental effects (e.g. corrosion, ionizing radiation) or by the electrical parameter which was exceeded (e.g. electrostatic discharge, overvoltage, overcurrent, etc.).

Failures most commonly occur at the beginning and near the end of the lifetime of the parts. Burn-in procedures are used to detect early failures.

Presence of parasitic structures, irrelevant for normal operation, may become important in the context of failures; such structures can be both a source of failure and a protective device.

A sudden fail-open failure can cause multiple secondary failures, when the event is fast and the circuit contains an inductance. The suddenly interrupted current flow in combination with the inductance then causes large voltage spikes, which for very fast events may exceed 500 volts. A burned metallization on a chip may then cause secondary overvoltage damage.

Packaging-related failures

Electronic packaging, acting as the barrier between the materials of the electronic parts and the environment, is very susceptible to environmental factors. Thermal cycling may cause material fatigue due to mechanical stresses induced by thermal expansion, especially when the thermal expansion coefficients of the materials are different. Humidity or presence of aggressive chemicals can cause corrosion of the packaging materials, leads, or cause failure of encapsulation and following damage to the part inside, leading to electrical failure. Exceeding the allowed environmental temperature, whether too high or too low, can cause overstressing of the wire bonds inside the package, tearing the connections, cracking the semiconductor dies, or causing cracks to the packaging itself. Absorption of humidity into the packaging material and subsequent heating to high temperature (e.g. during soldering) may also cause cracking. Mechanical damage could also fall here.

Majority of failure of electronics parts is typically packaging-related.

Bonding wires can be severed or shorted together during encapsulation, or touch the chip die, usually its edge. Dies can crack due to mechanical overstress or thermal shock; initial defects introduced during e.g. wafer sawing or scribing can develop to fractures later. Lead frame may contain excessive material or burrs, causing shorts. Ions of e.g. alkali metals or halogens can be released from the packaging materials and migrate to the semiconductor dies, causing corrosion or parameter deterioration.

Glass-metal seals commonly fail by forming radial cracks. The cracks originate at the pin-glass interface and continue outwards; other failure causes are weak oxide layer on the pin-glass interface and poor formation of glass meniscus around the pin.

Moisture and other gases may be present in the package cavity, either as impurities trapped during manufacture, or from outgassing of the materials used (monomers, curing agents, etc.), or even from chemical reactions (e.g. when the packaging material gets overheated; the released reaction products, often of ionic nature, then can facilitate corrosion and cause a delayed failure). Helium is often added into the inert atmosphere of the packagings, as a tracer gas to detect leaks during testing. Carbon dioxide may result from oxidation of organic materials with residual oxygen. Hydrogen may be released from some organic materials. Moisture can be outgassed by polymers. Amine-cured epoxies may outgas ammonia.

Formation of cracks and growth of intermetallics in die attachment materials may lead to formation of voids and delamination, impairing the heat transfer from the chip die to the substrate and heatsink, and cause a thermal-related failure.

As some semiconductors, notably silicon and gallium arsenide, are transparent in infrared, infrared microscopy can be used to check the integrity of die bonding and under-die structures.

Red phosphorus, used as a charring-promoter flame retardant, facilitates silver migration when present in device packaging materials. The phosphorus particles are normally coated with aluminium hydroxide. If this coating is incomplete, the phosphorus particles oxidize to phosphorus pentoxide, which is strongly hygroscopic and reacts with moisture to phosphoric acid. Phosphoric acid acts as a corrosive electrolyte, which together with electric fields facilitates dissolution and migration of silver, forming shorts between adjanced packaging pins, lead frame leads, tie bars and chip mount structures, and/or chip pads. The silver bridge may be interrupted with thermal expansion of the package; disappearance of the failure when the chip is heated with a heat gun and its reappearance after several minutes later is an indication of this problem.

Delamination and thermal expansion may move the chip die relative to the packaging, deforming and possibly shorting and/or cracking the bonding wires.

Electrical contacts

- Failures of soldered joints

- Failures of electrical contacts – mechanical faults, corrosion
- Failures of cables – fraying, breaking of the conductors, corrosion, fire damage

Soldered joints, whether on boards, cables, or inside the electronic parts themselves, can fail in many ways; by electromigration, mechanical overstress, formation of brittle intermetallic layers, or material fatigue due to excessive thermal cycling. Such failures can be apparent only at high or low joint temperatures, hindering the debugging.

Thermal expansion mismatch between the part package and the printed circuit board material stresses the part-to-board bonds. Leaded parts are able to absorb the strain by bending. Leadless chip packages rely on the properties of the solder to absorb the stresses. Thermal cycling may lead to fatigue cracking of the solder joints, especially with less plastic solders. Various approaches are used to alleviate the temperature-induced strains.

Loose particles can form in the device cavity; a piece of bonding wire, a fragment of the chip die, flakes of plating, particles of die attachment material, fragments of the case, weld flash, and other materials may migrate inside the packaging cavity and cause shorts, often intermittent and sensitive to mechanical shocks.

Corrosion may cause buildup of oxides and other nonconductive products on the electrical contact surfaces. The contacts, when closed, then show unacceptably high resistance. Corrosion products may migrate and cause shorts.

Tin whiskers can form on the tin-coated metals, e.g. on the internal side of the packagings. Loose whiskers then can cause intermittent short circuits inside the packaging.

Printed circuit boards

Printed circuit boards are vulnerable to environmental influences. The traces are prone to corrosion, the vias can be insufficiently plated-through or insufficiently filled with solder. The traces may be improperly etched, either etched-through entirely, weakened, or insufficiently etched and leaving shorts between traces. The traces may crack under mechanical loads; the thin crack then often causes unreliable circuitboard operation, dependent on the physical warping on the board. Residues of solder flux may facilitate corrosion.

Residues of other materials on the surface of the boards can cause leaks. Polar nonionic compounds can attract water molecules from the atmosphere, forming a thin layer of conductive moist coating between the traces (some antistatic agents act the same way). Ionic compounds, especially chlorides, tend to facilitate corrosion. Alkali metal ions may migrate through plastic packaging and influence the operations of the semiconductors. Chlorinated hydrocarbon residues may hydrolyze, with release of corrosive chlorides; chlorinated solvent residues trapped in the packagings may cause problems years later. Polar molecules may dissipate high-frequency energy, causing dielectric losses.

Above the glass transition temperature of the boards, the resin matrix softens and becomes significantly susceptible to diffusion of contaminants. As an example, polyglycols from the flux can enter the board and increase its humidity intake, with corresponding deterioration of dielectric and corrosion properties.

Multilayer substrates, using ceramics instead of fiber-reinforced polymers, suffer from mostly the same problems.

Conductive anodic filaments (CAF) may grow within the boards, along the fibers of the composite material. The metal is introduced to the vulnerable surface typically from plating the vias, then migrates in presence of ions, moisture, and electrical potential. Drilling damage and poor glass-resin bonding promotes such failures. The formation of CAF usually begins by bonding failure between the glass fiber and the resin matrix, and a layer of adsorbed moisture then provides a channel through which ions and corrosion products migrate. In presence of chloride ions, the precipitated material is atacamite (copper chloride hydroxide); its semiconductive properties then may lead to increased leaks, deteriorated dielectric strength, and short circuits between the traces. Absorbed glycols from the flux residues aggravate the problem. The difference in thermal expansion of the fibers and the matrix also weakens the bond when the board is subjected to high temperature during soldering; the lead-free solders, which generally require higher soldering temperatures, are expected to increase the incidence of CAF. CAF incidence depends on absorbed humidity; below certain threshold it does not occur.

Delamination can occur, separating the board layers, cracking the vias and conductors and introducing pathways for corrosive contaminants and migration of conductive species.

Semiconductors

- Reliability (semiconductor)

Many failures result in generation of large amount of hot carriers in the chip structure, namely hot electrons. These are observable under an optical microscope, as they generate near-infrared photons detectable by e.g. a CCD camera. Latchups can be observed this way.

The location of the failure site, if visible, on the chip die may present clues to the nature of the overstress; whether the site is located at the place with highest current density, highest temperature, the highest electric field gradient, etc., size of the damage, secondary damage (fused leads, cracked die, reflowed die attachment...).

Liquid crystal coatings can be used for localization of faults. Cholesteric liquid crystals respond to temperature, are thermochromic; these are used for visualisation of locations of heat production on the chips. Nematic liquid crystals respond to voltage; these are used for visualising current leaks through oxide defects, and for visualising of charge states on the chip surface, allowing seeing the logical states on the individual structures and conductors.

During laser marking of plastic-encapsulated packages, the laser beam may reach and damage the chip die, if the glass spheres used as fillers in the epoxy resin packaging material line up in such way that they conduct the laser light to the chip.

Parameter-related

- GaAs MMICs:
    - Degradation of I_{DSS}: caused by gate sinking and hydrogen effects ("hydrogen poisoning"). Most common and easiest to detect. Affected by reduction of the active channel of the transistor (gate sinking) or depletion of the donor density in the active channel (hydrogen poisoning).
    - Degradation in gate leakage current: occurs at accelerated life tests or high operation temperatures; suspected to be caused by surface-state effects.
    - Degradation in pinch-off voltage": common failure mode for GaAs devices, operating at high temperature. Primarily results from semiconductor-metal interactions and degradation of gate metal structures. Can be hindered by suitable barrier metal inhibiting diffusion between gold and GaAs. Also can be caused by presence of hydrogen.
    - Increase in drain-to-source resistance: observed in devices operating at high temperature, caused by metal-semiconductor interactions. Caused by gate sinking and ohmic contact degradation.
    - Degradation in RF performance: caused by multiple factors. Surface-state density and material related effects play major roles.

In some cases, the normal presence of tolerances in the circuits can cause erratic behavior difficult to trace. For example a combination of a weak driver transistor with a higher series resistance together with the capacitance of the gate of the subsequent transistor, each within the normal "good" specifications, can significantly increase the propagation delay of the signal. Such faults can manifest only at very specific environmental conditions, high clock speeds in combination with low (but within specifications) power supply voltages, and/or specific circuit signal states. Significant variations can occur on a single die. Overstress induced damage, by e.g. creating ohmic shunts or lowering transistor output current, can potentially increase such delays, leading to erratic behavior of the circuit. As the propagation delays show significant dependence on power supply voltage, normally allowed fluctuations of power supply voltage can trigger such erratic behavior.

Vias are a common source of unwanted serial resistance on chips. Defective vias show higher resistance than they should have and therefore increase propagation delays. As their resistivity drops with increasing temperature, degradation of maximum operating frequency of the chip with decreasing temperature is an indicator of such fault.

Mousebites are regions of partially missing metalization. The conductor is still present but its width is locally decreased. Such defects usually do not show during electrical testing, but present a major reliability risk. The increased current density in the damaged region may exacerbate electromigration problems. A very significant degree of voiding is needed to create a temperature-sensitive propagation delay.

Metalization and wire bonding related

Metalization-related faults are more common and more serious cause of FET transistor degradation than bulk semiconductor material processes. Amorphous materials are promising; the lack of grain boundaries hinders interdiffusion and corrosion.

- Electromigration, caused by high current density can move atoms out of the active regions, leading to emergence of dislocations and point defects, acting as nonradiative recombination centers and producing heat instead of light.
    - Al-gate electromigration in power MESFETs can occur with large RF signals. The current densities in the "fingers" of the gate can be sufficient to cause electromigration, leading to voids and interruptions of the gate fingers and consequent loss of control of drain current. Gold metallization is less susceptible, the issue is therefore limited to aluminium.
    - Drain contacts in power FET transistors are depleted on their end while source contacts get material deposited on them.
    - In structures using aluminium metalization over a refractory metal barrier layer, the electromigration affects primarily the aluminium layer. The underlying refractory metal is highly resistant to electromigration, so the conductor does not fail entirely; its resistance just somewhat, often erratically, increases. The displaced aluminium can however cause shorts to neighbouring structures. Addition of 0.5-4% of copper to the aluminium metal significantly increases resistance to electromigration; copper accumulates on the alloy grain boundaries, increasing the energy needed to dislodge the metal atoms from them.

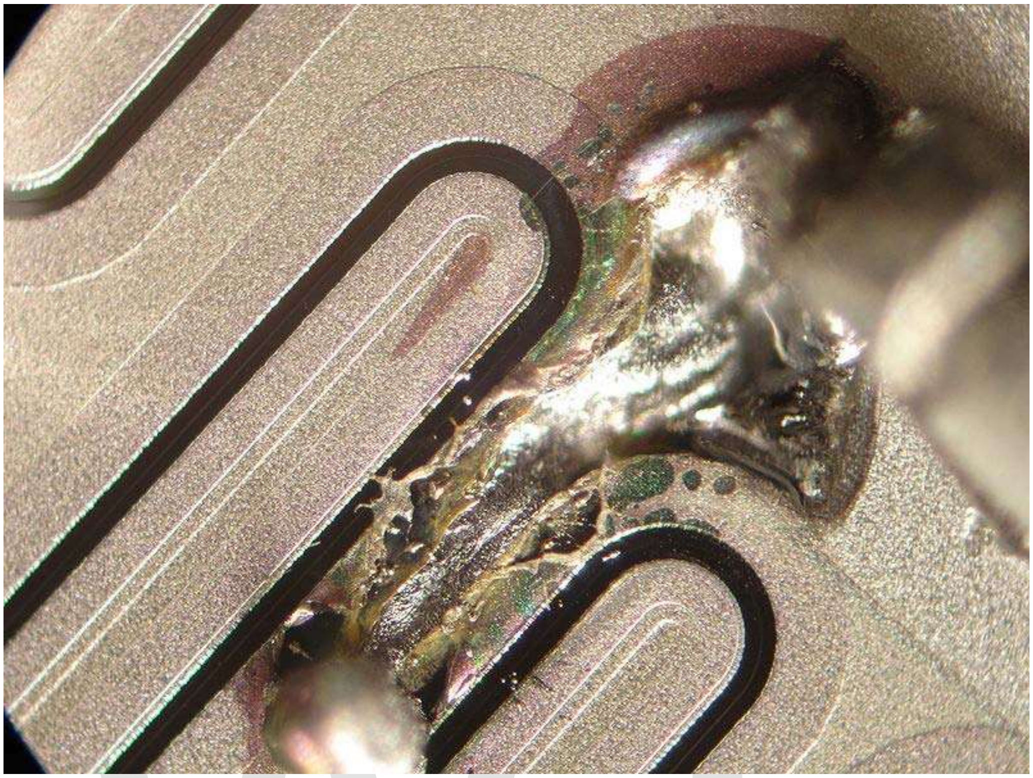
- Metal diffusion caused by high electrical currents or voltages at elevated temperatures can move metal atoms from the electrodes into the active regions. Some materials, notably indium tin oxide and silver, are subject to electromigration which causes leakage current and, in LEDs, nonradiative recombination along the chip edges. A barrier metal layer can be used to hinder the electromigration effects. Metal diffusion can cause changes in dimensions (and therefore parameters) of the transistor gates and other semiconductor junctions. The migration of gate layer in MESFET transistors is known as **gate sinking**; it reduces the dimensions of the active channel and causes change in its effective level of doping, leading to deterioration of electrical parameters.
    - Al/GaAs: Gallium arsenide in contact with aluminium, a common construction of MESFET gates, is susceptible to interdiffusion. Arsenic and especially gallium migrate into the aluminium layer, creating a zone with depleted stoichiometry, and buildup of a $Al_xGa_{1-x}As$ region with

gradient of concentration of aluminium. AlAs crystals are formed. This manifests e.g. as an increase of the Schottky barrier height. Forward current accelerates the process in comparison with thermal-only effect. Ti/GaAs shows similar effects. Aluminium-metalized GaAs devices two decades ago had lifetimes of thousands hours; contemporary TiPtAu metalization has lifetimes reaching millions of hours.

  - Au/GaAs: Other interface layers are used with gold; examples are Au/TiW/GaAs, Au/TiPt/GaAs, and Au/Ta/GaAs. However TiW, while best for high temperatures, has different coefficient of thermal expansion than GaAs, which leads to growth of crystal defects under the metalization and reduced carrier mobility. The industry standard structures for gates on GaAs are based on Au/Pt/Ti or Au/Pd/Ti; Ti serves as a thin interlayer to facilitate adhesion, Pt or Pd is a barrier metal hindering diffusion of Au, and Au is the thick layer conductor. Grain boundaries in the barrier metal can however facilitate increased diffusion of gold into GaAs, leading to gate sinking.

- Ohmic contact degradation. The boundary between a metalization layer and the semiconductor can degrade. In case of GaAs, a layer of gold-germanium-nickel alloy (or gold-germanium alloy) is used to achieve low contact resistance. The ohmic contact is achieved by diffusion of germanium into GaAs, forming a highly n-doped region under the metal that facilitates the connection. A thick layer of gold is then deposited over the thin layer of AuGe. Gallium atoms can migrate through the thin layer and get scavenged by the gold above, creating a defect-rich Ga-depleted zone under the contact. Gold (and oxygen) migrate in the other direction, resulting in increased resistance of the ohmic contact and depletion of effective doping level. Formation of intermetallic compounds also plays a role.

- Short circuits; mechanical stresses, high currents, and corrosive environment can lead to formation of whiskers, causing short circuits. These effects can occur both within the packaging of individual devices and on the level of circuit boards.



Micro-photograph of a failed TO3 power transistor due to short circuit

- Formation of intermetallic compounds, e.g. the well-known gold-aluminium intermetallics (the dreaded white and purple plagues), leading to increased contact resistance and vastly decreased mechanical reliability. This is limited to older devices.

- Formation of silicon nodules. Aluminium interconnects may be doped with silicon to saturation during deposition, to prevent alloy spikes. During thermal cycling, the silicon atoms, originally homogeneously distributed, may migrate and clump together, forming nodules. The nodules act as voids in the metallization, increasing its local resistance and lowering the device lifetime.

- Corrosion of metallization. Aluminium is highly susceptible to humidity; especially negatively biased structures can be interrupted easily. Degradation of aluminium metalization was a common cause of failures of early plastic-encapsulated integrated circuits. Gold is susceptible to anodic corrosion in presence of humidity, forming voluminous conductive gold(III) hydroxide. Nickel can be extruded from metalization in presence of humidity and electric field, forming filaments along the electric field gradient that can short the electrodes. Arsenic can be leached from GaAs in presence of moisture. To cause **dry corrosion** of aluminium, only trace amounts of water and ionic contaminants are required; all plastics are somewhat permeable. Phosphate ions can be leached from phosphosilicate glasses used for passivation of the chips; overcoat with oxide or oxynitride layer can be used in modern processes to create a moisture barrier on top of the phosphate glass layer, and/or a borophosphosilicate glass can be used to lower the phosphate content. Halogenides also rapidly corrode aluminium. Chlorides can be transported through the packaging from the outside. Bromides can be liberated from brominated flame retardants present in the packaging plastics when heated above 250 °C; overheating such package during storage, manufacture or use can increase the part's susceptibility to corrosion in the future.

- Sodium contamination, together with less common lithium and potassium. These ions are mobile in silicon dioxide layers even at normal temperature. (Their counterpart anions in contrast stay immobilized in the oxide structure.) The electric fields present during operation of the semiconductors cause migration of the mobile ions, leading to buildup of charged areas in the gate oxide. The NMOS gates are especially susceptible; a positive gate bias repels the cations towards the junction, depressing its threshold voltage. Even low concentrations of mobile ions can cause shifts by few millivolts, enough to cause trouble for analog circuits. The ion migration is a slow process, causing slow gradual drift of circuit parameters up to a possible failure. This effect was observed especially in early metal-gate CMOS logic circuits. Baking the affected chips at 200 °C for a few hours can temporarily reverse the effect by rediffusing the ions through the oxide. Doping the polysilicon of the base region with phosphorus is an effective way for immobilizing the ions. Alkali ions can migrate from the outside of the package; plastics can hinder their movement but can not slow them entirely. Nitride or phosphate glass layers are used as a chip die protection against externally originated contaminants.

- Very narrow interconnection cracks can maintain functionality by electron tunneling across the crack. Cold temperatures cause contraction of the metal, widening the crack and reducing the tunneling. Devices suffering this failure can run at higher frequencies at higher temperatures; this unusual behavior is symptomatic for an interconnection crack.

- Metalization step coverage, or microcracking, is an insidious unscreenable manufacturing failure. The metallization layer at some places, where their height differs, due to geometrical constraints and deposition technique forms locally weakened sites. The metal on such steps is thinner than required, or even develops a microcrack. The increased current density then leads to other effects leading to interruption of the layer and premature failure of the device.

Semiconductor-related

- Nucleation and growth of dislocations are known mechanisms for degradation of the semiconductor junctions. This requires a presence of an existing defect in the crystal and is accelerated by heat, high current density, and emitted light. In case of LEDs, Gallium arsenide and aluminium gallium arsenide are more susceptible to this mechanism than gallium arsenide phosphide and indium phosphide.

Because of different properties of the active regions, gallium nitride and indium gallium nitride are virtually insensitive to this kind of defect.

- Accumulation of charge carriers trapped in the gate oxide of MOSFETs. This introduces permanent gate biasing, influencing the transistor's threshold voltage. This effect may be caused by hot carriers injection, ionizing radiation (one of the total dose effects), or even nominal use; in case of EEPROM cells and related structures this is the major wear mechanism limiting the number of erase-write cycles.

- Ionizing radiation and neutron radiation have multiple effects, both transient and permanent. They can cause defects in the semiconductor, creating recombination centers and shortening the lifetime of minority carriers, degrading the performance of bipolar junction transistors; it also causes accumulation of charge carriers discussed above. More details are described in the problematics of radiation hardening.

- Migration of charge carriers from floating gates, limiting the lifetime of stored data in EEPROM and flash EPROM structures.

- Improper passivation; corrosion effects are a significant source of delayed failures. Semiconductor materials, metallic interconnects, and passivation glasses are all susceptible to corrosion. The surface of semiconductor, subjected to moisture, develops a layer of oxide; the hydrogen liberated from water then reacts with deeper layers of the material, yielding volatile hydrides.

- Stress Induced Leakage Current is an increase in the gate leakage current of a MOSFET, due to defects created in the gate oxide during electrical stressing.

- Hot carrier injection, occurring in MOSFET transistors. At high field gradients, the charge carriers (electrons in NMOS, holes in PMOS) are accelerated to high speeds. When the transistor operates at saturation with high drain-source voltage, the pinched-off region under the gate (created by the electric field) produces hot carriers near the drain end. Due to lower mobility of holes, the NMOS devices produce hot carriers at lower field gradients than PMOS; e.g. a 3 micrometer NMOS will generate hot electrons at 10 V, while a PMOS of the same size will not generate hot holes below 20 V. Some of the hot carriers, after collision with atoms in the drain region, are deflected into the gate oxide. Most return back, but some become trapped in the oxide imperfections. These then accumulate and cause a persistent gate bias, increasingly shifting the threshold voltage as more hot carriers accumulate. Like with alkali ions, the damage can be partially or fully annealed by heating the chip without power at 200-250 °C for a couple of hours. Avalanche breakdown is also a major source of hot carriers; in case of diffused junction it usually occurs near the surface, where the dopant concentration is highest.

- Parasitic channels can form anywhere where a source-drain structure exists. The role of the gate can be played by a metallization trace above, or even by static charge built up or migrated into the overlaying insulator, protective overcoat, or passivation layer. As the migrating charge carriers are usually electrons, parasitic PMOS transistors tend to be formed. The bipolar chips are more sensitive, likely because of more stringent requirements regarding absence of mobile ionic species for CMOS and BiCMOS technologies.

- Burnt fuses, used for programming the integrated circuits, can under certain conditions reform. Polysilicon fuses may crack instead of vaporizing, which prematurely interrupts the current and leaves enough material to allow reforming the fuse later. Programming both polysilicon and metalization fuses after packaging the circuit prevents the metal from dispersing; the metal stays in the vicinity of the burnt fuse and can migrate back.

- Random access memory chips suffer from several types of failures:
  - Address fault, where the memory address decoder is faulty
  - Stuck-at fault, where a data line somewhere on the chip is shorted to H or L
  - Bridging fault, where two inputs or outputs are shorted together
  - Transition fault
  - n-cell coupling fault
  - Delay fault, where a propagation delay somewhere in the circuit is unacceptably high, causing faulty operation at high speeds
  - Retention fault, where a DRAM cell capacitor does not reliably hold charge for sufficient time

The failures are also classed as bit failures (single, double, triple, quadruple, multiple), row failures (where the entire row of bits fails), column failure (same for the memory array column), cross (where both a row and a column failure occurs), continuous block, peripheral logic, and systemic defect.

Stress-related

Most stress-related failures are electrothermal in nature. The locally increased temperature can lead to immediate failure by melting or vaporizing metalization layers, melting the semiconductor, or creating other structural changes. The diffusion and electromigration effects tend to be accelerated by high temperature, which shortens the lifetime of the device. Damages to junctions that do not lead to immediate failure manifest as altered current-voltage characteristics of the junctions.

Electrical overstress failures can be classified as thermally induced failures, electromigration related failures, and electric field related failures.

- Thermal runaway: Nonhomogenities in the substrate, causing localized loss of thermal conductivity, can cause thermal runaway where heat causes damage which causes more heat etc. Most common ones are voids caused by incomplete soldering, or by electromigration effects and Kirkendall voiding.

- Current crowding, non-homogenous distribution of the current density over the junction, formation of current filaments. This may lead to creation of localized hot spots, which poses risk of thermal runaway.

- Reverse bias: Although e.g. the LED is based on a diode junction and is nominally a rectifier, the reverse-breakdown mode for some types can occur at very low voltages and essentially any excess reverse bias causes immediate degradation, and may lead to vastly accelerated failure. 5 V is a typical, "maximum reverse bias voltage" figure for ordinary LEDs, some special types may have lower limits.

- Overcurrent can cause failures of the bonding wires. In some cases the semiconductor junctions can withstand high enough current to melt the bonding wires.

- Zener diodes in reverse bias, when severely overloaded, fail as a short circuit. A sufficiently high voltage causes an avalanche breakdown of the Zener junction; the voltage across the junction together with a significant current being forced through causes extreme localized heating; the junction and metallization melt, and an alloy of silicon and aluminium shorts the diode's terminals. This is sometimes intentionally used in semiconductors as a type of programming fuses.

- Latchups can occur when the device is subjected to an overvoltage or undervoltage pulse. The opened SCR parasitic structure then can cause an overcurrent-based permanent failure. In integrated circuits, latchups are divided by their cause to internal (transmission line reflections, ground bounces, power supply overshoots) and external (signals injected to the chip via its I/O pins from the outside). External latchups can be triggeded by an electrostatic discharge. Effects of ionizing radiation and cosmic rays are also included in external effects. Susceptibility to latchup is tested according to the JEDEC78 test standard. Latchup can be triggered by charge carriers injected into the chip substrate by e.g. a current flowing through an ESD protection diode, or through another latchup.

- FET transistors are sensitive to dV/dt failures; excessively fast voltage transients can cause the transistor to open.

- Bipolar transistors are more thermally sensitive than FETs; the thermal runaway phenomenons limit their operation margins at higher ambient temperatures. Bipolar transistors are also more sensitive to degradation of die cooling mechanisms (die bond defects, part-heatsink attachment degradation).

Electrostatic discharge

Electrostatic discharge (ESD), a subclass of electrical overstress (EOS), may cause immediate failure of the semiconductor device, a permanent shift of its parameters, or a latent damage causing increased rate of degradation. ESD failure has at least one of three components: localized heat generation, high current density, and high electric field gradient. Currents of several amperes can be present for several hundreds of nanoseconds; the energy deposited to the device structure then causes the damage.

ESD discharge in a real circuit containing capacitances and/or inductances causes a "ringing" waveform, a damped wave with rapidly alternating polarity. Affected junctions are rapidly stressed in alternating forward and reverse polarization.

ESD damage has four basic mechanisms:

- **Oxide rupture/breakdown**, occurring at field strengths above 6–10 MV/cm. Designing the circuit so the oxide layers are protected by junctions with lower avalanche breakdown thresholds than the oxide breakdown voltage can protect against this failure mode.
- Junction damage manifests as increased reverse-bias leakage, up to total shorting.
  - **Junction filamentation**, also known as second or thermal breakdown. Localized overheating causes melting of the silicon; molten silicon has 30 or more times lower resistance; the current through the junction therefore flows in few narrow filaments of high current, resulting in a thermal runaway. In MOSFET devices, the filamentation region is usually near the surface, in the gate-drain overlap region, where the insulator layer serves as a thermal insulation. Devices where the hot spots occur deeper in the structure are less susceptible and are often used as ESD protection structures. Dopants diffuse easily in the molten region; shorter events lead to localized thinning of the base region, resulting in increased leakage, longer events may provide enough time for formation of an ohmic channel across the base, resulting in emitter-collector (or source-drain) short.
  - **Junction spiking**, where metalization is involved; the aluminium metalized devices are more susceptible as the eutectic melting point of aluminium-silicon alloy is 577 °C instead of 1415 °C for silicon. Refractory barrier metals inhibit this effect. When the molten silicon region reaches the metalization, a violent exchange of materials occurs.
- **Metalization and polysilicon burn-out**, where the damage is focused to the conductive and resistive elements – metal and polysilicon interconnects, thin film and diffused resistors. Thin film resistors are the most susceptible. The thermal damage causes localized degradation to destruction of the conductive structure. The critical current densities can be lower in structures surrounded by thermal insulators where the Joule heat can not easily dissipate.
- **Charge injection**, where the hot carriers generated by an avalanche breakdown are injected into the oxide layer. This mechanism is the same as for hot carrier

injection. The result is an increased leakage current from the FETs prebiased by the injected charges in the gate oxide.

For **catastrophic failures**, there are three basic ESD-related mechanisms:

- **Junction burnout**, formation of a conductive path through the junction, shorting it
- **Metalization burnout**, melting or vaporizing part of the metal interconnect, interrupting it

- **Oxide punch-through**, formation of a conductive path through the insulating layer between two conductors or semiconductors; the gate oxides are thinnest and therefore most sensitive. The damaged transistor shows a low-ohmic junction between gate and drain terminals.

ESD can cause a **parametric performance failure**; the device still operates, but its parameters are shifted. The failure may manifest in stress testing. In some cases, the degree of damage can lower over time (so called *cold healing*).

ESD can also cause **latent failures**, which manifest themselves in a delayed fashion and are difficult to impossible to test for. They have these main reasons:

- Damage to insulators: weakening of the insulator structures, leading to accelerated breakdown and/or increased leakage
- Damage to junctions: lowering the lifetime of minority carriers with consequent bipolar transistor gain loss; increasing resistance in forward biased state; increasing leakage in reverse biased state
- Damage to metalization: weakening of the conductor, leading to increased resistance or increased rate of electromigration

Catastrophic failures require the highest discharge voltages; they are easiest to test for, and rarest to occur. Parametric performance failures occur at intermediate discharge voltages and occur more often. Latent failures occur at low voltages and are the most common; for each parametric performance failure there are 4–10 latent ones.

Modern high-integration circuits are more ESD sensitive. The features are smaller, their capacitance is lower and for the same amount of charge the deposited voltage is higher. The silicidation of the conductive layers makes them more conductive, reducing the ballast resistance that can have a degree of protective role.

The gate oxide of some MOSFET transistors can be damaged by as little as 50 volts potential. The gate is isolated from the junction; potential accumulated on it causes extreme stress on the thin dielectric layer. Stressed oxide can shatter and fail immediately (gate rupture – occurs in nanoseconds, does not require a sustained electric current, and is irreversible; usually the gate and backgate of the affected transistor end up connected together). The gate oxide does not have to fail immediately; the gate leakage can increase by stress induced leakage current, the oxide damage can lead to a delayed failure after hundreds or thousands of operation hours. On-chip capacitors using oxide or nitride dielectric are vulnerable to the same kind of damage. Small structures are more vulnerable than large ones by the virtue of their lower capacitance; the same amount of charge carriers will charge the capacitor the structure forms to a higher voltage. All thin layers of dielectric, e.g. the protective oxide layers over emitter regions of transistors or the insulation between two interconnections, are vulnerable; not just the MOS gates. Chips made by processes employing thicker oxide layers are less vulnerable.

The degree of gate oxide damage depends on the size of the gate. Older transistors had larger gate regions with higher capacitance; a discharge of the accumulated potential frequently caused a gate rupture, causing a **hard breakdown**. Newer ultrathin oxide layers are more commonly damaged by a **soft breakdown**; while the damage is still irreversible, its most significant effect is an increase of the noise voltage of the gate, by up to 4 orders of magnitude. Gate oxide breakdown does not always have to lead to failure; gate-to-channel breakdowns usually do not lead to a hard failure, unlike the gate-to-source or gate-to-drain breakdowns.

Gate oxides can also be stressed during manufacture, by so called **antenna effect**; charges introduced during e.g. ion implantation or dry etching accumulate in conductive structures, causing a voltage buildup across the vulnerable dielectrics.

Breakdown of the gate oxide can form a number of different structures with varying voltage-current characteristics. The simplest case is a connection between two regions with the same doping; a resistive path is then formed. In case of opposite doping regions, the resulting structure is a diode. A metal-to-semiconductor connection may create an ohmic contact or a Schottky diode.

The formation of an ohmic path across an insulator or a junction may not always lead to a hard failure. If the resistance of the path is higher than a critical value, only a degradation of parameters occurs.

Interruption of a conductor may also not always lead to a hard failure. Capacitive coupling can still occur between the structures, maintaining partially degraded functionality, reducing signal strength or affecting gate output voltage range. Faults in drain or source connections of CMOS transistors may lead to formation of quasi-memory cells, making the defect manifestation dependent on previous states of the logic circuit. Loss of an output driver can lead to high-impedance output at some circuit states; the load capacitance then can maintain previous state for some time as a dynamic memory cell.

Semiconductor junctions are more robust than thin dielectrics. A sufficiently high voltage leads to an avalanche breakdown. The electric current is usually concentrated to a small area, due to current crowding; a semiconductor heated above certain limit starts losing resistance with increasing temperature and the region concentrates more current in itself, leading to more heat production and a thermal runaway. The extreme current density can migrate the metalization and short the junction, the heat can melt and recrystallize or even shatter the junction. Such damage usually manifests as a shorted junction.

Current-induced failures are more common in bipolar junction devices, where Schottky or pn junctions are predominant. The high power of the discharge (often above 5 kilowatts for less than a microsecond) is capable of melting and vaporizing silicon and metal. Thin-film resistors can have their value altered by a discharge path forming a shunt across part of their length (value decrease) or get part of the layer vaporized (value increase); this can be problematic in precision analog applications where the values are critical.

A junction which suffered an avalanche breakdown typically has increased leakage. Avalanching the base-emitter junction of a NPN transistor permanently lowers its beta.

Device terminals connected only to MOS gates or capacitors are the most vulnerable to the effects of electrostatic discharge. In case both insulated gates and junctions are connected, avalanching usually takes precedence before gate oxide damage. Substrates of chips and large diffusion areas in the semiconductors, e.g. collectors of high-power transistors, are less vulnerable due to their large size and ability to dissipate more energy. Small diffusion areas, e.g. base and emitor regions of small NPN transistors, are more vulnerable.

Newer CMOS output buffers using a lightly doped drain and silicide source/drain are more ESD sensitive. The N-channel driver usually suffers damage, usually in the oxide layer or n+/p well junction. The damage is caused by current crowding during the snapback of the parasitic NPN transistor. In PMOS-NMOS totem-pole structures, the NMOS transistor is virtually always the one damaged.

Typical ESD-related failure distribution for bipolar/Schottky devices is 90% junction burnout and 10% metalization burnout. For MOSFET devices it is 63% metalization burnout and 27% oxide punchthrough. MIL-HDBK-263 offers more complete discussion.

The structure of the junction influences its ESD sensitivity. Corners and defects can lead to current crowding, reducing the damage threshold. Forward-biased junctions are less sensitive than reverse-biased; in the first case the Joule heat is dissipated through the thicker layer of the material while in the latter it is concentrated in the narrow depletion region.

LEDs and lasers grown on sapphire substrate are more susceptible to ESD damage.

In NMOS transistors, the ESD pulse can lead to formation of a metal filament between source and drain, by a phenomenon called *electrothermomigration*. The junction heat can also melt the polysilicon gate, forming polysilicon filaments between gate-source and gate-drain.

The damage can be observed on the I/V curve; undamaged devices show sharp knees, while damaged ones are significantly softened.

Optoelectronics

- List of LED failure modes
- Catastrophic optical damage of semiconductor lasers at high power, when the part is overstressed
- Hydrogen darkening of optical fibers in presence of hydrogen
- Phosphor degradation, in white LEDs, cathode ray tubes, plasma displays, fluorescent lights, etc.
- Corrosion of optical materials; silicate glasses are attacked by moisture. Lenses and optical fibers are prone to deterioration by condensed moisture. Phosphate glasses are more susceptible to corrosion than silica-based glasses, different formulations of glasses can have vastly different.

MEMS

Microelectromechanical systems suffer from specific types of failures.

- Stiction causes the moving parts stick to other surfaces on contact. An external impulse can sometimes release the adhesion and restore functionality. Non-stick coatings, reduction of surface contact area, and increased awareness virtually eliminated the problem in contemporary systems.
- Particle contamination can cause failures by particles lodged between the elements and blocking their movements. Conductive particles may short out circuits, namely electrostatic actuators.
- Wear damages the surfaces; debris, removed from the surfaces during their mutual movements, can be a source of particle contamination.
- Fractures may cause loss of mechanical parts
- Material fatigue may induce cracks in moving structures
- Electrostatic discharge
- Oxide charging, causing electrostatic attraction between parts, may lead to electrostatically mediated stiction
- Dielectric breakdown, causing a short circuit and irreversibly damaging the MEMS structure

Vacuum tubes

- Vacuum tubes and fluorescent lights are susceptible to degradation of hot cathodes, leading to gradual loss of emission of electrons. The vacuum inside the tubes can be compromised by outgassing of materials inside, diffusion of gases through the envelope, or envelope failure. Burnout of the hot cathode filaments leads to a sudden failure. Cold cathode devices tend to be more reliable.
- Multipactor effect
- Phosphor degradation

Passive elements

Resistors

Resistors can fail open (going to infinite resistance), fail short (going to close to zero resistance), or their value can increase or decrease under environmental conditions (e.g. corrosion, material aging) or because of exceeding of performance limits (e.g. overheating).

Thin film resistors are formed from a thin film of a suitable material, e.g. chromium or tantalum nitride.

- Resistor#Failure modes
- The value of certain thin film metal resistors can change when they overheat, due to annealing of their crystalline structure. Extreme overheating may lead to open circuit failure. In integrated circuits, diffused resistors are preferable for applications where high transient currents are to be encountered, e.g. ESD protection, as they are in close contact with the semiconductor substrate which serves as additional heatsinking.
- Mechanical defects from manufacturing can cause intermittent problems or failures. Improperly crimped caps on carbon or metal resistors can become loose and lose contant permanently or intermittently, or the resistor-to-cap interface resistance increase can shift the value of the resistor. Leads to the through-hole resistors are welded to the caps by spark or by butt weld; a faulty weld can become loose. Rough handling can cause such defects or lead to manifestation of latent defects. Thin cracks in the ceramic substrate may cause an open fault, often only annoyingly intermittent.
- Deformation of wire-wound resistors can lead to shorting of adjanced loops of resistive wire, causing partial loss of resistance.
- Operation in oxidizing atmosphere causes oxidation to the outer layer of the resistive wire, reducing its active diameter and increasing its resistance. Operation in reducing atmosphere, e.g. in presence of hydrogen, has the opposite effect.
- Ceramic and carbon resistor cores are prone to cracking under mechanical loads and shocks.
- Under overload, the power dissipated on the resistor can cause heating above the maximum rating, and melting or oxidizing of the resistive element. The protective lacquer or polymer is charred and pyrolyzed, with release of the characteristic smell and possible formation of a conductive path. The resistive element may be interrupted or weakened.
- At high voltages, arcing can occur between parts of the resistor surface, with possible formation of conductive paths or vaporization of the resistive material.
- Damage to insulation layer of the end caps can lead to short between the cap and an underlying circuit board trace.
- Carbon composition resistors, rods of carbon/ceramic composite with embedded leads, may absorb water in humid environments during operation or storage, with resistance changes of as much as 15% up or down.

- Wire-wound resistors are prone to corrosion of the resistive wire. Cap crimping and welds between the wire, caps, and leads are also weak points of this type.
- Metal and carbon film resistors, composed of a thin film of resistive material (typically cut into a spiral) on a (typically) ceramic core, are prone to corrosion and electrolysis of the resistive layer, when the protective layer is penetrated by moisture and contaminants. The electrolytic damage can erode the layer up to an open circuit failure; externally applied voltage is needed for this deterioration mode. Conductive contaminants may form bridges between the loops of the spiral and lead to resistance drop.
- Surface-mount resistors can suffer delamination of their structure where dissimilar materials join, e.g. between the ceramic substrate and the resistive layer, or between the resistive layer and the contact terminations.
- Nichrome thin-film resistors in integrated circuits can be attacked by phosphorus from the passivation glass, which corrodes them and increases their resistance.
- Laser-trimmed resistors may develop instabilities due to the heat damage and microcracks in the resistive layer adjanced to the site of the kerf.
- SMD resistors with silver metallization of termination contacts may suffer open-circuit failure in sulfur-rich environment (e.g. close to sulfur-vulcanized rubber), due to buildup of silver sulfide. Presence of either sulfur dioxide (emitted e.g. by heated rubber) and hydrogen sulfide can be the cause. Conformal coatings do not prevent the failure. The failure is however fairly rare.
- Materials may migrate through the resistor structure and alter resistance or cause shorts. Silver can migrate from the electrodes of thick film chip resistors. Silver atoms were detected as far as 100 micrometers from the electrodes. In addition to lowering the resistance, the silver atoms create nonhomogenities in current distribution and lead to current crowding, instability of characteristics over time, and increase of noise.
- Copper dendrites may grow from copper oxide present in some materials (e.g. from the layer facilitating adhesion of metallization to a ceramic substrate), and bridge the trimming kerf.
- The resistive layer of thick film chip resistors may degrade when subjected to overvoltage or overcurrent.
- The electrodes may crack or corrode.

Potentiometers and trimmers

Potentiometers and trimmers are three-terminal electromechanical parts, containing a resistive path with a wiper contact with adjustable position. All the failure modes listed for resistors apply to these parts too. Additionally, mechanical wear on the wiper and the resistive layer, together with corrosion, surface contamination, and mechanical deformations, may lead to intermittent path-wiper resistance changes. These are especially annoying with audio amplifiers. Many types are not perfectly sealed, and contaminants and moisture may readily enter the parts. An especially common contaminant is the solder flux. Mechanical deformations, with impaired wiper-path contact, can occur by warpage of the part housing during soldering or mechanical stress during mounting. Excess stress on leads can cause substrate cracking and open failure when the crack penetrates the resistive path.

Capacitors

Capacitors are characterized by their capacitance, parasitic resistance in series (equivalent series resistance) and parallel (leakage) – both often frequency-dependent and voltage-dependent, breakdown voltage, and dissipation factor. Structurally, capacitors consist of electrodes separated by the dielectric (sometimes soaked with liquid electrolyte), the leads, and the housing. Deterioration of any of these structures may cause shift of parameters or open or short failure. Short failures and increase of leakage are the most common failure modes of capacitors, followed by open failures.

- Dielectric breakdown occurs due to overvoltage or aging of the dielectric material leading to breakdown voltage falling below the operating voltage; some types of capacitors are "self-healing", the internal arcing vaporizes parts of the electrodes around the failed spot of the dielectric and breaks the contact, others form a conductive pathway through the dielectric, leading to a short or to partial loss of dielectric resistance.
- Electrode materials may migrate across the dielectric, forming conductive paths.
- Leads can be separated from the capacitor by rough handling during storage, assembly or operation; this leads to an open failure. The failure can occur inside the packaging, invisible from the outside but measurable.
- Dissipation factor may increase due to contamination of the capacitor materials, whether from manufacture or by penetration along the leads or through the packaging. Flux and solvent residues are a common source of problems.
- Leakage can result from contaminants forming conductive paths across the capacitor plates or from altering the dielectric parameters. Moisture or solvents can be absorbed by the dielectric if the coating cracks. Cracks in dielectric can also form leakage paths.
- Excessive charging or discharging current may cause partial fusing of the electrodes, leading to open failure or to capacitance shift. Excessive charging or discharging currents may also fuse the leads to the electrodes.
- Partial debonding of leads can cause loss of capacitance in multilayer capacitors.
- Delamination of multilayer capacitors may cause partial capacitance loss and deterioration of other electrical parameters.
- Voids, cracks, pores and other defects may impair the parameters of the dielectric, causing leakages, loss of capacitance, and loss of maximum breakdown voltage.
- Cracks in the dielectric may present pathways for metal migration and arcing.
- Chemical attack on the dielectric may degrade the insulation resistance.
- Thermal shocks, especially during soldering, may cause mechanical defects in the parts, and alter (temporarily or permanently) the properties of the dielectrics.
- Capacitors in low-voltage high-impedance applications generally fail due to lowered dielectric resistance. Capacitors in high voltage and/or low impedance applications tend to fail due to dielectric breakdown.

- Pinholes and voids cause increased leakage in film capacitors. When sufficient current is available, the defect site can be burned away and the capacitor "self-heals". When insufficient current is available, the current can cause migration of electrode material through the defect, gradually decreasing the leakage path resistance.
- Film capacitors are sensitive to excessive ripple voltage; the thin metallization on the dielectric may not sustain high currents, and partially fuse with a loss of capacitance.
- Polymer film dielectrics tend to become brittle with age, especially at increased temperature. Fractures may develop. Moisture may cause degradation of the polymer.
- Corrosion and cracking of the internal leads can cause open failure.
- Overvoltage in a ceramic capacitor may cause an avalanche breakdown in a weak spot of the dielectric, resulting in a thermal runaway. The damaged area may show localized melting and cracking. The energy deposited in the failure point can be high enough to cause the capacitor to explode, create a miniature fireball, and char the encapsulation and possibly the underlying circuitboard.
- Mica capacitors may show intermittent open failures from separation of the electrode plates from the termination at certain temperatures. Reconstituted mica capacitors may fail short due to weaknesses in the dielectric. As mica capacitors are typically used in high voltage applications, the results may be severe.

Electrolytic capacitors

- Aluminium electrolytic capacitors suffer from gradual increase of leakage and equivalent series resistance, and loss of capacitance due to drying out of the electrolyte. Power dissipation due to high ripple currents and high internal resistance cause rise of the internal temperature of the capacitor, which accelerates the deterioration rate, especially when the internal temperature exceeds maximum design temperature. Such capacitors usually fail short.
- Electrolyte contamination, especially with moisture, can lead to corrosion of the electrodes. The deterioration of their area can lead to capacitance loss, loose particles of conductive material can cause shorts.
- Electrolyte may under certain conditions evolve gas. This leads to increased pressure inside the capacitor housing, in extreme but frequent cases leading to an explosion.
  - Capacitor plague is an extreme kind of premature electrolytic capacitor degradation and failure.
- Tantalum capacitors are susceptible to electrical overstress. Exceeding of the maximum surge voltage may permanently degrade the dielectric and even cause open or short failure.
- The most common failure mode of solid tantalum capacitors is an electrical short. The most common manufactured-related cause of this failure is a defect in the dielectric oxide.
- The failure locations may be visible as discolored dielectric, or as locally melted anode.

- Defects in the tantalum chip anode material, e.g. impurities, are transferred to the dielectric oxide layer formed during the capacitor manufacture. These locations then present weak spots that may fail prematurely. Thermal shock e.g. during soldering may generate new defects.
- The thickness of the oxide layer is limited by the size of the powder from which the tantalum slug is made.
- On voltage transients, tantalum capacitors may show a momentary short or increase of leakage current, called "scintillation". If the current through the failure site is small, the site self-heals; if the current is sufficient to heat the site above 400 °C, thermal runaway occurs and the part shorts.

Crystals

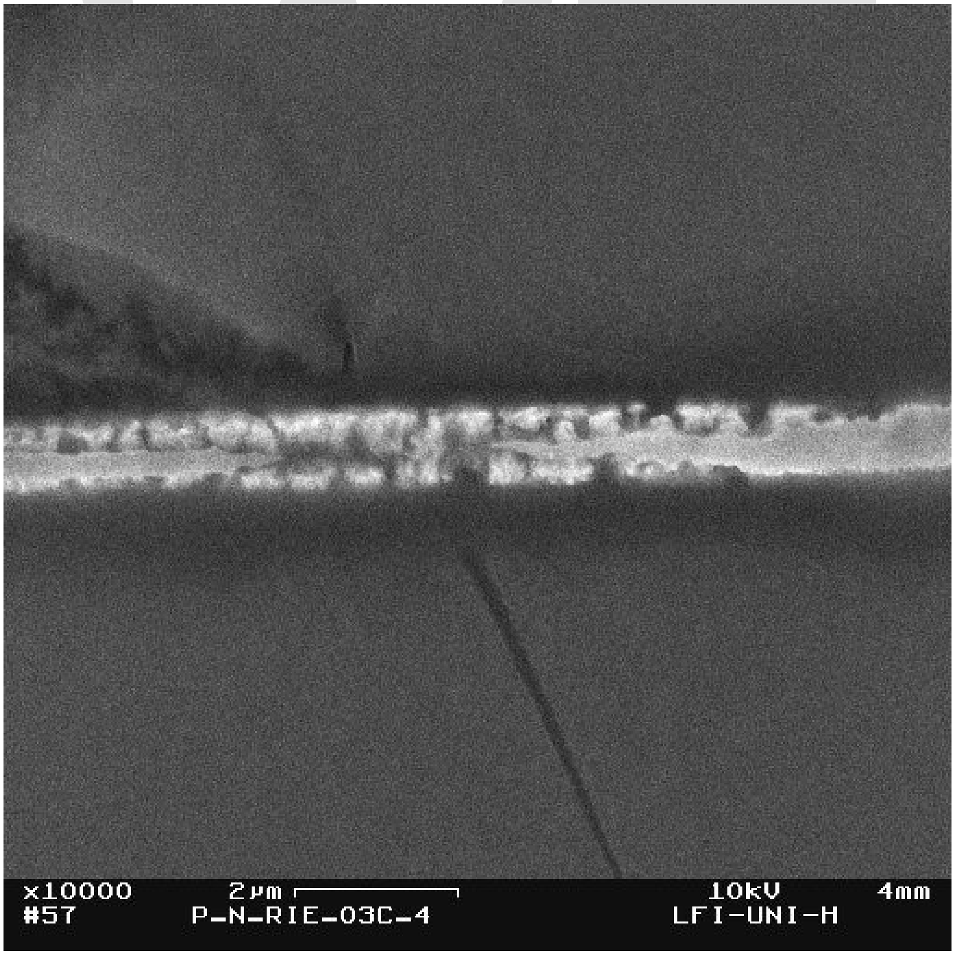
Crystals for crystal oscillators are thin slabs of a piezoelectric material, typically quartz, with deposited electrodes on the slab surfaces, mounted in a hermetically sealed housing, providing the circuit with oscillations at a stable frequency.

- Cracks of crystal slab can occur at extremely high drive levels (especially with low-frequency crystals), or due to mechanical shocks strong enough to make the crystal impacts the inside of the housing.
- Failure of lead connection can lead to an open circuit failure.
- Outgassing of materials inside the housing, most often improperly cured epoxy, can lead to frequency shifts of the crystal.
- Moisture present in the housing can condense on the crystal at low temperatures, causing significant frequency shifts.
- Ionizing radiation and neutron radiation can cause frequency shifts. Swept quartz, from which the mobile alkali metal ions were removed, is used for rad-hard crystals.
- Dehermetization of the housing, by rough handling or corrosion, may lead to penetration of contaminants and moisture into the housing, with consequent changes in resistance and frequency, and possible lead-to-lead and lead-to-packaging leakage paths.

Chapter 9

Electromigration

Electromigration is the transport of material caused by the gradual movement of the ions in a conductor due to the momentum transfer between conducting electrons and diffusing metal atoms. The effect is important in applications where high direct current densities are used, such as in microelectronics and related structures. As the structure size in electronics such as integrated circuits (ICs) decreases, the practical significance of this effect increases.



SEM image of a failure caused by electromigration in a copper interconnect. The passivation has been removed by RIE and HF

History

The phenomenon of electromigration has been known for over 100 years, having been discovered by the French scientist Gerardin. The topic first became of practical interest in 1966 when the first integrated circuits became commercially available. Research in this field was pioneered by a number of investigators throughout the fledgling semiconductor industry. One of the most important engineering studies was performed by Jim Black of Motorola, after whom Black's equation is named. At the time, the metal interconnects in ICs were still about 10 micrometres wide. Currently interconnects are only hundreds to tens of nanometers in width, making research in electromigration increasingly important.

Practical implications of electromigration

Electromigration decreases the reliability of ICs. In the worst case it leads to the eventual loss of one or more connections and intermittent failure of the entire circuit. Since the reliability of interconnects is not only of great interest in the field of space travel and for military purposes but also with civilian applications like for example the anti-lock braking system of cars, high technological and economic values are attached to this effect.

Due to the relatively high life span of interconnects and the short product lifecycle of most consumer ICs, it is not practical to characterize a product's electromigration under real operating conditions. A mathematical equation, the Black's equation, is commonly used to predict the life span of interconnects in integrated circuits tested under "stress", that is external heating and increased current density, and the model's results can be extrapolated to the device's expected life span under real conditions. Such testing is known as high temperature operating life (HTOL) testing.

Although electromigration damage ultimately results in failure of the affected IC, the first symptoms are intermittent glitches, and are quite challenging to diagnose. As some interconnects fail before others, the circuit exhibits seemingly random errors, which may be indistinguishable from other failure mechanisms (such as electrostatic discharge damage). In a laboratory setting, electromigration failure is readily imaged with an electron microscope, as interconnect erosion leaves telltale visual markers on the metal layers of the IC.

With increasing miniaturization the probability of failure due to electromigration increases in VLSI and ULSI circuits because both the power density and the current density increase. In advanced semiconductor manufacturing processes, copper has replaced aluminium as the interconnect material of choice. Despite its greater fragility in the fabrication process, copper is preferred for its superior conductivity. It is also intrinsically less susceptible to electromigration. However, electromigration (EM) continues to be an ever present challenge to device fabrication, and therefore the EM research for copper interconnects is ongoing (though a relatively new field).

A reduction of the structure (scaling) by a factor *k* increases the power density proportional to *k* and the current density increases by k^2 whereby EM is clearly strengthened.

In modern consumer electronic devices, ICs rarely fail due to electromigration effects. This is because proper semiconductor design practices incorporate the effects of electromigration into the IC's layout. Nearly all IC design houses use automated EDA tools to check and correct electromigration problems at the transistor layout-level. When operated within the manufacturer's specified temperature and voltage range, a properly designed IC device is more likely to fail from other (environmental) causes, such as cumulative damage from gamma-ray bombardment.

Nevertheless, there have been documented cases of product failures due to electromigration. In the late 1980s, one line of Western Digital's desktop drives suffered widespread, predictable failure 12–18 months after field usage. Using forensic analysis of the returned bad units, engineers identified improper design-rules in a third-party supplier's IC controller. By replacing the bad component with that of a different supplier, WD was able to correct the flaw, but not before significant damage to the company's reputation.

Electromigration can be a cause of degradation in some power semiconductor devices such as low voltage power MOSFETs, in which the lateral current through the source contact metallisation (often aluminium) can reach the critical current densities during overload conditions. The degradation of the aluminium layer causes an increase in on-state resistance, and can eventually lead to complete failure.

Fundamentals

The material properties of the metal interconnects have a strong influence on the life span. The characteristics are predominantly the composition of the metal alloy and the dimensions of the conductor. The shape of the conductor, the crystallographic orientation of the grains in the metal, procedures for the layer deposition, heat treatment or annealing, characteristics of the passivation and the interface to other materials also affect the durability of the interconnects. There are also grave differences with time dependent current: direct current or different alternating current forms cause different effects.

Forces on ions in an electrical field

Two forces affect ionized atoms in a conductor. The direct electrostatic force F_e as a result from the electric field therefore having the same direction. The force from the exchange of momentum with other charge carriers F_p showing toward the flow of charge carriers. In metallic conductors F_p is caused by a so-called "electron wind" or "Ion wind".

The resulting force F_{res} on an activated ion in the electrical field is

$$F_{res} = F_e - F_p = q \cdot Z^* \cdot E = q \cdot Z^* \cdot j \cdot \rho$$

Electromigration occurs when some of the momentum of a moving electron is transferred to a nearby activated ion. This causes the ion to move from its original position. Over time this force knocks a significant number of atoms far from their original positions. A break or gap can develop in the conducting material, preventing the flow of electricity. In narrow interconnect conductors, such as those linking transistors and other components in integrated circuits, this is known as a **void** or **internal failure** open circuit. Electromigration can also cause the atoms of a conductor to pile up and drift toward other nearby conductors, creating an unintended electrical connection known as a **hillock failure** or **whisker failure** (short circuit). Both of these situations can lead to a malfunction of the circuit.

Failure mechanisms

Diffusion mechanisms

In a homogeneous crystalline structure, because of the uniform lattice structure of the metal ions, there is hardly any momentum transfer between the conduction electrons and the metal ions. However, this symmetry does not exist at the grain boundaries and material interfaces, and so here momentum is transferred much more vigorously. Since the metal ions in these regions are bonded more weakly than in a regular crystal lattice, once the electron wind has reached a certain strength, atoms become separated from the grain boundaries and are transported in the direction of the current. This direction is also influenced by the grain boundary itself, because atoms tend to move along grain boundaries.

Diffusion processes caused by electromigration can be divided into grain boundary diffusion, bulk diffusion and surface diffusion. In general, grain boundary diffusion is the major electromigration process in aluminum wires, whereas surface diffusion is dominant in copper interconnects.

Thermal effects

In an ideal conductor, where atoms are arranged in a perfect lattice structure, the electrons moving through it would experience no collisions and electromigration would not occur. In real conductors, defects in the lattice structure and the random thermal vibration of the atoms about their positions causes electrons to collide with the atoms and scatter, which is the source of electrical resistance. Normally, the amount of momentum imparted by the relatively low-mass electrons is not enough to permanently displace the atoms. However, in high-power situations (such as with the increasing current draw and decreasing wire sizes in modern VLSI microprocessors), if many electrons bombard the atoms with enough force to become significant, this will accelerate the process of electromigration by causing the atoms of the conductor to vibrate further from their ideal lattice positions, increasing the amount of electron scattering. High current density increases the number of electrons scattering against the atoms of the conductor, and hence the speed at which those atoms are displaced.

In integrated circuits, electromigration does not occur in semiconductors directly, but in the metal interconnects deposited onto them.

Electromigration is exacerbated by high current densities and the Joule heating of the conductor, and can lead to eventual failure of electrical components. Localized increase of current density is known as current crowding.

Balance of atom concentration

A governing equation which describes the atom concentration evolution throughout some interconnect segment, is the conventional mass balance (continuity) equation

$$\frac{\partial N}{\partial t} + \nabla \cdot \vec{J} = 0$$

where $N(\vec{x}, t)$ is the atom concentration at the point with a coordinates $\vec{x} = (x, y, z)$ at the moment of time *t*, and *J* is the total atomic flux at this location. The total atomic flux *J* is a combination of the fluxes caused by the different atom migration forces. The major forces are induced by the electric current, and by the gradients of temperature, mechanical stress and concentration. $\vec{J} = \vec{J}_c + \vec{J}_T + \vec{J}_\sigma + \vec{J}_N$. Define the fluxes mentioned above. $\vec{J}_c = \frac{NeZD\rho}{kT}\vec{j}$. Here *e* is the electron charge, *eZ* is the effective charge of the migrating atom, ρ the resistivity of the conductor where atom migration takes place, $\vec{j}$ is the local current density, *k* is Boltzmann's constant, *T* is the absolute temperature. $D(\vec{x}, t)$ is the time and position dependent atom diffusivity. $\vec{J}_T = -\frac{NDQ}{kT^2}\nabla T$. We use *Q* the heat of thermal diffusion. $\vec{J}_\sigma = \frac{ND\Omega}{kT}\nabla H$ here $\Omega = 1 / N_0$ is the atomic volume and N_0 is initial atomic concentration, $H = (\sigma_{11} + \sigma_{22} + \sigma_{33}) / 3$ is the hydrostatic stress and $\sigma_{11}, \sigma_{22}, \sigma_{33}$ are the components of principal stress. $\vec{J}_N = -D\nabla N$.

Assuming a vacancy mechanism for atom diffusion we can express *D* as a function of the hydrostatic stress $D = D_0 \exp(\frac{\Omega H - E_A}{kT})$ where E_A is the effective activation energy of the thermal diffusion of metal atoms. The vacancy concentration represents availability of empty lattice sites, which might be occupied by a migrating atom.

Electromigration-aware design

Electromigration reliability of a wire (Black's equation)

At the end of the 1960s J. R. Black developed an empirical model to estimate the MTTF (mean time to failure) of a wire, taking electromigration into consideration:

$$\text{MTTF} = A(J^{-n})e^{\frac{E_a}{kT}}$$

Here A is a constant based on the cross-sectional area of the interconnect, J is the current density, E_a is the activation energy (e.g. 0.7 eV for grain boundary diffusion in aluminum), k is the Boltzmann's constant, T is the temperature and n a scaling factor (usually set to 2 according to Black). It is clear that **current density *J* and (less so) the temperature *T* are deciding factors** in the design process that affect electromigration.

The temperature of the conductor appears in the exponent, i.e. it strongly affects the MTTF of the interconnect. For an interconnect to remain reliable in rising temperatures, the maximum tolerable current density of the conductor must necessarily decrease.

Wire material

The most common conductor used in integrated circuits is aluminium, due to its good adherence to substrate, good conductivity, and formation of ohmic contacts with silicon. However, it soon appeared that pure aluminium is susceptible to electromigration. Research shown adding 2-4% of copper to aluminium increases resistance to electromigration about 50 times. The effect is attributed to grain boundary segregation of copper, which greatly inhibits the diffusion of aluminium atoms across grain boundaries.

It is known that pure copper used for Cu-metallization is more electromigration-robust than aluminum. Copper wires can withstand approximately five times more current density than aluminum wires while assuming similar reliability requirements. This is mainly due to the higher electromigration activation energy levels of copper, caused by its superior electrical and thermal conductivity as well as its higher melting point. Further improvements can be achieved by alloying copper with about 1% palladium, which, similar to copper in aluminium, inhibits diffusion of copper atoms along grain boundaries.

Bamboo structure and metal slotting

It is obvious that a wider wire results in smaller current density and, hence, less likelihood of electromigration. Also, the metal grain size has influence; the smaller grains, the more grain boundaries and the higher likelihood of electromigration effects. However, if you reduce wire width to below the average grain size of the wire material, the resistance to electromigration increases, despite an increase in current density. This apparent contradiction is caused by the position of the grain boundaries, which in such narrow wires as in a bamboo structure lie perpendicular to the width of the whole wire.

Because the grain boundaries in these so-called “bamboo structures” are at right angles to the current, the boundary diffusion factor is excluded, and material transport is correspondingly reduced.

However, the maximum wire width possible for a bamboo structure is usually too narrow for signal lines of large-magnitude currents in analog circuits or for power supply lines. In these circumstances, slotted wires are often used, whereby rectangular holes are carved in the wires. Here, the widths of the individual metal structures in between the slots lie within the area of a bamboo structure, while the resulting total width of all the metal structures meets power requirements.

Blech length

There is a lower limit for the length of the interconnect that will allow electromigration to occur. It is known as “Blech length”, and any wire that has a length below this limit will not fail by electromigration. Here, a mechanical stress buildup causes a reversed migration process which reduces or even compensates the effective material flow towards the anode. The Blech length must be considered when designing test structures for electromigration.

Via arrangements and corner bends

Particular attention must be paid to vias and contact holes, because generally the ampacity of a (tungsten) via is less than that of a metal wire of the same width. Hence multiple vias are often used, whereby the geometry of the via array is very significant: Multiple vias must be organized such that the resulting current is distributed as evenly as possible through all the vias.

Attention must also be paid to bends in interconnects. In particular, 90-degree corner bends must be avoided, since the current density in such bends is significantly higher than that in oblique angles (e.g., 135 degrees).

Electromigration in Solder Joints

The typical current density at which electromigration occurs in Cu or Al interconnects is 10^6 to 10^7 A/cm^2. For solder joints (SnPb or SnAgCu lead-free) used in IC chips, however, electromigration occurs at much lower current densities, e.g. 10^4 A/cm^2. It causes a net atom transport along the direction of electron flow. The atoms pile up at the anode, voids are generated at the cathode and back stress is induced during electromigration. The typical failure of a solder joint due to electromigration will occur at the cathode side. Due to the current crowding effect, voids form first at the corner of the solder joint. Then the voids extend and cause a failed circuit. Electromigration also influences formation of intermetallic compounds.

Chapter 10

Electrostatic Discharge, Overvoltage & Overcurrent

Electrostatic discharge (**ESD**) is the sudden and momentary electric current that flows between two objects at different electrical potentials. The term is usually used in the electronics and other industries to describe momentary unwanted currents that may cause damage to electronic equipment.

ESD is a serious issue in solid state electronics, such as integrated circuits. Integrated circuits are made from semiconductor materials such as silicon and insulating materials such as silicon dioxide. Either of these materials can suffer permanent damage when subjected to high voltages; as a result, there are now a number of antistatic devices that help prevent static build up.

Causes of ESD

One of the causes of ESD events is static electricity. Static electricity is often generated through tribocharging, the separation of electric charges that occurs when two materials are brought into contact and then separated. Examples of tribocharging include walking on a rug, rubbing plastic comb against dry hair, ascending from a fabric car seat, or removing some types of plastic packaging. In all these cases, the friction between two materials results in tribocharging, thus creating a difference of electrical potential that can lead to an ESD event.

Another cause of **ESD** damage is through electrostatic induction. This occurs when an electrically charged object is placed near a conductive object isolated from ground. The presence of the charged object creates an electrostatic field that causes electrical charges on the surface of the other object to redistribute. Even though the net electrostatic charge of the object has not changed, it now has regions of excess positive and negative charges. An ESD event may occur when the object comes into contact with a conductive path. For example, charged regions on the surfaces of styrofoam cups or plastic bags can induce potential on nearby ESD sensitive components via electrostatic induction and an ESD event may occur if the component is touched with a metallic tool.

Types of ESD

The most spectacular form of ESD is the **spark**, which occurs when a strong electric field creates an ionized conductive channel in air. This can cause minor discomfort to people, severe damage to electronic equipment, and fires and explosions if the air contains combustible gases or particles.

However, many ESD events occur without a visible or audible spark. A person carrying a relatively small electric charge may not feel a discharge that is sufficient to damage sensitive electronic components. Some devices may be damaged by discharges as small as 10 V. These invisible forms of ESD can cause device outright failures, or less obvious forms of degradation that may affect the long term reliability and performance of electronic devices. The degradation in some devices may not become evident until well into their service life.

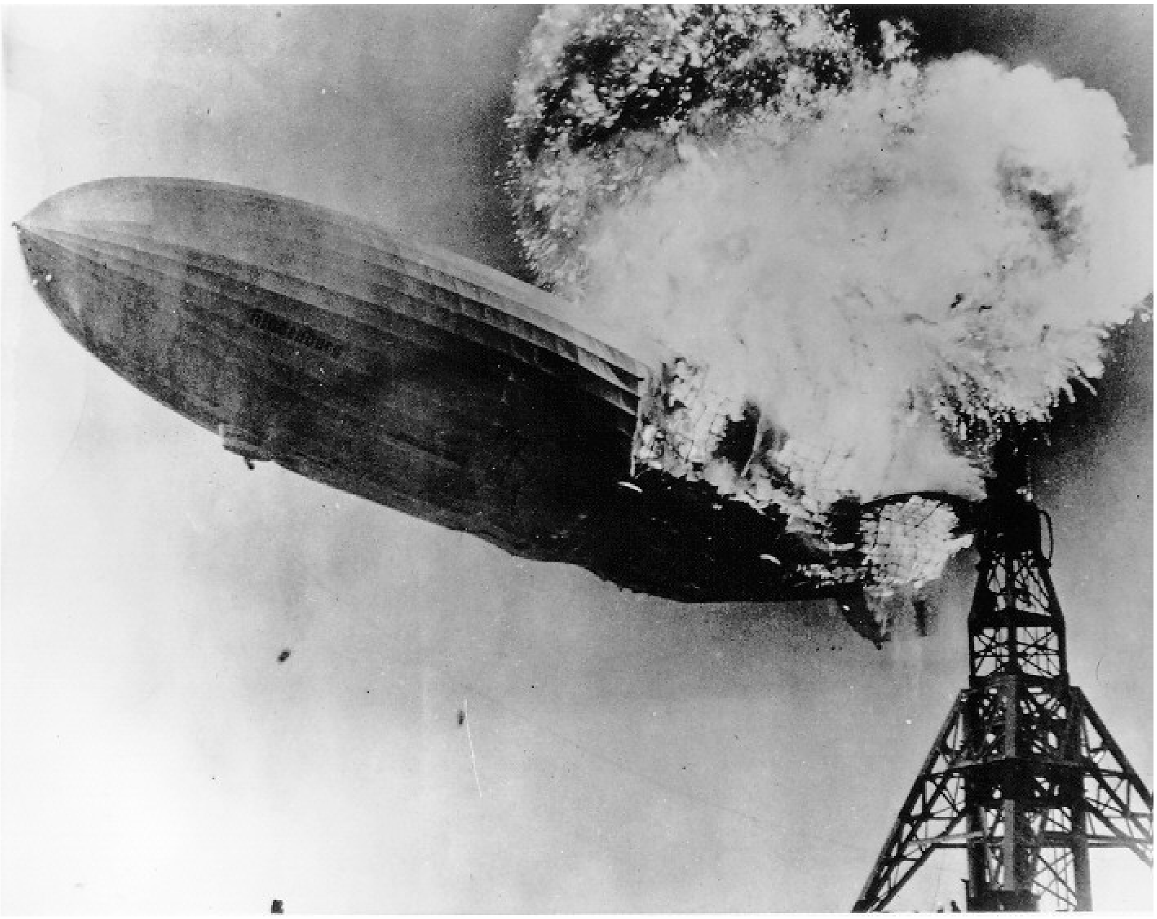
Sparks

A spark is triggered when the electric field strength exceeds approximately 4–30 kV/cm — the dielectric field strength of air. This may cause a very rapid increase in the number of free electrons and ions in the air, temporarily causing the air to abruptly become an electrical conductor in a process called dielectric breakdown.

Lightning over Rymań. Northern Poland

Perhaps the best known example of a natural spark is a lightning strike. In this case the potential difference between a cloud and ground, or between two clouds, is typically hundreds of millions of volts. The resulting current that flows through the ionized air causes an explosive release of energy. On a much smaller scale, sparks can form in air during electrostatic discharges from charged objects that are charged to as little as 380 V (Paschen's law).

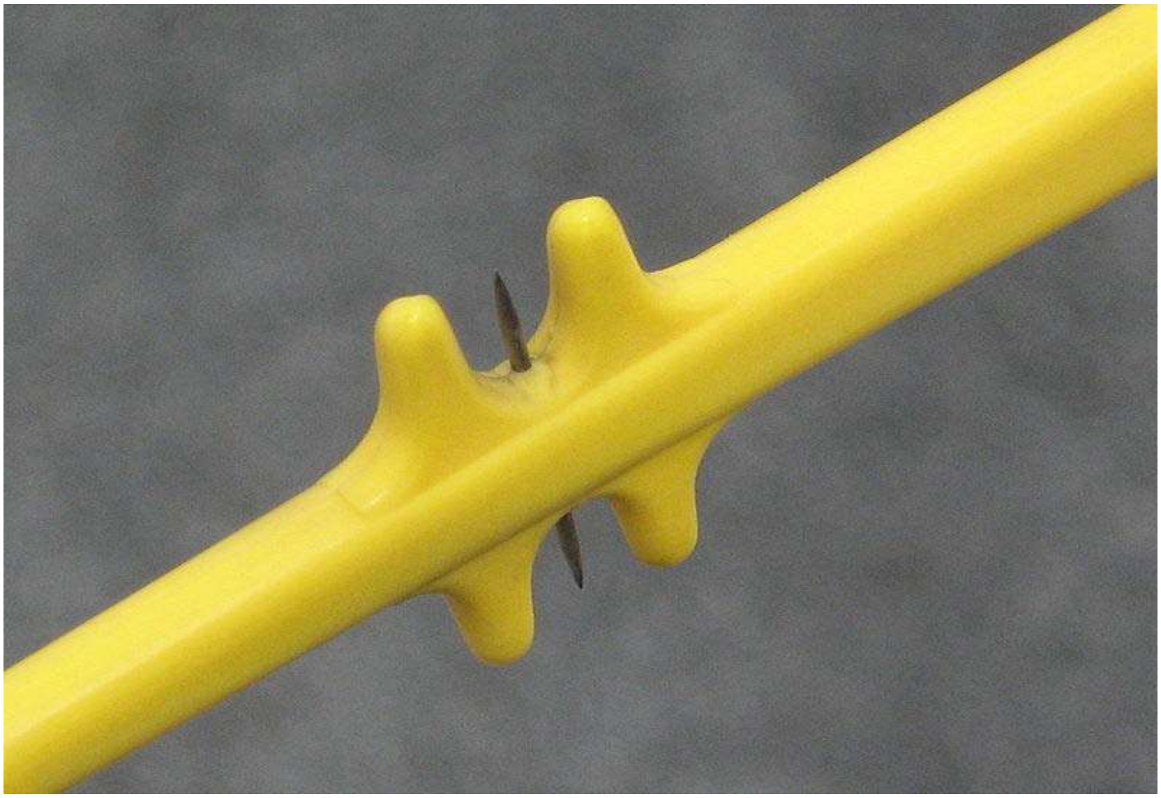
Earth's atmosphere consists of 21% oxygen (O_2) and 78% nitrogen (N_2). During an electrostatic discharge, the intervening atmosphere become electrically overstressed. The diatomic oxygen molecules are split, and then recombine to form ozone (O_3), which is unstable, or reacts with metals and organic matter. If the electrical stress is high enough, nitrogen oxides (NOx) can form. Both products are toxic to animals, and nitrogen oxides are essential for nitrogen fixation. Ozone attacks all organic matter by ozonolysis and is used in water purification.



The *Hindenburg*, moments after catching fire

Sparks can cause serious explosions because of the high temperatures reached in a spark. Methane and coal dust explosions have been caused by electrostatic discharges. The Hindenburg disaster has been attributed to spark discharge igniting flammable panels tainted with thermite, which burned vigorously, violently, and extremely swiftly, which ultimately led to the ignition of hydrogen gas held in or leaking from the airship at the time. The ship had just passed through a thunderstorm and therefore may have acquired a large charge. Discharge supposedly occurred when mooring ropes were dropped as it came in to land in New Jersey in 1937.

Damage prevention in electronics



A portion of a static discharger on an aircraft. Note the two sharp 3/8" metal micropoints and the protective yellow plastic.

Prevention of ESD bases on Electrostatic Protective Area (EPA). EPA can be a small working station or a large manufacturing area. The main principle of an EPA is that there are no highly charging materials in the vicinity of ESD sensitive electronics, all conductive materials are grounded, workers are grounded, and charge build-up on ESD sensitive electronics is prevented. International standards are used to define typical EPA and can be found for example from International Electrotechnical Commission (IEC) or American National Standards Institute (ANSI).

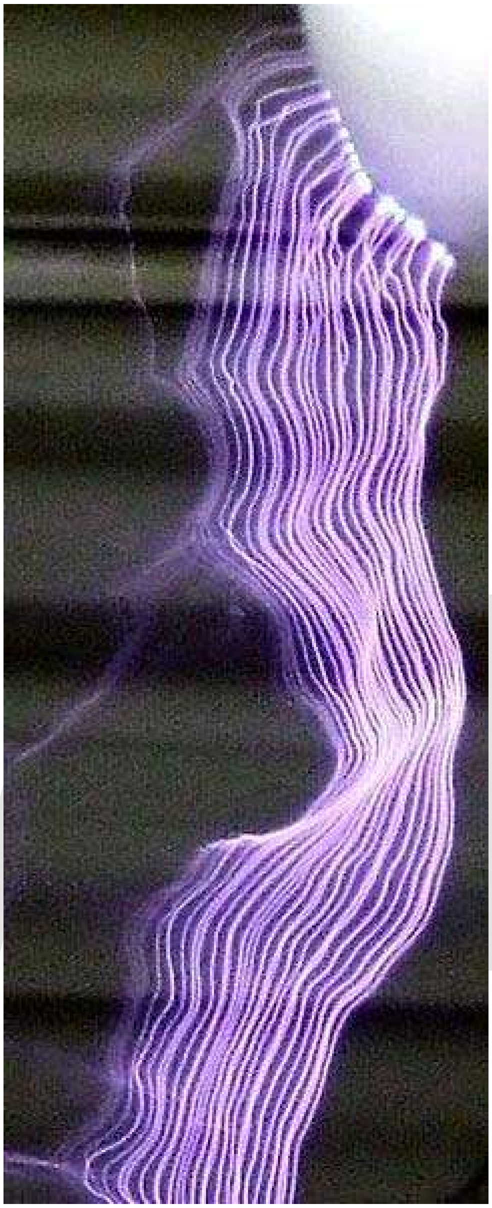
ESD prevention within an EPA may include using appropriate ESD-safe packing material, the use of conductive filaments on garments worn by assembly workers, conducting wrist straps and foot-straps to prevent high voltages from accumulating on workers' bodies, anti-static mats or conductive flooring materials to conduct harmful electric charges away from the work area, and humidity control. Humid conditions prevent electrostatic charge generation because the thin layer of moisture that accumulates on most surfaces serves to dissipate electric charges.

Ion generators are sometimes used to inject ions into the ambient airstream. Ionization systems help to neutralize charged surface regions on insulative or dielectric materials. Insulating materials prone to triboelectric charging should be kept away from sensitive devices to prevent accidental charging of devices through induction. On aircraft, static dischargers are used on the trailing edges of wings and other surfaces.

Manufacturers and users of integrated circuits must take precautions to avoid ESD. ESD prevention can be part of the device itself and include special design techniques for device input and output pins. External protection components can also be used with circuit layout.

Due to dielectric nature of electronics component and assemblies, electrostatic charging can not be completely prevented during handling of devices. Most of ESD sensitive electronic assemblies and components are also so small that manufacturing and handling is made with automated equipment. ESD prevention activities are therefore important with those processes where component is touching on equipment surfaces. In addition, it is important to prevent ESD when electrostatic discharge sensitive component is connected with other conductive parts of the product itself. An efficient way to prevent ESD is to use materials that are not too conductive but will slowly conduct static charges away. These materials are called static dissipative and have resistivity values in the range of 10^5 to 10^{11} ohm-meters. Materials in automated manufacturing which will touch on conductive areas of ESD sensitive electronic should be made of dissipative material, and the dissipative material must be grounded.

Simulation and testing for electronic devices



Electric discharge showing the ribbon-like plasma filaments from multiple discharges from a Tesla coil.

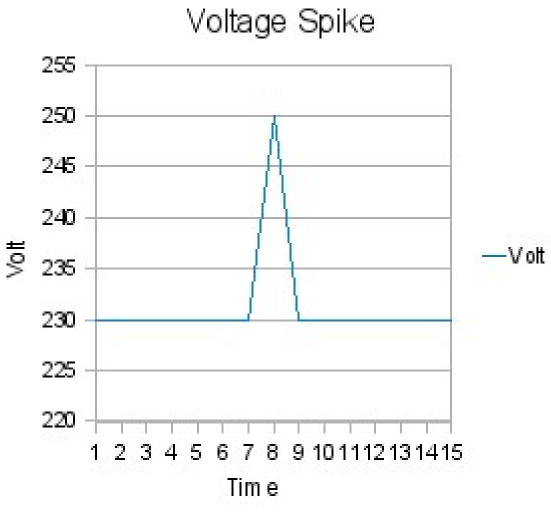
For testing the susceptibility of electronic devices to ESD from human contact, an ESD Simulator with a special output circuit, called the human body model (HBM) is often used. This consists of a capacitor in series with a resistor. The capacitor is charged to a specified high voltage from an external source, and then suddenly discharged through the resistor into an electrical terminal of the device under test. One of the most widely used models is defined in the JEDEC 22-A114-B standard, which specifies a 100 picofarad capacitor and a 1500 ohm resistor. Other similar standards are MIL-STD-883 Method 3015, and the ESD Association's ESD STM5.1. For comportment to European Union standards for Information Technology Equipment, the IEC-61000-4-2 test specification is used. Guidelines and requirements are given for test cell geometries, generator specifications, test levels, discharge rate and waveform, types and points of discharge on the "victim" product, and functional criteria for gauging product survivability.

A Charged Device Model (CDM) test is used to define the ESD a device can withstand when the device itself has an electrostatic charge and discharges due to metal contact. This discharge type is the most common type of ESD in electronic devices and causes most of the ESD damages in their manufacturing. CDM discharge depends mainly on parasitic parameters of the discharge and is strongly depending on size and type of component package. One of the most widely used CDM simulation test models is defined by the JEDEC.

Other standardized ESD test circuits include the following:

- Machine model (MM)
- Transmission line pulse (TLP)

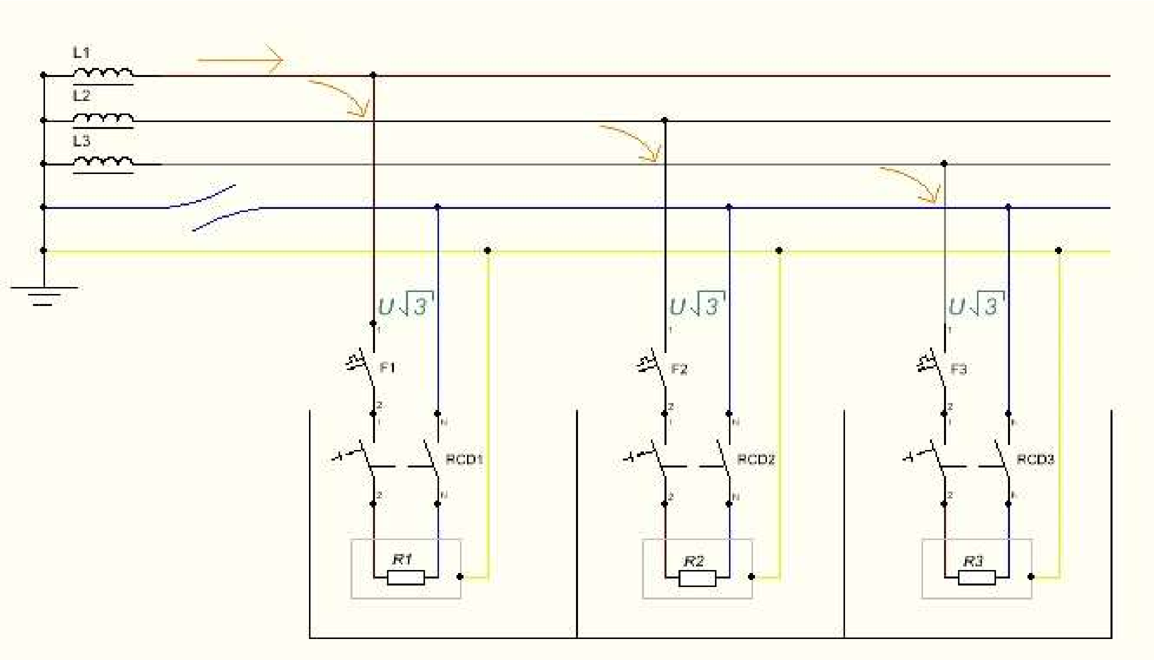
Overvoltage



Voltage spike

When the voltage in a circuit or part of it is raised above its upper design limit, this is known as **overvoltage**. The conditions may be hazardous. Depending on its duration, the overvoltage event can be transient—a voltage spike—or permanent, leading to a power surge.

Explanation



Lack of 3-phase electric system connected by star. If zero breakes off, small-power appliances will be destroyed by overvoltage

Electronic and electrical devices are designed to operate at a certain maximum supply voltage, and considerable damage can be caused by voltage that is higher than that for which the devices are rated.

For example an electric light bulb has a wire in it that at the given rated voltage will carry a current just large enough for the wire to get very hot (giving off light and heat), but not hot enough for it to melt. The amount of current in a circuit depends on the voltage supplied: if the voltage is too high, then the wire may melt and the light bulb would have "burned out". Similarly other electrical devices may stop working, or may even burst into flames if an overvoltage is delivered to the circuit of which these devices are part.

Sources

Natural

A typical natural source of transient overvoltage events is lightning. Bursts of solar wind following solar flares are also known to cause overvoltage in electrical circuits, especially onboard space satellites.

Man made

Man-made sources of spikes are usually caused by electromagnetic induction when switching on or off inductive loads (such as electric motors or electromagnets), or by switching heavy resistive AC loads when zero-crossing circuitry is not used - anywhere a large change of current takes place. One of the purposes of electromagnetic compatibility compliance is to eliminate such sources.

An important potential source of dangerous overvoltage is electronic warfare. There is intensive military research in this field, whose goal is to produce various transient electromagnetic devices designed to generate electromagnetic pulses that will disable an enemy's electronic equipment. A recent military development is that of the exploding capacitor designed to radiate a high voltage electromagnetic pulse. Another intense source of an electromagnetic pulse is a nuclear explosion.

Conduction path

The transient pulses can get into the equipment either by power or data lines, or directly through space from a strong electromagnetic field change - an electromagnetic pulse (EMP). Filters are used to prevent spikes entering or leaving the equipment through wires, and the devices coupled electromagnetically to space (such as radio-frequency pick-up coils in MRI scanners) are protected by shielding.

Overvoltage protection devices

- Arcing horns
- Zener diode
- Avalanche diode
- Transil
- Trisil
- Spark gap
- Gas filled tube
- Metal Oxide Varistor

Overcurrent

In electricity supply, **overcurrent** or **excess current** is a situation where a larger than intended electric current exists through a conductor, leading to excessive generation of heat, and the risk of fire or damage to equipment. Possible causes for overcurrent include short circuits, excessive load, and incorrect design. Fuses, circuit breakers, temperature sensors and current limiters are commonly used protection mechanisms to control the risks of overcurrent.

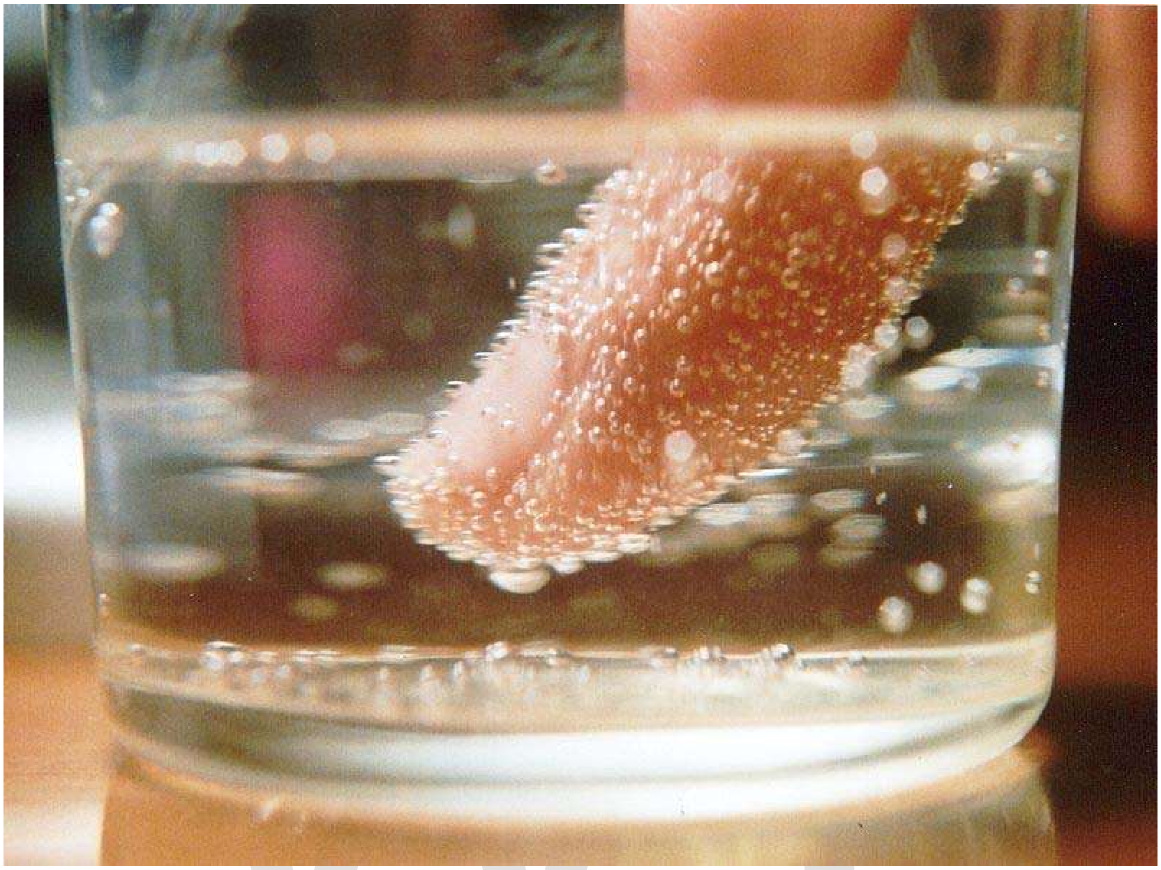
Chapter 11

Semiconductor-Related Failures

Nucleation

Nucleation is the extremely localized budding of a distinct thermodynamic phase. Some examples of phases that may form via nucleation in liquids are gaseous bubbles, crystals or glassy regions. Creation of liquid droplets in saturated vapor is also characterized by nucleation. Nucleation of crystalline, amorphous and even vacancy clusters in solid materials is also important, for example to the semiconductor industry. Most nucleation processes are physical, rather than chemical, but a few exceptions do exist (e.g. electrochemical nucleation). A good example would be the famous Diet Coke and Mentos eruption. Nucleation normally occurs at *nucleation sites* on surfaces contacting the liquid or vapor. Suspended particles or minute bubbles also provide nucleation sites. This is called *heterogeneous nucleation*. Nucleation without preferential nucleation sites is *homogeneous nucleation*. Homogeneous nucleation occurs spontaneously and randomly, but it requires superheating or supercooling of the medium. Nucleation is involved in such processes as cloud seeding and in instruments such as the bubble chamber and the cloud chamber.

Examples of nucleation

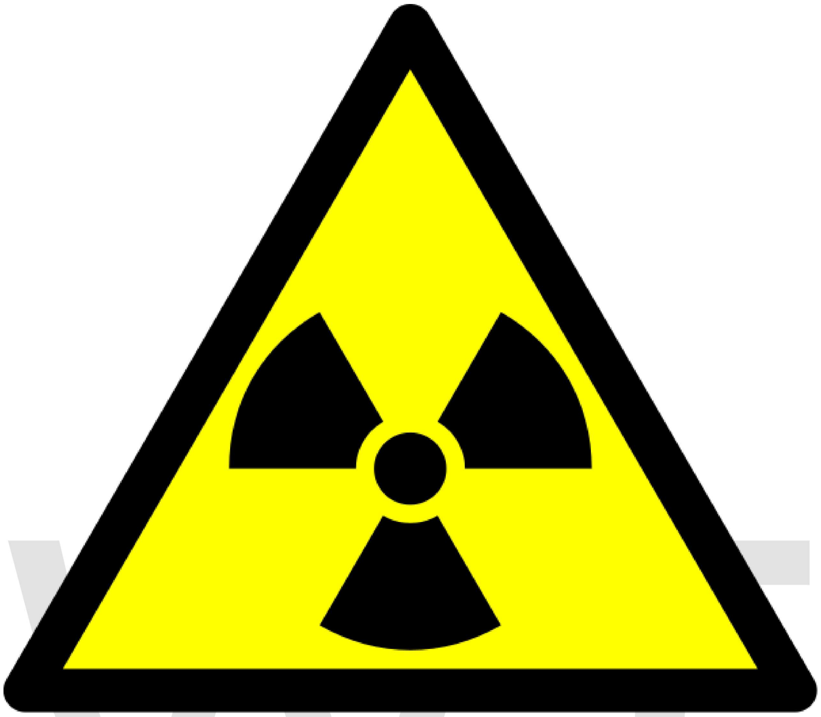


Nucleation of carbon dioxide bubbles around a finger

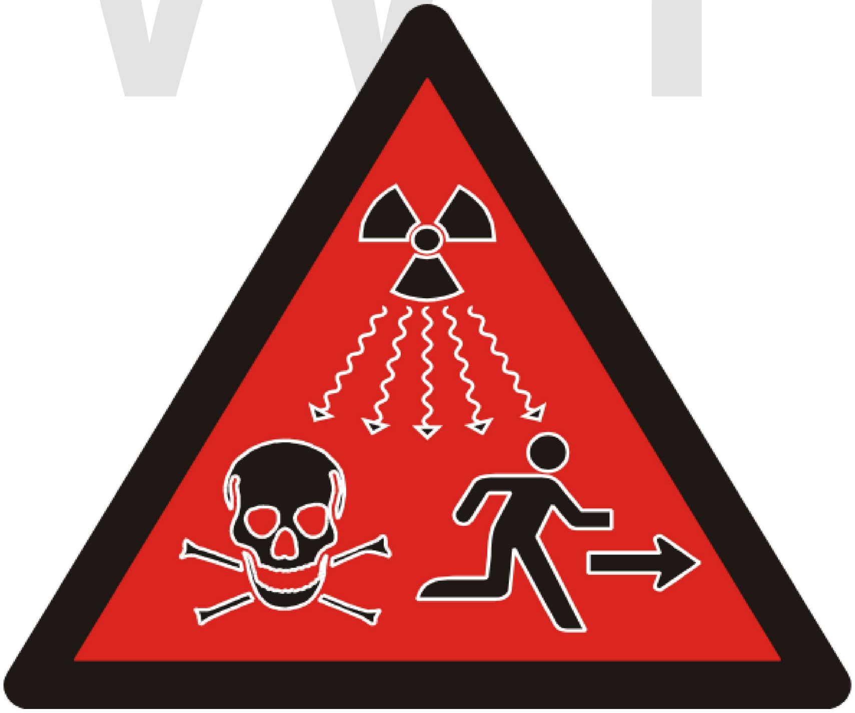
- Pure water freezes at −42°C rather than at its freezing temperature of 0°C if no crystal nuclei, such as dust particles, are present to form an ice nucleus.
- Presence of cloud condensation nuclei is important in meteorology because they are often in short supply in the upper atmosphere.
- All natural and artificial crystallization process (of formation of solid crystals from a homogeneous solution) starts with a *nucleation* event.
- Bubbles of carbon dioxide *nucleate* shortly after the pressure is released from a container of carbonated liquid. Nucleation often occurs more easily at a pre-existing interface (*heterogeneous nucleation*), as happens on boiling chips and string used to make rock candy. So-called Diet Coke and Mentos eruptions are a dramatic example.
- Nucleation in boiling can occur in the bulk liquid if the pressure is reduced so that the liquid becomes superheated with respect to the pressure-dependent boiling point. More often nucleation occurs on the heating surface, at *nucleation sites*. Typically, nucleation sites are tiny crevices where free gas-liquid surface is maintained or spots on the heating surface with lower wetting properties. Substantial superheating of a liquid can be achieved after the liquid is de-gassed and if the heating surfaces are clean, smooth and made of materials well wetted by the liquid.

- Nucleation is relevant in the process of crystallization of nanometer sized materials and plays an important role in atmospheric processes.
- Nucleation is a key concept in polymer, alloy and ceramic systems.
- In chemistry and biophysics, nucleation can also refer to the phaseless formation of multimers which are intermediates in polymerization processes. This sort of process is believed to be the best model for processes such as crystallization and amyloidogenesis.
- In molecular biology, nucleation is used to term the critical stage in the assembly of a polymeric structure, such as a microfilament, at which a small cluster of monomers aggregates in the correct arrangement to initiate rapid polymerization. For instance, two actin molecules bind weakly, but addition of a third stabilizes the complex. This trimer then adds additional molecules and forms a nucleation site. The nucleation site serves the slow, or lag phase of the polymerization process.
- Some champagne stirrers operate by providing many nucleation sites via high surface area and sharp corners, speeding the release of bubbles and removing carbonation from the wine.
- Sodium acetate heating pads use a metal disk as a nucleation centre for exothermic crystallization

Ionizing radiation



Radiation hazard symbol



Ionizing radiation hazard symbol (recently introduced)

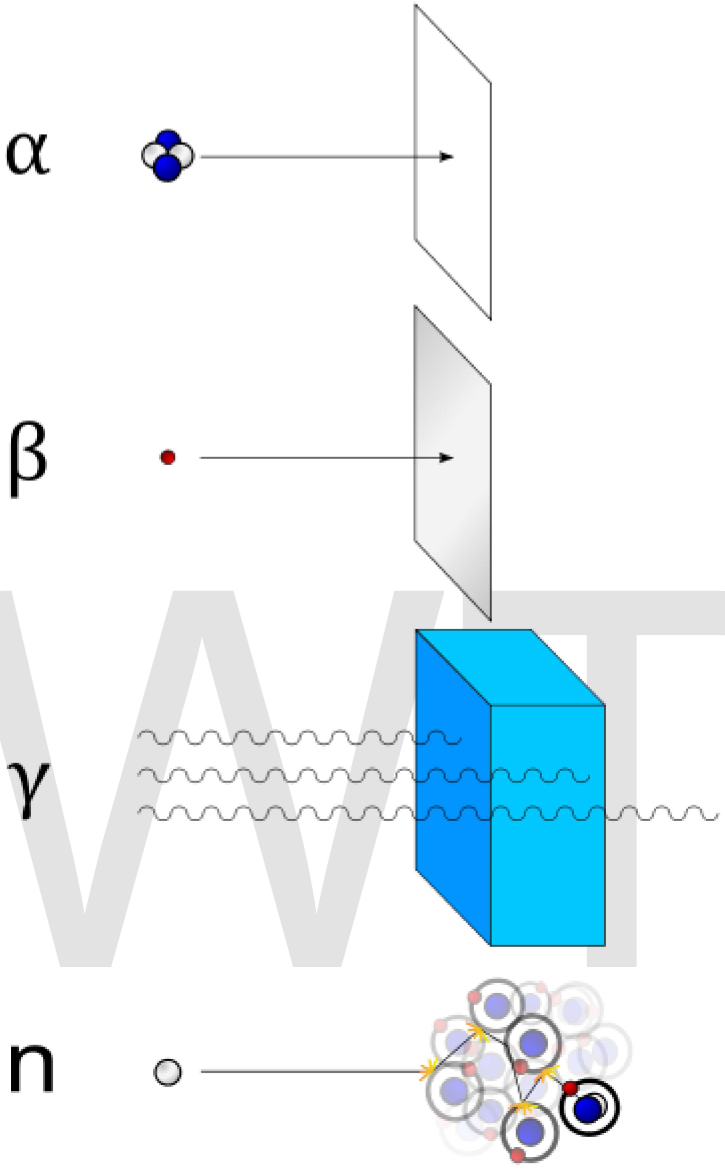
Ionizing radiation consists of particles or electromagnetic waves energetic enough to detach electrons from atoms or molecules, thus ionizing them. The degree and nature of such ionization depends on the energy of the individual particles or waves, and not on their number. An intense flood of particles or waves will not cause ionization if these particles or waves do not carry enough energy to be ionizing. Roughly speaking, particles or photons with energies above a few electron volts (eV) are ionizing. Ionization produces free radicals, which are atoms or molecules containing unpaired electrons, that tend to be especially chemically reactive due to their electronic structure.

Examples of ionizing particles are alpha particles, beta particles, neutrons, and cosmic rays. The ability of an electromagnetic wave (photons) to ionize an atom or molecule depends on its frequency. Radiation on the short-wavelength end of the electromagnetic spectrum—high frequency ultraviolet, x-rays, and gamma rays—is ionizing. Lower-energy radiation, such as visible light, infrared, microwaves, and radio waves, are not ionizing.

The latter types of lower energy electromagnetic radiation may damage molecules, but the effect is generally indistinquishable from the effects of simple heating. Such heating does not produce free radicals until higher temperatures (for example, flame temperatures or "browning" temperatures, and above) are attained. In contrast, damage done by ionizing radiation produces free radicals, even at room temperatures and below, and production of such free radicals is the reason these and other ionizing radiations produce quite different types of chemical effects from (low-temperature) heating. Free radical production is also a primary basis for the particular danger to biological systems of relatively small amounts of ionizing radiation that are far smaller than needed to produce significant heating. Free radicals easily damage DNA, and ionizing radiation may also directly damage DNA by ionizing or breaking DNA molecules.

Ionizing radiation is ubiquitous in the environment, and also comes from radioactive materials, x-ray tubes, and particle accelerators. It is invisible and not directly detectable by human senses, so instruments such as Geiger counters are usually required to detect its presence. In some cases it may lead to secondary emission of visible light upon interaction with matter, as in Cherenkov radiation and radioluminescence. It has many practical uses in medicine, research, construction, and other areas, but presents a health hazard if used improperly. Exposure to radiation causes damage to living tissue, and high doses can result in mutation, radiation sickness, cancer, and death.

Types



Alpha (**α**) radiation consists of a fast moving helium-4 (^{4}He) nucleus and is stopped by a sheet of paper. Beta (**β**) radiation, consisting of electrons, is halted by an aluminium plate. Gamma (**γ**) radiation, consisting of energetic photons, is eventually absorbed as it penetrates a dense material. Neutron (**n**) radiation consists of free neutrons which are blocked using light elements, like hydrogen, which slow and/or capture them.

Various types of ionizing radiation may be produced by radioactive decay, nuclear fission and nuclear fusion, and by particle accelerators and naturally occurring cosmic rays. Muons and many types of mesons (in particular charged pions) are also ionizing.

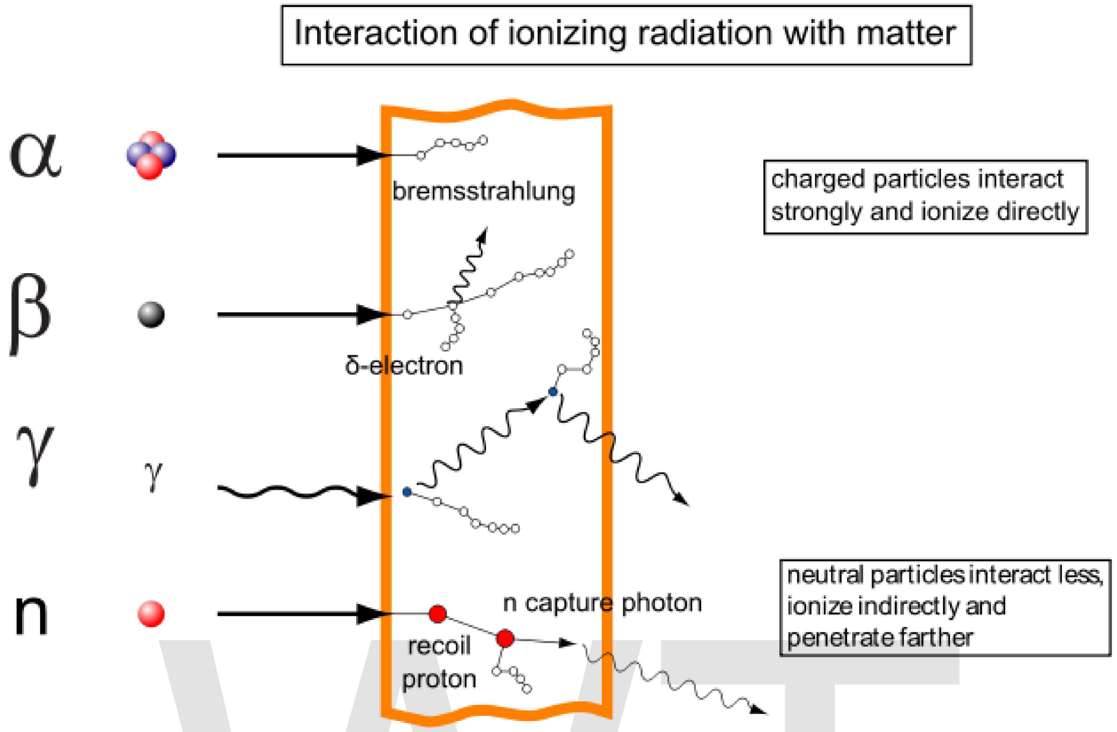
In order for a particle to be ionizing, it must both have a high enough energy and interact with the atoms of a target.

Photons interact electromagnetically with charged particles, so photons of sufficiently high energy also are ionizing. The energy at which this begins to happen with photons (light) is in the high frequency end of the ultraviolet region of the electromagnetic spectrum.

Charged particles such as electrons, positrons, muons, protons, alpha particles, and heavy atomic nuclei from accelerators or cosmic rays also interact electromagnetically with electrons of an atom or molecule. Muons contribute to background radiation due to cosmic rays, but by themelves are thought to be of little hazard importance due to their relatively low dose. Pions (another very short-lived sometimes-charged particle) may be produced in large amounts in the largest particle accelerators, but are not a theoretical biological hazard except near such machines, which are in practice are subject to heavy security while in use.

Neutrons, on the other hand, having zero electrical charge, do not interact electromagnetically with electrons, and so they cannot directly cause ionization by this mechanism. However, fast neutrons will interact with the protons in hydrogen (in the manner of a billiard ball hitting another, head on, sending it away with all of the first ball's energy of motion), and this mechanism produces proton radiation (fast protons). These protons are ionizing because they are charged, and interact with the electrons in matter.

A neutron can also interact with other atomic nuclei, depending on the nucleus and the neutron's velocity; these reactions happen with fast neutrons and slow neutrons, depending on the situation. Neutron interactions in this manner often produce radioactive nuclei, which produce ionizing radiation when they decay, then they can produce chain reactions in the mass that is decaying, sometimes causing a larger effect of ionization.


Interaction of ionizing radiation with matter
α
β
γ
γ
n
bremsstrahlung
δ-electron
n capture photon
recoil
proton
charged particles interact strongly and ionize directly
neutral particles interact less, ionize indirectly and penetrate farther

Types of radiation - gamma rays are represented by wavy lines, charged particles and neutrons by straight lines. The little circles show where ionization processes occur.

An ionization event normally produces a positive atomic ion and an electron. High-energy beta particles may produce bremsstrahlung when passing through matter, or secondary electrons (δ-electrons); both can ionize in turn. Energetic Beta-particles. like those emitted by ^{32}P, are quickly decelerated when passing through matter. The energy lost to deceleration is emitted in the form of X-rays called "Bremsstrahlung" which translates "Braking Radiation". Bremsstrahlung is of concern when shielding beta emitters. The intensity of bremsstrahlung increases with the increase in energy of the electrons or the atomic number of the absorbing medium.

Unlike alpha or beta particles, gamma rays do not ionize all along their path, but rather interact with matter in one of three ways: the photoelectric effect, the Compton effect, and pair production. By way of example, the **figure** shows Compton effect: two Compton scatterings that happen sequentially. In every scattering event, the gamma ray transfers energy to an electron, and it continues on its path in a different direction and with reduced energy.

In the same figure, the neutron collides with a proton of the target material, and then becomes a fast recoil proton that ionizes in turn. At the end of its path, the neutron is captured by a nucleus in an (n,γ)-reaction that leads to a neutron capture photon.

Neutron radiation

Neutron radiation is a kind of non-ionizing radiation which consists of free neutrons. A result of nuclear fission or nuclear fusion, it consists of the release of free neutrons from both stable molecules and isotopes, and these free neutrons react with nuclei of other stable molecules to form new isotopes of previously non-isotopic molecules, which in turn produce radiation. This will result in a chain reaction of nuclear radiation, which makes radiation dangerous and harmful over great areas of space.

Sources

Neutrons may be emitted from nuclear fusion or nuclear fission, or from any number of different nuclear reactions such as from radioactive decay or reactions from particle interactions (such as from cosmic rays or particle accelerators). Large neutron sources are rare, and are usually limited to large-sized devices like nuclear reactors or particle accelerators (such as the Spallation Neutron Source).

Neutron radiation was discovered as a result of observing a beryllium nucleus reacting with an alpha particle thus transforming into a carbon nucleus and emitting a neutron, Be(α, n)C. The combination of an alpha particle emitter and an isotope with a large (α, n) nuclear reaction probability is still a common neutron source.

Uses

Cold, *thermal* and *hot* neutron radiation is most commonly used for scattering and diffraction experiments in order to access the properties and the structure of materials in crystallography, condensed matter physics, biology, solid state chemistry, materials science, geology, mineralogy and related sciences. Neutron radiation is also used in select facilities to treat cancerous tumors due to its highly penetrating and damaging nature to cellular structure. Neutrons can also be used for imaging of industrial parts termed neutron radiography when using film, neutron radioscopy when taking a digital image, such as through image plates, and neutron tomography for three dimensional images. Neutron imaging is commonly used in the nuclear industry, the space and aerospace industry, as well as the high reliability explosives industry.

Ionization mechanisms and properties

Neutron radiation is often called *indirectly ionizing radiation*. It does not ionize atoms in the same way that charged particles such as protons and electrons do (exciting an electron), because neutrons have no charge. However, neutron interactions are largely ionizing, for example when neutron absorption results in gamma emission and the gamma subsequently removes an electron from an atom, or a nucleus recoiling from a neutron interaction is ionized and causes more traditional subsequent ionization in other atoms. Because neutrons are uncharged, they are more penetrating than alpha radiation or beta radiation. In some cases they are more penetrating than gamma radiation, which is impeded in materials of high atomic number. In hydrogen, a low energy neutron may not be as penetrating as a high energy gamma.

Health hazards and protection

In health physics neutron radiation is considered a fourth radiation hazard alongside the other types of radiation. Another, sometimes more severe hazard of neutron radiation, is neutron activation, the ability of neutron radiation to induce radioactivity in most substances it encounters, including the body tissues of the workers themselves. This occurs through the capture of neutrons by atomic nuclei, which are transformed to another nuclide, frequently a radionuclide. This process accounts for much of the radioactive material released by the detonation of a nuclear weapon. It is also a problem in nuclear fission and nuclear fusion installations, as it gradually renders the equipment radioactive; eventually the hardware must be replaced and disposed of as low-level radioactive waste.

Neutron radiation protection relies on radiation shielding. In comparison with conventional ionizing radiation based on photons or charged particles, neutrons are repeatedly bounced and slowed (absorbed) by light nuclei, so a large mass of hydrogen-rich material is needed. Neutrons readily pass through most material, but interact enough to cause biological damage. Due to the high kinetic energy of neutrons, this radiation is considered to be the most severe and dangerous radiation available. The most effective materials are e.g. water, polyethylene, paraffin wax, or concrete, where a considerable amount of water molecules are chemically bound to the cement. The light atoms serve to slow down the neutrons by elastic scattering, so they can then be absorbed by nuclear reactions. However, gamma radiation is often produced in such reactions, so additional shielding has to be provided to absorb it.

Because the neutrons that strike the hydrogen nucleus (proton, or deuteron) impart energy to that nucleus, they in turn will break from their chemical bonds and travel a short distance, before stopping. Those protons and deuterons are high linear energy transfer particles, and are in turn stopped by ionization of the material through which they travel. Consequently, in living tissue, neutrons have a relatively high relative biological effectiveness, and are roughly ten times more effective at causing cancers or LD-50s compared to photon or beta radiation of equivalent radiation exposure.

Effects on materials

Neutrons also degrade materials; bombardment of materials with neutrons creates collision cascades that can produce point defects and dislocations in the materials. At high neutron fluences this can lead to embrittlement of metals and other materials, and to swelling of some of them. This poses a problem for nuclear reactor vessels, and significantly limits their lifetime (which can be somewhat prolonged by controlled annealing of the vessel, reducing the number of the built-up dislocations). Graphite moderator blocks are especially susceptible to this effect, known as Wigner effect, and have to be annealed periodically; the well-known Windscale fire was caused by a mishap during such an annealing operation.

Neutron radiation and nuclear fission

The neutrons in reactors are generally categorized as slow (thermal) neutrons or fast neutrons depending on their energy. Thermal neutrons are similar to a gas in thermodynamic equilibrium but are easily captured by atomic nuclei and are the primary means by which elements undergo atomic transmutation.

In order to achieve an effective fission chain reaction, the neutrons produced during fission must be captured by fissionable nuclei, which then split, releasing more neutrons. In most fission reactor designs, the nuclear fuel is not sufficiently refined to be able to absorb enough fast neutrons to carry on the fission chain reaction, due to the lower cross section for higher-energy neutrons, so a neutron moderator must be introduced to slow the fast neutrons down to thermal velocities to permit sufficient absorption. Common neutron moderators include graphite, light water and heavy water. A few reactors (fast neutron reactors) and all nuclear weapons rely on fast neutrons. This requires certain changes in the design and in the required nuclear fuel. The element beryllium is particularly useful due to its ability to act as a neutron reflector or lens. This allows smaller quantities of fissile material to be used and is a primary technical development that led to the creation of neutron bombs.

Cosmogenic neutrons

Cosmogenic neutrons, neutrons produced from cosmic radiation in the Earth's atmosphere or surface, and those produced in particle accelerators can be significantly higher energy than those encountered in reactors. Most of them activate a nucleus before reaching the ground; a few react with nuclei in the air. The reactions with nitrogen 14 lead to the formation of carbon 14, widely used in radiocarbon dating.

SILC (semiconductors)

Stress Induced Leakage Current (SILC) is an increase in the gate leakage current of a MOSFET, due to defects created in the gate oxide during electrical stressing. SILC is perhaps the largest factor inhibiting device miniturization. Increased leakage is a common failure mode of electronic devices.

Oxide Defects

The most well studied defects assisting in the leakage current are those produced by charge trapping in the oxide. This model provides a point of attack and has stimulated researchers to develop methods to decrease the rate of charge trapping by mechanisms such as N_2O nitridation of the oxide.

Hot carriers injection

Hot carriers injection (**HCI**) is the phenomenon in solid-state or semiconductor electronic devices where either an electron or a “hole” gains sufficient kinetic energy to overcome a potential barrier necessary to break an interface state. The term hot electron comes from the effective temperature used to model carrier density (with fermi dirac function). One should not think that the term "hot" refers to the actual temperature of the MOS transistor. At higher temperatures, the mean free path (distance between two collisions with atoms in the substrate) is shorter, which decreases the energy gain by the carrier. As a result, Hot Carrier Degradation is more important at low temperature.

HCI and CMOS Semiconductor Technology

Semiconductor Physics of HCI

The term “hot carrier injection” usually refers to the effect in MOSFETs, where a carrier is injected from the conducting channel in the silicon substrate to the gate dielectric, which usually is made of silicon dioxide (SiO_2).

To become “hot” and enter the conduction band of SiO_2, an electron must gain a kinetic energy of 3.3 eV. For holes, the valence band offset in this case dictates they must have a kinetic energy of 4.6 eV.

When electrons are accelerated in the channel, they gain energy along the mean free path. This energy is lost in two different ways:

1. The carrier hit an atom in the substrate. Then the collision create a cold carrier and an additional electron-hole pair. In the case of nMOS transistors, additional electrons are collected by the channel and additional holes are evacuated by the substrate.
2. The carrier hit a Si-H bond and break the bond. An interface state is created and the Hydrogen atom is released in the substrate.

The probability to hit either an atom or a Si-H bond is random. And the average energy involved in each process is the same in both case.

This is the reason why the substrate current is monitored during HCI stress. A high substrate current means a large number of created electron-hole pairs and thus an efficient Si-H bond breakage mechanism.

When interface states are created, the threshold voltage is modified and the subthreshold slope is degraded. This lead to lower current, and degrade the operating frequency of integrated circuit.

Scaling and HCI

Advances in semiconductor manufacturing techniques and ever increasing demand for faster and more complex integrated circuits (ICs) have driven the associated Metal–Oxide–Semiconductor field-effect transistor (MOSFET) to scale to smaller dimensions.

However, it has not been possible to scale the supply voltage used to operate these ICs proportionately due to factors such as compatibility with previous generation circuits, noise margin, power and delay requirements, and non-scaling of threshold voltage, subthreshold slope, and parasitic capacitance.

As a result internal electric fields increase in aggressively scaled MOSFETs, which comes with the additional benefit of increased carrier velocities (up to velocity saturation), and hence increased switching speed, but also presents a major reliability problem for the long term operation of these devices, as high fields induce hot carrier injection which affects device reliability.

Large electric fields in MOSFETs imply the presence of high-energy carriers, referred to as "**hot carriers**". These hot carriers that have sufficiently high energies and momenta to allow them to be injected from the semiconductor into the surrounding dielectric films such as the gate and sidewall oxides as well as the buried oxide in the case of silicon on insulator (SOI) MOSFETs.

CMOS Reliability Impact of HCI

The presence of such mobile carriers in the oxides triggers numerous physical damage processes that can drastically change the device characteristics over prolonged periods. The accumulation of damage can eventually cause the circuit to fail as key parameters such as threshold voltage shift due to such damage. The accumulation of damage resulting degradation in device behavior due to hot carrier injection is called "**hot carrier degradation**".

The useful life-time of circuits and integrated circuits based on such a MOS device are thus affected by the life-time of the MOS device itself. To assure that integrated circuits manufactured with minimal geometry devices will not have their useful life impaired, the life-time of the component MOS devices must have their HCI degradation well understood. Failure to accurately characterize HCI life-time effects can ultimately affect business costs such as warranty and support costs and impact marketing and sales promises for a foundry or IC manufacturer.

Relationship to Radiation Effects

Hot carrier degradation is fundamentally same as the ionization radiation effect known as the total dose damage to semiconductors, as experienced in space systems due to solar proton, electron, X-ray and gamma ray exposure.

HCI and NOR Flash Memory Cells

HCI is the basis of operation for a number of non-volatile memory technologies such as Electrically Erasable Programmable Read-Only Memory (EEPROM) cells. As soon as the potential detrimental influence of HC injection on the circuit reliability was recognized, several fabrication strategies were devised to reduce it without compromising the circuit performance.

NOR flash memory exploits the principle of hot carriers injection by deliberately injecting carriers across the gate oxide to charge the floating gate. This charge alters the MOS transistor threshold voltage to represent a logic '0' state. An uncharged floating gate represents a '1' state. Erasing the NOR Flash memory cell removes stored charge through the process of Fowler–Nordheim tunneling.

Because of the damage to the oxide caused by normal NOR Flash operation, HCI damage is one of the factors that cause the number of write-erase cycles to be limited. Because the ability to hold charge and the formation of damage traps in the oxide affects the ability to have distinct '1' and '0' charge states, HCI damage results in the closing of the non-volatile memory logic margin window over time. The number of write-erase cycles at which '1' and '0' can no longer be distinguished defines the endurance of a non-volatile memory.

Chapter 12

Stress-Related Failures

Most stress-related failures are electrothermal in nature. The locally increased temperature can lead to immediate failure by melting or vaporizing metalization layers, melting the semiconductor, or creating other structural changes. The diffusion and electromigration effects tend to be accelerated by high temperature, which shortens the lifetime of the device. Damages to junctions that do not lead to immediate failure manifest as altered current-voltage characteristics of the junctions.

Electrical overstress failures can be classified as thermally induced failures, electromigration related failures, and electric field related failures.

Thermal runaway

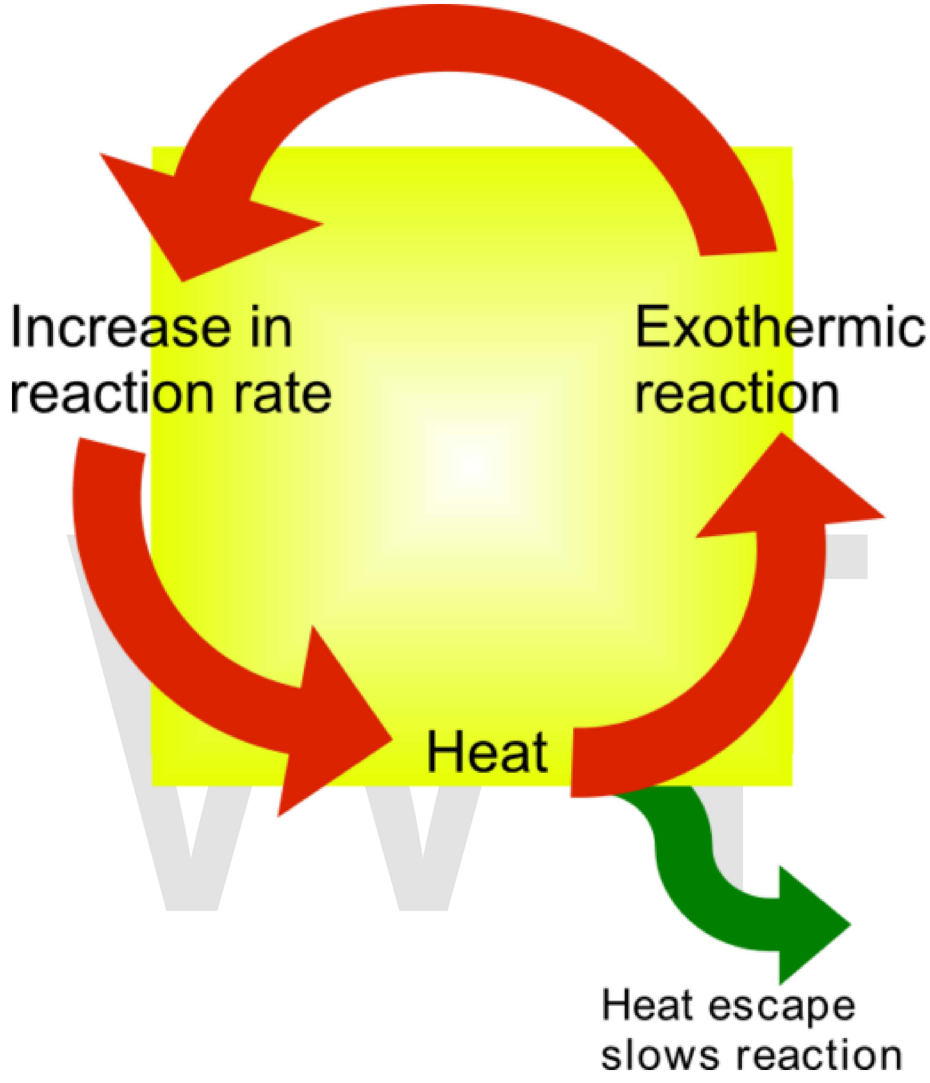


Diagram of thermal runaway

Thermal runaway refers to a situation where an increase in temperature changes the conditions in a way that causes a further increase in temperature, leading (in the normal case of an exothermic reaction) to a destructive result. It is a kind of positive feedback.

Chemical engineering

In chemical engineering, thermal runaway is a process by which an exothermic reaction goes out of control, often resulting in an explosion. It is also known as a "runaway reaction" in organic chemistry.

Thermal runaway occurs when the reaction rate increases due to an increase in temperature, causing a further increase in temperature and hence a further increase in the reaction rate. It has contributed to industrial chemical accidents, most notably the 1947 Texas City disaster from overheated ammonium nitrate in a ship's hold, and the disastrous release of a large volume of methyl isocyanate gas from a Union Carbide plant in Bhopal, India in 1984. Thermal runaway is also a concern in hydrocracking, an oil refinery process. Thermal runaway may result from exothermic side reaction(s) that begin at higher temperatures, following an initial accidental overheating of the reaction mixture. This scenario was behind the Seveso disaster, where thermal runaway heated a reaction to temperatures such that in addition to the intended 2,4,5-trichlorophenol, poisonous 2,3,7,8-tetrachlorodibenzo-p-dioxin was also produced, and was vented into the environment after the reactor's rupture disk burst.

Thermal runaway is most often caused by failure of the reactor vessel's cooling system. Failure of the mixer can result in localized heating, which initiates thermal runaway. Similarly, in flow reactors, localized insufficient mixing causes hotspots to form, where thermal runaway conditions occur, which causes blowouts of reactor contents and catalysts. Incorrect component installation is also a common cause. Many chemical production facilities are designed with high-volume emergency venting to limit the extent of injury and property damage when such accidents occur.

Some laboratory reactions must be run under extreme cooling, because they are prone to hazardous thermal runaway. For example, in Swern oxidation, the formation of the sulfonium chloride must be performed in a cooled system (–30 °C), because at room temperature the reaction undergoes thermal runaway explosively.

The UK Chemical Reaction Hazards Forum publishes previously unreported chemical accidents to assist the education of the scientific community, with the aim of preventing similar occurrences elsewhere. Almost 150 such reports are available to view at the present time (Jan 2009).

Microwave heating

Microwaves are used for heating of various materials in cooking and various industrial processes. The rate of heating of the material depends on the energy absorption, which depends on the dielectric constant of the material. The dependence of dielectric constant on temperature varies for different materials; some materials display significant increase with increasing temperature. This behavior, when the material gets exposed to microwaves, leads to selective local overheating, as the warmer areas are better able to accept further energy than the colder areas—potentially dangerous especially for thermal insulators, where the heat exchange between the hot spots and the rest of the material is slow. These materials are called *thermal runaway materials*. This phenomenon occurs in some ceramics.

Semiconductors

Silicon shows a peculiar instability. Its electrical resistance increases with temperature up to about 160 °C, then starts decreasing, and drops further when the melting point is reached. This leads to thermal runaway phenomena within the semiconductor junction areas; the resistance of the area which becomes hot above this threshold decreases, more of current flows through it causing yet more heating in comparison with the surrounding areas, which leads to further temperature increase and resistance decrease. This leads to the phenomenon of current crowding and formation of current filaments, and is one of the underlying principles behind many semiconductor junction failure mechanisms.

Bipolar transistors

Leakage current increases significantly in bipolar transistors (notably germanium-based bipolar transistors) as they increase in temperature. Depending on the design of the circuit, this increase in leakage current can increase the current flowing through the transistor and with it the power dissipation. This causes a further increase in C–E current. This is frequently seen in a push–pull stage of a class AB amplifier. If the transistors are biased to have minimal crossover distortion at room temperature, and the biasing is not made temperature dependent, as the temperature rises, both transistors will be increasingly turned on, causing current and power to further increase, eventually destroying one or both devices.

To avoid thermal runaway the operating point of BJT should be $V_{ce} \leq 1/2V_{cc}$

The proper practice is to mount a thermal feedback transistor or other device on the heat sink, which generates the crossover bias voltage, so that as the output transistors heat up, so does the thermal feedback transistor, causing the thermal feedback transistor to also start to turn on at a slightly lower voltage, and thus reduce the crossover bias, and reduce the heat released by the output transistors. It can be surprising what professional sound equipment does not have thermal feedback and would thus need modification.

If multiple bipolar transistors are connected in parallel (which is typical in high current applications) one device will enter thermal runaway first, taking the current which originally was distributed across all the devices and exacerbating the problem. This effect is called ***current hogging***. Eventually one of two things will happen, either the circuit will stabilize or the transistor in thermal runaway will be destroyed by the heat. Hence current hogging term is related to thermal runaway.

Power MOSFETs

Power MOSFETs display increase of the on-resistance with temperature. Power dissipated in this resistance causes more heating of the junction, which further increases the junction temperature, in a positive feedback loop. (However, the increase of on-resistance with temperature helps balance current across multiple MOSFETs connected in parallel and current hogging does not occur). If the transistor produces more heat than the heatsink can dissipate, the thermal runaway happens and destroys the transistor. This problem can be alleviated to a degree by lowering the thermal resistance between the transistor die and the heatsink.

- Java applet demo of MOSFET thermal runaway

Digital electronics

The leakage currents of transistors increase with temperature. In rare instances, this may lead to thermal runaway in digital electronics. This is not a common problem, since leakage currents make up a small portion of overall power consumption, so the increase in power is fairly modest - for an Athlon 64, the power dissipation increases by about 10% for every 30 degrees Celsius. For a device with a TDP of 100 W, for thermal runaway to occur, the heat sink would have to have a thermal resistivity of over 3 K/W (kelvins per watt), which is about 6 times worse than a stock Athlon 64 heat sink. (A stock Athlon 64 heat sink is rated at 0.34 K/W, although the actual thermal resistance to the environment is somewhat higher, due to the thermal boundary between processor and heatsink, rising temperatures in the case, and other thermal resistances..) A heat sink with a thermal resistance of over 0.5 to 1 K/W would result in the destruction of a 100 W device even without thermal runaway effects.

Electronics and tropical environments

Many electronic circuits contain provisions against thermal runaway. This is most often seen in transistor biasing arrangements. However when equipment is used above its designed ambient temperature, thermal runaway can in some cases still occur. This occasionally causes equipment failures in tropical countries, and when air cooling vents are blocked.

Tantalum capacitors

Tantalum capacitors are under some conditions prone to self-destruction by thermal runaway. The capacitor typically consists of a sintered tantalum sponge acting as the anode, a manganese dioxide cathode, and a dielectric layer of tantalum pentoxide created on the tantalum sponge surface by anodizing. The tantalum oxide layer may have weak spots that undergo dielectric breakdown during a voltage spike. The tantalum then comes to direct contact with manganese dioxide and the leakage current causes a local heating; a chemical reaction then produces manganese(III) oxide and regenerates (self-heals) the tantalum oxide layer.

If the energy dissipated at the failure point is high enough, a self-sustaining exothermic reaction may occur, similar to the thermite reaction, with tantalum as fuel and manganese dioxide as oxidizer, destroying the capacitor and occasionally producing smoke and possibly flame.

Batteries

When handled improperly, some rechargeable batteries can experience thermal runaway, resulting in overheating. Sealed cells will sometimes explode. Especially prone to thermal runaway are lithium-ion batteries. Reports of exploding cellphones occasionally appear in newspapers. In 2006, laptop batteries from Apple, HP, Toshiba, Lenovo, Dell and other notebook manufacturers were recalled because of fire and explosions.   The Pipeline and Hazardous Materials Safety Administration (PHMSA) of the U.S. Department of Transportation recently established regulations regarding the carrying of certain types of batteries on airplanes because of their instability in certain situations. This action was partially inspired by a fire on a UPS airplane.

Current crowding

Current crowding (also **current crowding effect**, or **CCE**) is a nonhomogenous distribution of current density through a conductor or semiconductor, especially at the vicinity of the contacts and over the PN junctions.

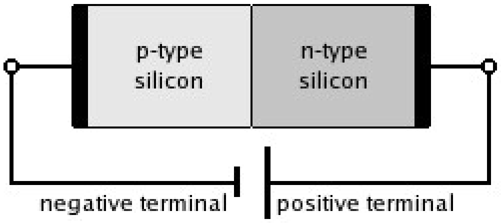
Current crowding is one of the limiting factors of efficiency of light emitting diodes. Materials with low mobility of charge carriers, eg. AlGaInP, are especially prone to current crowding phenomena. It is a dominant loss mechanisms in some LEDs, where the current densities especially around the P-side contacts reach the part of the emission characteristics with lower brightness/current efficiency.

Current crowding can lead to localized overheating and formation of thermal hot spots, in catastrophic cases leading to thermal runaway. Nonhomogenous distribution of current also aggravates electromigration effects and formation of voids. Formation of voids causes localized nonhomogenity of current density, and the increased resistance around the void causes further localized temperature rise, which in turn accelerates the formation of the void. Conversely, localized lowering of current density may lead to deposition of the migrated atoms, leading to further lowering of current density and further deposition of material and formation of hillocks, which may cause short circuits.

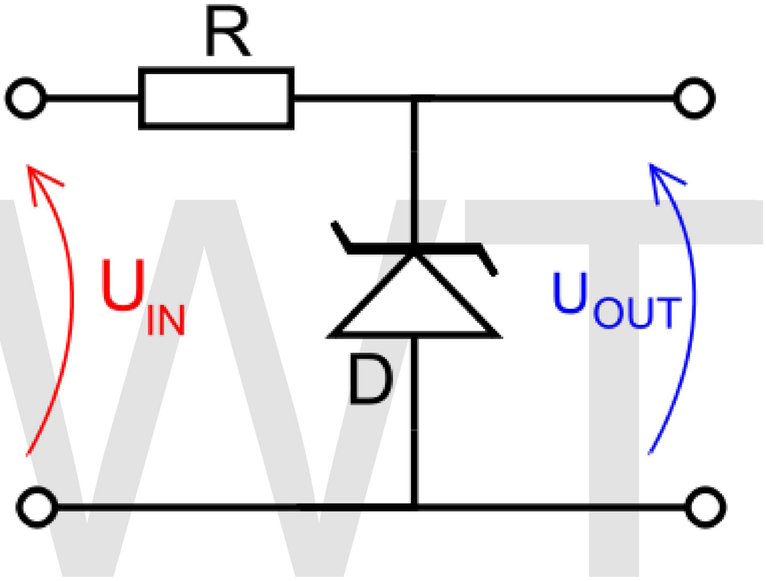
In large bipolar transistors, the resistance of the base layer influences the distribution of current density through the base region, especially at the emitter side.

Current crowding occurs especially at the areas of localized lowered resistance, or in areas where the field strength is concentrated (eg. at the edges of layers).

Reverse bias



A silicon p–n junction in reverse bias



Reverse biased usually refers to how a diode is used in a circuit. If a diode is reverse biased, the voltage at the cathode is higher than that at the anode. Therefore, no current will flow until the diode breaks down. Connecting the *P-type* region to the *negative* terminal of the battery and the *N-type* region to the *positive* terminal, corresponds to reverse bias. The connections are illustrated in the following diagram:

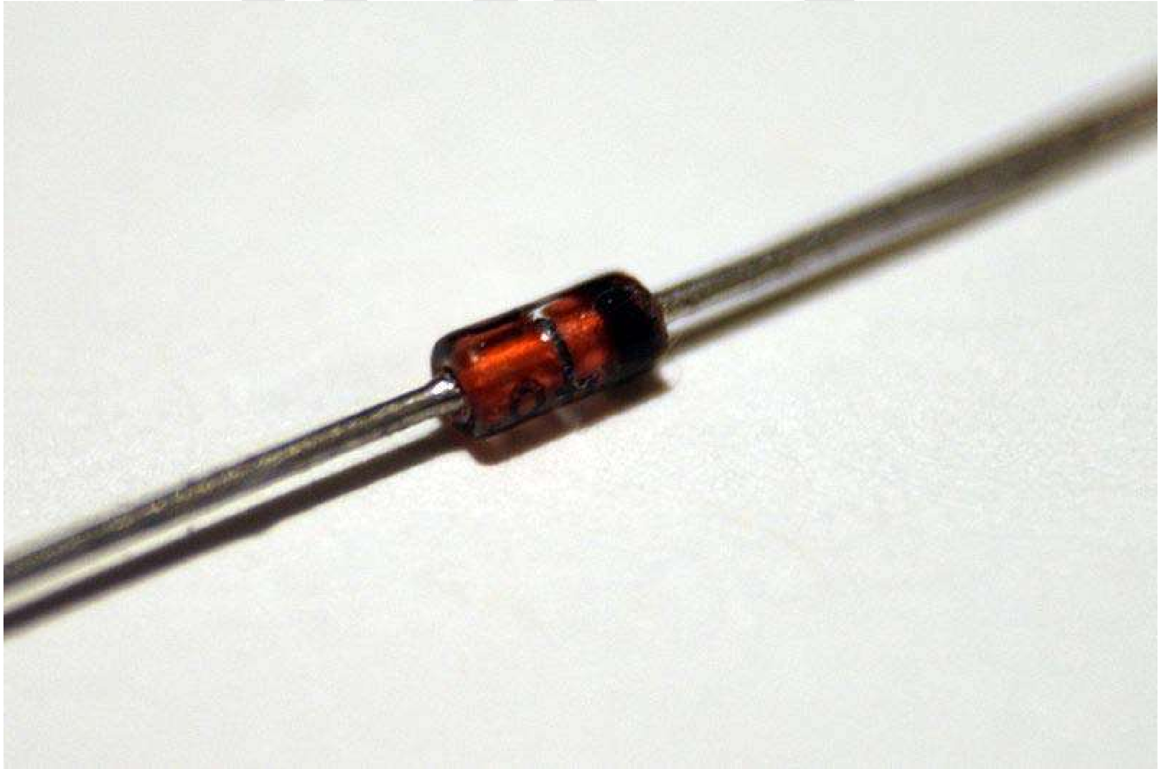
Because the p-type material is now connected to the negative terminal of the power supply, the 'holes' in the P-type material are pulled away from the junction, causing the width of the depletion zone to increase. Similarly, because the N-type region is connected to the positive terminal, the electrons will also be pulled away from the junction. Therefore the depletion region widens, and does so increasingly with increasing reverse-bias voltage. This increases the voltage barrier causing a high resistance to the flow of charge carriers thus allowing minimal electric current to cross the p–n junction. The increase in resistance of the p-n junction results in the junction to behave as an insulator. This is important for radiation detection because if current was able to flow, the charged particles would just dissipate into the material. The reverse bias ensures that charged particles are able to make it to the detector system.

The strength of the depletion zone electric field increases as the reverse-bias voltage increases. Once the electric field intensity increases beyond a critical level, the p–n junction depletion zone breaks-down and current begins to flow, usually by either the Zener or avalanche breakdown processes. Both of these breakdown processes are non-destructive and are reversible, so long as the amount of current flowing does not reach levels that cause the semiconductor material to overheat and cause thermal damage.

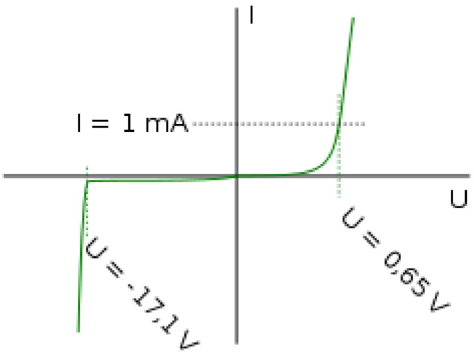
This effect is used to one's advantage in zener diode regulator circuits. Zener diodes have a certain - low - breakdown voltage. A standard value for breakdown voltage is for instance 5.6V. This means that the voltage at the cathode can never be more than 5.6V higher than the voltage at the anode, because the diode will break down - and therefore conduct - if the voltage gets any higher. This effectively regulates the voltage over the diode.

Another application where reverse biased diodes are used is in Varicap diodes. The width of the depletion zone of any diode changes with voltage applied. This varies the capacitance of the diode. For more information, refer to the Varicap article.

Zener diode



Zener diode



Current-voltage characteristic of a Zener diode with a breakdown voltage of 17 volt. Notice the change of voltage scale between the forward biased (positive) direction and the reverse biased (negative) direction.

A **Zener diode** is a type of diode that permits current not only in the forward direction like a normal diode, but also in the reverse direction if the voltage is larger than the breakdown voltage known as "Zener knee voltage" or "Zener voltage". The device was named after Clarence Zener, who discovered this electrical property.

A conventional solid-state diode will not allow significant current if it is reverse-biased below its reverse breakdown voltage. When the reverse bias breakdown voltage is exceeded, a conventional diode is subject to high current due to avalanche breakdown. Unless this current is limited by circuitry, the diode will be permanently damaged. In case of large forward bias (current in the direction of the arrow), the diode exhibits a voltage drop due to its junction built-in voltage and internal resistance. The amount of the voltage drop depends on the semiconductor material and the doping concentrations.

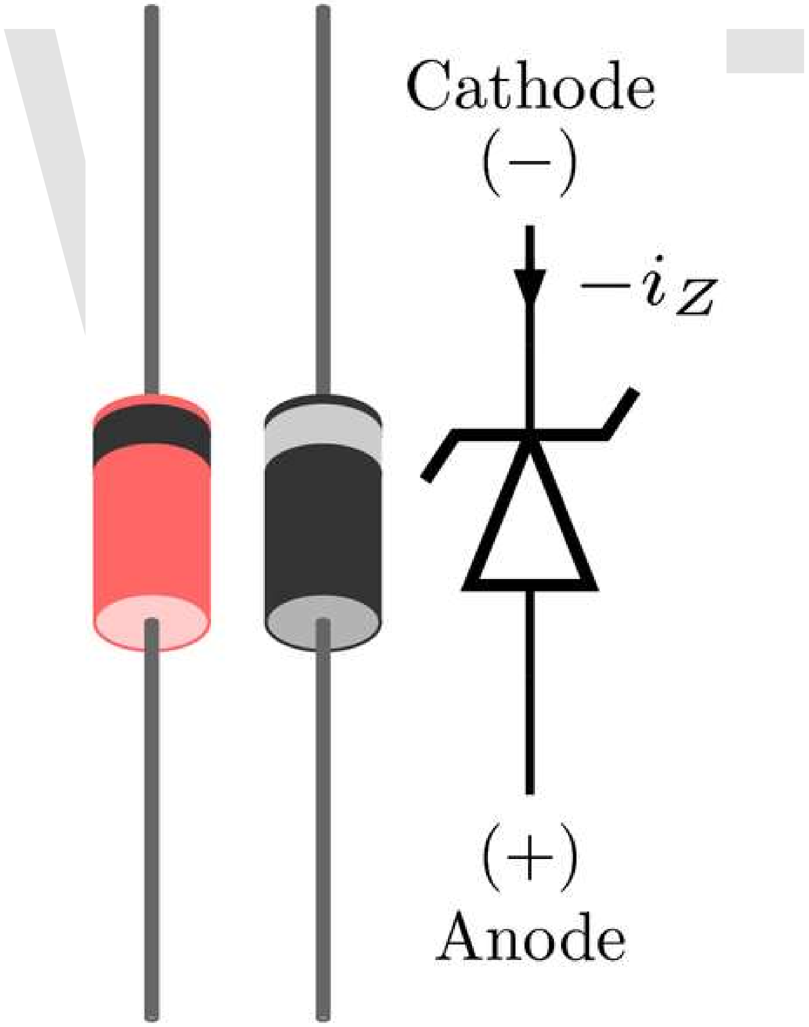
A Zener diode exhibits almost the same properties, except the device is specially designed so as to have a greatly reduced breakdown voltage, the so-called Zener voltage. By contrast with the conventional device, a reverse-biased Zener diode will exhibit a controlled breakdown and allow the current to keep the voltage across the Zener diode at the Zener voltage. For example, a diode with a Zener breakdown voltage of 3.2 V will exhibit a voltage drop of 3.2 V even if reverse bias voltage applied across it is more than its Zener voltage. The Zener diode is therefore ideal for applications such as the generation of a reference voltage (e.g. for an amplifier stage), or as a voltage stabilizer for low-current applications.

The Zener diode's operation depends on the heavy doping of its p-n junction allowing electrons to tunnel from the valence band of the p-type material to the conduction band of the n-type material. In the atomic scale, this tunneling corresponds to the transport of valence band electrons into the empty conduction band states; as a result of the reduced barrier between these bands and high electric fields that are induced due to the relatively high levels of dopings on both sides. The breakdown voltage can be controlled quite accurately in the doping process. While tolerances within 0.05% are available, the most widely used tolerances are 5% and 10%. Breakdown voltage for commonly available zener diodes can vary widely from 1.2 volts to 200 volts.

Another mechanism that produces a similar effect is the avalanche effect as in the avalanche diode. The two types of diode are in fact constructed the same way and both effects are present in diodes of this type. In silicon diodes up to about 5.6 volts, the Zener effect is the predominant effect and shows a marked negative temperature coefficient. Above 5.6 volts, the avalanche effect becomes predominant and exhibits a positive temperature coefficient. In a 5.6 V diode, the two effects occur together and their temperature coefficients neatly cancel each other out, thus the 5.6 V diode is the component of choice in temperature-critical applications. Modern manufacturing techniques have produced devices with voltages lower than 5.6 V with negligible temperature coefficients, but as higher voltage devices are encountered, the temperature coefficient rises dramatically. A 75 V diode has 10 times the coefficient of a 12 V diode.

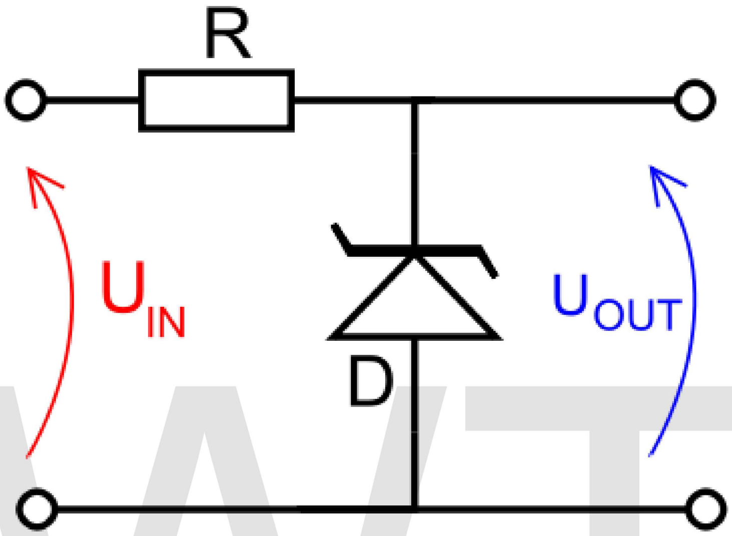
All such diodes, regardless of breakdown voltage, are usually marketed under the umbrella term of "Zener diode".

Uses



Zener diode shown with typical packages. *Reverse* current $-i_Z$ is shown

Zener diodes are widely used as voltage references and as shunt regulators to regulate the voltage across small circuits. When connected in parallel with a variable voltage source so that it is reverse biased, a Zener diode conducts when the voltage reaches the diode's reverse breakdown voltage. From that point on, the relatively low impedance of the diode keeps the voltage across the diode at that value.



In this circuit, a typical voltage reference or regulator, an input voltage, U_{IN}, is regulated down to a stable output voltage U_{OUT}. The intrinsic voltage drop of diode D is stable over a wide current range and holds U_{OUT} relatively constant even though the input voltage may fluctuate over a fairly wide range. Because of the low impedance of the diode when operated like this, Resistor R is used to limit current through the circuit.

In the case of this simple reference, the current flowing in the diode is determined using Ohms law and the known voltage drop across the resistor R. $I_{Diode} = (U_{IN} - U_{OUT}) / R_{\Omega}$

The value of *R* must satisfy two conditions:

1. *R* must be small enough that the current through D keeps D in reverse breakdown. The value of this current is given in the data sheet for D. For example, the common BZX79C5V6 device, a 5.6 V 0.5 W Zener diode, has a recommended reverse current of 5 mA. If insufficient current exists through D, then U_{OUT} will be unregulated, and less than the nominal breakdown voltage (this differs to voltage regulator tubes where the output voltage will be higher than nominal and could rise as high as U_{IN}). When calculating *R*, allowance must be made for any current through the external load, not shown in this diagram, connected across U_{OUT}.

2. *R* must be large enough that the current through D does not destroy the device. If the current through D is I_D, its breakdown voltage V_B and its maximum power dissipation P_{MAX}, then $I_D V_B < P_{MAX}$.

A load may be placed across the diode in this reference circuit, and as long as the zener stays in reverse breakdown, the diode will provide a stable voltage source to the load.

Shunt regulators are simple, but the requirements that the ballast resistor be small enough to avoid excessive voltage drop during worst-case operation (low input voltage concurrent with high load current) tends to leave a lot of current flowing in the diode much of the time, making for a fairly wasteful regulator with high quiescent power dissipation, only suitable for smaller loads.

Zener diodes in this configuration are often used as stable references for more advanced voltage regulator circuits.

These devices are also encountered, typically in series with a base-emitter junction, in transistor stages where selective choice of a device centered around the avalanche/Zener point can be used to introduce compensating temperature co-efficient balancing of the transistor PN junction. An example of this kind of use would be a DC error amplifier used in a regulated power supply circuit feedback loop system.

Zener diodes are also used in surge protectors to limit transient voltage spikes.

Another notable application of the zener diode is the use of noise caused by its avalanche breakdown in a random number generator that never repeats.

Latchup

Latchup is a term used in the realm of integrated circuits (ICs) to describe a particular type of short circuit which can occur in an improperly designed circuit. More specifically it is the inadvertent creation of a low-impedance path between the power supply rails of a MOSFET circuit, triggering a parasitic structure which disrupts proper functioning of the part and possibly even leading to its destruction due to overcurrent. A power cycle is required to correct this situation.

The parasitic structure is usually equivalent to a thyristor (or SCR), a PNPN structure which acts as a PNP and an NPN transistor stacked next to each other. During a latchup when one of the transistors is conducting, the other one begins conducting too. They both keep each other in saturation for as long as the structure is forward-biased and some current flows through it - which usually means until a power-down. The SCR parasitic structure is formed as a part of the totem-pole PMOS and NMOS transistor pair on the output drivers of the gates.

The latchup does not have to happen between the power rails; it can happen at any place where the required parasitic structure exists. A spike of positive or negative voltage on an input or output pin of a digital chip, exceeding the rail voltage by more than a diode drop, is a common cause of latchup. Another cause is the supply voltage exceeding the absolute maximum rating, often from a transient spike in the power supply, leading to a breakdown of some internal junction. This frequently happens in circuits which use multiple supply voltages that do not come up in the proper order after a power-up, leading to voltages on data lines exceeding the input rating of parts that have not yet reached a nominal supply voltage.

Yet another common cause of latchups is ionizing radiation which makes this a significant issue in electronic products designed for space (or very high-altitude) applications.

Latchup prevention

It is possible to design chips that are latchup-resistant, where a layer of insulating oxide (called a *trench*) surrounds both the NMOS and the PMOS transistors. This breaks the parasitic SCR structure between these transistors. Such parts are important in the cases where the proper sequencing of power and signals cannot be guaranteed (e.g., in hot swap devices).

Devices fabricated in lightly doped epitaxial layers grown on heavily doped substrates are also less susceptible to latchup. The heavily doped layer acts as a current sink where excess minority carriers can quickly recombine.

Another possibility for a latchup prevention is the *Latchup Protection Technology* circuit. When a latchup is detected, the LPT circuit shuts down the chip and holds it powered-down for a preset time.

Most silicon-on-insulator devices are inherently latchup-resistant. Latchup is the low resistance connection between tub and power supply rails.

Chapter 13

Optoelectronics Failures

Catastrophic optical damage

Catastrophic optical damage (**COD**), or **catastrophic optical mirror damage** (**COMD**), is a failure mode of high-power semiconductor lasers. It occurs when the semiconductor junction is overloaded by exceeding its power density and absorbs too much of the produced light energy, leading to melting and recrystallization of the semiconductor material at the facets of the laser. This is often colloquially referred to as "blowing the diode." The affected area contains a large number of lattice defects, negatively affecting its performance. If the affected area is sufficiently large, it can be observable under optical microscope as darkening of the laser facet, and/or as presence of cracks and grooves. The damage can occur within a single laser pulse, in less than a millisecond. The time to COD is inversely proportional to the power density.

Catastrophic optical damage is one of the limiting factors in increasing performance of semiconductor lasers. It is the primary failure mode for AlGaInP/AlGaAs red lasers.

Short-wavelength lasers are more susceptible to COD than long-wavelength ones.

The typical values for COD in industrial products range between 12 and 20 MW/cm^2.

Causes and mechanisms

At the edge of a diode laser, where light is emitted, a mirror is traditionally formed by cleaving the semiconductor wafer to form a specularly reflecting plane. This approach is facilitated by the weakness of the [110] crystallographic plane in III-V semiconductor crystals (such as GaAs, InP, GaSb, etc.) compared to other planes. A scratch made at the edge of the wafer and a slight bending force causes a nearly atomically perfect mirror-like cleavage plane to form and propagate in a straight line across the wafer.

But it so happens that the atomic states at the cleavage plane are altered (compared to their bulk properties within the crystal) by the termination of the perfectly periodic lattice at that plane. Surface states at the cleaved plane have energy levels within the (otherwise forbidden) band gap of the semiconductor.

The absorbed light causes generation of electron-hole pairs. These can lead to breaking of chemical bonds on the crystal surface followed by oxidation, or to release of heat by nonradiative recombination. The oxidized surface then shows increased absorption of the laser light, which further accelerates its degradation. The oxidation is especially problematic for semiconductor layers containing aluminium.

Essentially, as a result when light propagates through the cleavage plane and transits to free space from within the semiconductor crystal, a fraction of the light energy is absorbed by the surface states whence it is converted to heat by phonon-electron interactions. This heats the cleaved mirror. In addition the mirror may heat simply because the edge of the diode laser—which is electrically pumped—is in less-than-perfect contact with the mount that provides a path for heat removal. The heating of the mirror causes the band gap of the semiconductor to shrink in the warmer areas. The band gap shrinkage brings more electronic band-to-band transitions into alignment with the photon energy causing yet more absorption. This is thermal runaway, a form of positive feedback, and the result can be melting of the facet, known as *catastrophic optical damage*, or COD.

Deterioration of the laser facets with aging and effects of the environment (erosion by water, oxygen, etc.) increases light absorption by the surface, and decreases the COD threshold. A sudden catastrophic failure of the laser due to COD then can occur after many thousands hours in service.

Improvements

One of the methods of increasing the COD threshold in AlGaInP laser structures is the sulfur treatment, which replaces the oxides at the laser facet with chalcogenide glasses. This decreases the recombination velocity of the surface states.

Reduction of recombination velocity of surface states can be also achieved by cleaving the crystals in ultrahigh vacuum and immediate deposition of a suitable passivation layer.

A thin layer of aluminium can be deposited over the surface, for gettering the oxygen.

Another approach is doping of the surface, increasing the band gap and decreasing absorption of the lasing wavelength, shifting the absorption maximum several nanometers up.

Current crowding near the mirror area can be avoided by prevention of injecting charge carriers near the mirror region. This is achieved by depositing the electrodes away from the mirror, at least several carrier diffusion distances.

Energy density on the surface can be reduced by employing a waveguide broadening the optical cavity, so the same amount of energy exits through a larger area. Energy density of 15–20 MW/cm^2 corresponding to 100 mW per micrometer of stripe width are now achievable. A wider laser stripe can be used for higher output power, for the cost of transverse mode oscillations and therefore worsening of spectral and spatial beam quality.

In the 1970s, this problem, which is particularly nettlesome for GaAs-based lasers emitting between 1 µm and 0.630 µm wavelengths (less so for InP based lasers used for long-haul telecommunications which emit between 1.3 µm and 2 µm), was identified. Michael Ettenberg, a researcher and later Vice President at RCA Laboratories' David Sarnoff Research Center in Princeton, New Jersey, devised a solution. A thin layer of aluminum oxide was deposited on the facet. If the aluminum oxide thickness is chosen correctly, it functions as an anti-reflective coating, reducing reflection at the surface. This alleviated the heating and COD at the facet.

Since then, various other refinements have been employed. One approach is to create a so-called non-absorbing mirror (NAM) such that the final 10 µm or so before the light emits from the cleaved facet are rendered non-absorbing at the wavelength of interest. Such lasers are called **window lasers**.

In the very early 1990s, SDL, Inc. began supplying high power diode lasers with good reliability characteristics. CEO Donald Scifres and CTO David Welch presented new reliability performance data at, e.g., SPIE Photonics West conferences of the era. The methods used by SDL to defeat COD were considered to be highly proprietary and have still not been disclosed publicly as of June, 2006.

In the mid-1990s IBM Research (Ruschlikon, Switzerland) announced that it had devised its so-called "E2 process" which conferred extraordinary resistance to COD in GaAs-based lasers. This process, too, has never been disclosed as of June, 2006.

Hydrogen darkening

Hydrogen darkening is a physical degradation of the optical properties of glass. Free hydrogen atoms are able to bind to the SiO_2 silica glass compound forming hydroxyl (OH) - a chemical compound that interferes with the passage of light through the glass.

The problem is particularly relevant to fiber optic cables — particularly in oil and gas wells where fiber optic cables are used for distributed temperature sensing. Hydrogen can be present due to the cracking of hydrocarbons in the well. The darkening of the fiber can distort the DTS reading and possibly render the DTS system inoperable due to the optical loss budget being exceeded.

To prevent this, coatings such as carbon are applied to the fiber, and hydrogen capturing gels are used to buffer the fiber, and other proprietary techniques may be used to prevent hydrogen atoms from reaching the glass fiber via the cable sheath.