Notes for Microelectronics Fabrication I

Basic Semiconductor Material Science and Solid-State Physics

All terrestrial materials are made up of atoms. Indeed, the ancient Greeks put this hypothesis forward over two millennia ago. However, it was not until the twentieth century that the atomic theory of matter became firmly established as an unassailable, demonstrated fact. Moreover, it is now known that properties of all common forms of matter (excluding such exotic forms as may exist under conditions only found in white dwarfs, neutron stars, or black holes) are, in principle, completely determined by the properties of individual constituent atoms and their mutual interactions. Indeed, there are just over one hundred different types of atoms, *viz.*, the chemical elements, as summarized on a standard periodic chart. Most of these are quite rare and, worse yet, many are fundamentally unstable; only about two dozen are common and these make up the bulk of the natural world. Fortunately, for the modern electronics industry, silicon is one of the most common elements found on planet Earth.

Naturally, atoms were originally thought of as exceedingly small indivisible bits of solid matter. Moreover, it would seem trivially obvious that the simplest form such a particle could assume is that of a miniscule "billiard ball". Even so, in addition to simple spherical form, early philosophers and scientists variously conceptualized atoms as having different sizes and geometrical shapes, *e.g.*, cubes, polyhedra, *etc.* Accordingly, these differences between the atoms themselves were thought to account for the wide variation of physical properties apparent in all material substances. Furthermore, to account for chemical reactions and the formation of definite compounds, during the early development of modern chemistry it was even proposed that each type of atom might have a characteristic number and arrangement of "hooks" on its surface. Of course, all of these early speculations have been superseded by modern atomic theory based on quantum mechanics in which an atom appears as a spherical structure having a central positively-charged, massive nucleus (composed of protons and neutrons) surrounded by a "cloud" of orbiting negatively-charged, light electrons. Nevertheless, the primitive idea that physical differences and geometrical arrangements of constituent atoms are fundamentally significant to determine bulk properties of material substances has proven substantially correct. Accordingly, each type of atom has a unique electronic configuration, which is determined by nuclear charge or atomic number, Z, and the quantum mechanical behavior of electrons bound in a Coulomb potential. For a particular atomic species, four quantum numbers are required to specify the quantum state of any electron. These are, n, the principal quantum number (corresponding broadly to electronic energy), l, the azimuthal quantum number (corresponding to the magnitude of electron orbital angular momentum), m, the magnetic quantum number (corresponding to a specific, but arbitrarily chosen component of electron orbital angular momentum), and, s, the spin quantum number (corresponding to one of two possible spin states of an electron). The principal quantum number assumes strictly positive integer values, the azimuthal quantum number assumes non-negative integer values increasing from 0 to $n-1$, the magnetic quantum number takes integer values running consecutively from $-l$ to $+l$, and the spin quantum number takes only discrete half-integer values, $+½$ and $-½$. Therefore, each principal quantum shell (or energy level) is characterized by n azimuthal sub-shells and each azimuthal sub-shell is characterized by $2l+1$ magnetic sub-levels. In this way the three quantum numbers, *n*, *l*, and *m*, serve to define specific *atomic orbitals*. (The role of the *s* quantum number will be considered subsequently.)

Atomic Orbitals

Although orbitals are defined mathematically over all space, one can visualize a particular orbital (if occupied) as a finite region in space for which the probability of observing an electron associated with a particular set of quantum numbers significantly differs from zero. As such, it follows from quantum mechanical principles that an orbital does not have absolute significance, but depends on details of particular measurements or observations. Moreover, as a practical matter, the most convenient physical variables to observe are conserved quantities, *i.e.*, constants of motion, such as total energy, angular momentum, *etc*. For this reason, the usual atomic quantum numbers, *n*, *l*, and *m*, are often treated as essential; however, this is really just useful convention and, in general, orbitals can be defined in terms of any dynamically complete set of variables. Within this context, allowed values of the principal quantum number, *n*, can be thought of as defining a set of concentric spherical electron shells centered on the nucleus. With respect to increasing energy, *i.e.*, increasing *n*, each principal shell is characterized by the appearance of a new kind of orbital corresponding to the highest value of the azimuthal quantum number (which increases by unit value for each "new" principal shell) and the number of possible magnetic quantum numbers determines the number of the orbitals of each kind. Thus, for $l = 0$, there is only one kind of orbital of spherical shape called an *s*-orbital. For $l = 1$, there are three orbitals, called *p*-orbitals, which are shaped like dumbbells. Hence, each *p*-orbital is axially symmetric and oriented along a cartesian axis, *viz.*, *x*, *y*, or *z* axis. (Of course, coordinate axes can be chosen simply for convenience, hence illustrating the arbitrary nature of atomic orbitals as asserted previously.) Similarly, for $l = 2$, there are five orbitals of rosette shape which are called *d*-orbitals. These are also oriented with respect to specified axes; however, exact details are more complicated. For higher values of the azimuthal quantum number new kinds of orbitals exist, *e.g.*, *f*-orbitals in the case of $l = 3$, but, they are generally not as important to chemical interactions and the bonding of crystals as are *s*, *p*, and *d*-orbitals. By convention, atomic orbitals are generally designated by type (*s, p, d, f, etc.*) and principal quantum number. Therefore, in order of increasing energy, the standard atomic orbitals are 1*s*, 2*s*, 2*p*, 3*s*, 3*p*, 4*s*, 3*d*, *etc*.

The Pauli Exclusion Principle and Hund's Rule determine the occupancy of any particular orbital. Accordingly, the Pauli Exclusion Principle stipulates that no two electrons can be associated with exactly the same set of quantum numbers. Therefore, since only two values for the *s* quantum number are possible, maximum occupancy of any single orbital is two, *i.e.*, it can be occupied by one electron "spin up", *viz.*, spin quantum number equal to +½, and one electron "spin down", *viz.*, spin quantum number equal to –½. Clearly, this is of great importance, since if all electrons could simultaneously occupy the orbital of lowest energy, atoms would collapse and ordinary matter could not exist. In addition, Hund's Rule stipulates that as a multi-electron structure is built up, all available orbitals of a given energy are first occupied singly, *i.e.*, by electrons having the same spin quantum number, before they are "paired up". For atomic structures, this gives rise to Pauli's well-known *aufbau* or "building", principle.

For completeness, one should observe that in modified form these same rules generally apply to more complicated quantum mechanical systems, *e.g.*, molecules and crystals, as well as to atoms.

Chemical Bonding

Compound materials are formed by *chemical bonding*. This can be visualized as the "overlap" of two singly occupied, *i.e.*, half-filled, atomic orbitals to form *molecular orbitals* and is illustrated schematically by the following figure.

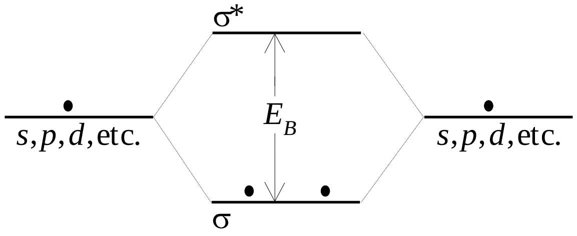


Fig. 1: Molecular orbital diagram illustrating formation of a chemical bond

Here, the horizontal dimension represents atomic separation and the vertical dimension represents electronic energy. Clearly, if a pair of atoms is widely separated, then the system just consists of two singly occupied atomic orbitals (*s*, *p*, *d*, *etc.*) of equal energy. (Valence electrons are indicated schematically by "big black dots".) In contrast, if the atoms are brought into close proximity as indicated by the slanted lines, then atomic orbitals "mix" to form two molecular orbitals of different energy. These are denoted conventionally as σ and σ*-orbitals. Naturally, the two electrons originally in separated atomic orbitals will both end up occupying the σ-orbital, *i.e.*, the lowest energy orbital, since this implies an overall reduction of electronic energy by a specific amount, *viz.*, E_B. Such a situation indicates formation of a *chemical bond* between the two original atoms. Therefore, E_B can be immediately identified with *bond energy* and, hence, the σ-orbital is called a *bonding orbital*. Moreover, it is useful to consider what happens if the original atomic orbitals had been doubly occupied, *i.e.*, filled. This situation is illustrated by the following figure:

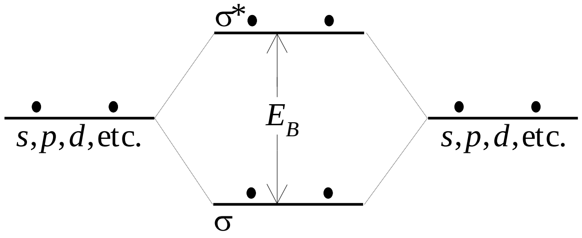


Fig. 2: Molecular orbital diagram illustrating non-bonding of filled atomic orbitals

In this case, there are four electrons to be considered and, therefore, due to the Pauli Exclusion Principle, both σ and σ*-orbitals must be fully occupied. Clearly, this results in no net lowering of overall electronic energy; hence, no stable chemical bond is formed. Therefore, the σ*-orbital is identified as an *anti-bonding orbital*. In a broad sense, electrons occupying an anti-bonding orbital counteract the effect of electrons occupying a corresponding bonding orbital.

A simple physical example of covalent chemical bonding is provided by formation of a hydrogen molecule from two separated hydrogen atoms. Clearly, overlap of 1*s* orbitals, each of which is half-filled, results in a filled bonding orbital and an empty anti-bonding orbital. Accordingly, the two electrons are localized between the two hydrogen nuclei as one expects from a classical picture of chemical bonding as sharing of electron pairs. Furthermore, this scheme also illustrates why helium does not form a diatomic species since obviously, any molecular helium structure has bonding and anti-bonding orbitals both completely filled. Moreover, with appropriate modification, a similar scheme accounts for the occurrence of a number of elemental species as diatomic molecules, *e.g.*, N_2, O_2, F_2, Cl_2, *etc.*, rather than as isolated gas atoms. In addition, formation of a chemical bond also illustrates a fundamental principle of quantum mechanics that any linear combination of some “old” set of orbitals to construct a “new” set of orbitals must conserve the total number of orbitals; however, in contrast, orbital energies generally do not remain the same in both sets. Indeed, by definition bonding orbitals have lower energy than corresponding anti-bonding orbitals, which, of course, merely accounts for the binding energy associated with the resulting chemical bond. Although some additional complexity is unavoidably introduced, essentially this same approach can be applied to the formation of bonds between any pair or even a larger group of atoms.

Semiconductors

Digressing briefly from general consideration of electronic structure, one observes that in the periodic chart, metals appear to the left side and non-metals to the right side; in between are elements having properties intermediate to those of metals and non-metals. Consequently, this is precisely the location of the elemental semiconductors, in particular, silicon and germanium (Si and Ge). Moreover, although germanium was the first semiconductor material to be successfully commercialized, its volume of use was soon exceeded by silicon, which is now dominant in the electronics industry and can be expected to remain so for the foreseeable future. Both silicon and germanium are Group IVB elements and both have the same cubic crystal structure with lattice parameters of 0.543 and 0.566 nm, respectively. Furthermore, in addition to elemental semiconductors, compound semiconductors also exist. The most commercially significant of these is *gallium arsenide* (GaAs), although more recently *indium phosphide* (InP) and *gallium nitride* (GaN) have significantly increased in importance. Gallium arsenide, indium phosphide, gallium nitride, and other materials such as indium antimonide (InSb), aluminum arsenide (AlAs), *etc.*, provide examples of III-V compound semiconductors. The origin of this designation is quite clear; one element comes from Group IIIB of the periodic chart and the other from Group VB. Furthermore, bonding in III-V compounds is very similar to that of elemental semiconductors since the one electron "deficiency" of the Group IIIB element is exactly compensated by an "extra" electron associated with the Group VB element. As a consequence, III-V semiconductors have substantially the same electronic structure as that of corresponding elemental semiconductors. Within this context, it would seem plausible to extend such a scheme further. Indeed, this is possible and additional materials called II-VI compound semiconductors are also found to exist. Obviously, this designation indicates the combination of Group IIB elements with Group VIB elements, in which case the Group IIB element is considered deficient by two electrons with this deficiency compensated by two extra electrons from the Group VIB element. Examples of II-VI semiconductors are cadmium selenide (CdSe), mercury telluride (HgTe), *etc.* Further consideration of compound semiconductors will not be entertained within the present context; however, it should be obvious that semiconductor materials are generally composed of elements from Group IVB or groups which are symmetric about Group IVB in the periodic chart. In addition, although materials such as carbon (C) in the form of diamond, cubic boron nitride (BN), silicon carbide (SiC), *etc.*, may behave as insulators at room temperature, these also become semiconductors at higher temperatures.

It is obvious from the structure of silicon that its atomic coordination number is four. This follows directly from the electron configuration, which for silicon is characterized by four valence electrons in the outer atomic shell. Moreover, one would expect on the basis of the primitive atomic electron configuration, that silicon should be characterized by a filled 3*s* orbital, two half-filled 3*p* orbitals and one empty 3*p* orbital. However, as asserted previously, orbitals should not be considered as absolute, but merely as provisional descriptions of electronic motion. In a mathematically precise sense, orbitals ultimately represent particular solutions of a linear partial differential equation in space and time, *e.g.*, a one electron Schrödinger wave equation. As is well-known from the mathematical theory of linear differential equations, the sum (or difference) of any two particular solutions of the equation is itself a "new" particular solution. This is called the Principle of Superposition. Therefore, if one constructs four independent linear combinations of a single 3*s* and three 3*p* orbitals, then one obtains a mathematically equivalent group of four new orbitals called sp^3 hybrids. These hybrid orbitals are no longer characterized by exact values of electronic angular momentum or energy; however, they do exhibit tetrahedral coordination and, as such, are particularly useful for description of bonding in a silicon crystal. Furthermore, one observes that each one of the sp^3 orbitals is exactly half-filled. Thus, the overlap of singly occupied sp^3 orbitals from two adjacent silicon atoms results in the formation of a doubly occupied bonding orbital in complete analogy to the elementary case of the hydrogen molecule. Hence, the covalently bonded crystal structure of silicon emerges naturally. The case of germanium is identical, except that 4*s* and 4*p* orbitals are to be considered instead of 3*s* and 3*p* orbitals. Similarly, gallium arsenide has a completely analogous structure. In this case though, the situation is slightly more complicated. For gallium atoms, one of the sp^3 orbitals can be regarded as empty with the remaining three half-filled. Conversely, for arsenic atoms, one of the sp^3 orbitals can be regarded as completely filled, again, with the remaining three half-filled. Of course, this description is purely formal since by definition, all of the sp^3 orbitals are mathematically equivalent. Clearly, when these orbitals are overlapped to form a bulk crystal, the total number of electrons and orbitals is just the same as in the case of the corresponding elemental semiconductors. In a formal sense, one may regard this as a consequence of the specific overlap of empty gallium sp^3 orbitals with filled arsenic sp^3 orbitals. Of course, the remaining half-filled orbitals from gallium and arsenic atoms also overlap just as in the case of silicon or germanium.

The Electronic Structure of Crystals

So far, consideration has been limited to the electronic structure of atoms and formation of covalent bonds between pairs of atoms. A solid crystal is, of course, a much larger and more extended assemblage of atoms. Nevertheless, quantum mechanical principles governing the formation of molecules are not essentially different when extended to whole crystals and is illustrated for silicon in the following figure:

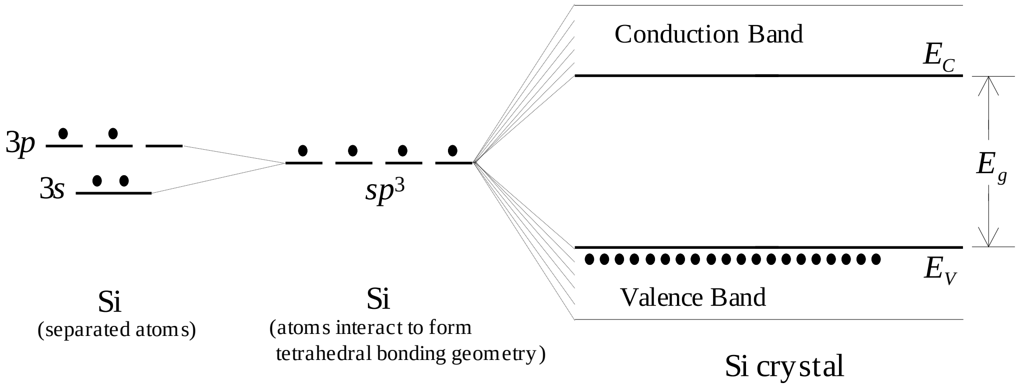


Fig. 3: Molecular orbital diagram illustrating formation of energy bands in crystalline silicon

This can be called the "molecular orbital approach" to the electronic structure of crystals. Although the mathematics is quite complicated and will not be considered further, one observes that orbitals for a whole crystal can be obtained, in principle, by combining all of the atomic orbitals, *e.g.*, *sp*^{3} hybrids, of constituent atoms in just the same way as atomic orbitals from two atoms are combined to form bonding and anti-bonding molecular orbitals, *i.e.*, a chemical bond. However, in the case of a crystal, linear combination of atomic orbitals results in *band orbitals* having energies falling in an essentially continuous range or *energy band*. Moreover, band orbitals are generally *delocalized* over the entire crystal. That is to say that the orbitals of a crystal are no longer necessarily identified with individual atoms or individual covalent bonds but generally exist throughout the entire body of the solid. Even so, the band structure of crystal, at least in a broad sense, still corresponds to the formation of bonding and anti-bonding molecular orbitals. To be more specific, the lower energy or *valence band* corresponds to bonding. Indeed, from a simplistic viewpoint, valence band orbitals can be regarded as linear combinations formed from all bonding orbitals constructed between pairs of constituent atoms of the crystal. Similarly, the higher energy or *conduction band* is the anti-bonding analog and, again, conduction band orbitals can be primitively regarded as linear combinations of all of anti-bonding orbitals associated with atomic pairs. In addition, a distinguishing feature of conduction band orbitals is that they are substantially more delocalized than corresponding valence band orbitals. Therefore, if an electron undergoes a transition from the valence band to the conduction band by, for example, thermal or photo excitation, then it becomes essentially free to wander throughout the body of the crystal, *i.e.*, it becomes a *mobile carrier* of electrical current. However, it must be cautioned that this picture is quite oversimplified. In a real crystal, more than two bands are generally formed when all relevant atomic orbitals are overlapped. In this case, the resulting band structure is generally quite complex and may have mixed bonding and anti-bonding character. Fortunately, for understanding the behavior of solid-state electronic devices as well as many other characteristics of semiconductors, a simple uniform two band picture is usually quite sufficient.

An obvious consequence of the construction of the electronic structure of a whole crystal from bonding and anti-bonding orbitals is the possible appearance of an *energy gap*, E_g. Of course, the size of the gap depends not only on the binding energy of the crystal, but also on “widths” (measured on an energy scale) of the valence and conduction bands. In insulating materials this gap is quite large, typically several electron-volts. Thus, electrons are promoted from the valence band to the conduction band only by expenditure of a large amount of energy, hence, very few if any, mobile carriers are ever present within an insulator at ordinary temperatures. In contrast, for some electrical conductors, *viz.*, *semimetals*, valence and conduction bands may overlap so that there is no energy gap. In this case, electronic transitions between valence and conduction bands require little or no energy. In other kinds of conductors, *viz., classical metals*, the valence band is only partially filled which, again, results in a significant internal concentration of mobile carriers. Of course, in general metals are characterized by large mobile carrier densities and are good electrical conductors. Thus, as one might have guessed, semiconductors have properties that are intermediate between metals and insulators. They have an energy gap, but it is relatively small, typically, on the order of one or two electron-volts or less. Indeed, the gap is small enough so that it is possible for thermal excitation alone to promote a significant number of electrons from the valence band to the conduction band and, thus, pure (*i.e.*, *intrinsic*) semiconductors exhibit a small but significant electrical conductivity.

Before proceeding, it should be noted that solid-state physicists take a completely different approach to the electronic structure of crystals and treat valence electrons as forming a “gas” that fills the entire volume of the solid. From a quantum mechanical point of view, energy states, *i.e.*, band orbitals, of such a system approximate plane waves. Indeed, if the crystal had no internal structure plane waves would provide an exact description of electronic structure. Accordingly, for a semiconductor crystal an energy gap appears as a consequence of explicit introduction of a periodic potential. Of course, this periodic potential derives directly from the periodic structure of the crystal lattice. To be more specific, spatial periodicity within the crystal causes standing wave states of specific energies to be allowed or forbidden depending on whether the waves concentrate electron density coincident or anti-coincident with extrema of the periodic potential (*i.e.*, coincident or anti-coincident with atomic nuclei). Furthermore, as should be expected, this picture is ultimately both complementary and equivalent to the molecular orbital approach in which bonding and anti-bonding orbitals also correspond to specific localizations of electron density.

Bands in Intrinsic Semiconductors

The band structure of any real crystalline semiconductor is quite complicated and allows for different types of behavior. For example, silicon and germanium are said to be *indirect* band gap semiconductors and gallium arsenide a *direct* band gap semiconductor. To be precise, in a direct band gap semiconductor an electron can be promoted from the valence band to the conduction band directly with no change in momentum, *e.g.*, by absorption of a photon. Physically, this occurs because of favorable alignment and curvature (or shape) of energy bands as constructed in a momentum representation. Conversely, in an indirect band gap semiconductor similar promotion of an electron from the valence band to the conduction band requires interaction with the crystal lattice in order to satisfy the principle of momentum conservation, *i.e.*, corresponding band alignment is unfavorable in “momentum space”. In passing, it is worthwhile to mention that distinction between direct and indirect band gap materials is of little importance for conventional electronic devices, but is technologically significant for optoelectronic devices, *e.g.*, light emitting diodes, laser diodes, *etc.*, which require direct band gap semiconductors for operation.

Even so, as asserted previously, for many (perhaps, most) practical situations detailed band structure is unimportant and can be greatly simplified into two aggregate bands, *viz.*, valence and conduction bands. Of course, just as in the case of atomic or molecular orbitals, electrons in a crystalline solid must satisfy the Pauli Exclusion Principle. Therefore, ignoring any effect of photo or thermal excitation, the valence band of a semiconductor can be regarded as completely filled and the conduction band as completely empty. Therefore, the conductivity of a pure semiconductor is expected to be quite low since mobility of electrons in the valence band is small. Naturally, this is merely a consequence of participation of the valence electrons in the bonding of the crystal lattice, which requires substantial localization of electron pairs between atomic nuclei. Nevertheless, as noted previously, conductivity of a pure semiconductor increases dramatically when electrons are promoted from the valence band into the conduction band since electrons in the conduction band are much freer to migrate than electrons in the valence band. Thus, electron density is much less localized for conduction band orbitals in comparison with valence band orbitals. Within this context, promotion of electrons into the conduction band leaves behind *holes* in the valence band that can be viewed as a sort of positively charged electron. Indeed, holes and electrons exhibit an approximate symmetry since valence band holes can move through the crystal almost as freely as conduction band electrons. Hence, within a semiconductor crystal, both electrons and holes can be treated formally as a distinct particles (or more correctly, quasi-particles) and can act as mobile carriers having opposite electrical charge.

For clarity, it is important to distinguish an *electron state* of a band and a band orbital. To be specific, in analogy to a primitive atomic orbital, a band orbital is formally associated with two band states, each corresponding to one of the possible spin quantum numbers, *viz.*, ±½. Therefore, on the basis of the Pauli Exclusion Principle, although maximum occupancy of a band orbital is two electrons, maximum occupancy of a band electron state is necessarily only one. Naturally, at ordinary temperatures in a pure semiconductor, some electrons will be promoted to the conduction band by thermal excitation alone. It has long been established that thermal equilibrium for a many electron system is described by the Fermi-Dirac distribution function:

$$f(E) = \frac{1}{1 + e^{(E - E_F)/kT}}$$

Here, $f(E)$, is the probability that an electron state of energy, E, is occupied. The quantity, E_F, is called *Fermi energy* and defines a characteristic energy for which it is equally likely that an electron state of precisely that energy will be either vacant or occupied, *i.e.*, the state has an occupation probability of exactly one half. Clearly, any band state must be either occupied or vacant and, moreover, the approximate symmetric behavior of electrons and holes implies that a vacant electronic state can just as well be regarded as an occupied *hole state* and vice versa. Accordingly, it follows that the occupation probability for holes, $f_h(-E)$, is trivially related to $f(E)$ by the simple formula:

$$f_h(-E) = 1 - f(E)$$

By convention, hole energy is written formally as the negative of electron energy since a hole deep in the valence band physically corresponds to a higher energy state than a hole at the top of the band. Hence, it is easily demonstrated that holes also obey a Fermi-Dirac distribution with corresponding Fermi energy of $-E_F$:

$$f_h(-E) = \frac{1}{1 + e^{(E_F - E)/kT}}$$

If one recalls that at finite temperatures the valence band is nearly full and the conduction band is nearly empty, then it is intuitively obvious that the Fermi energy must fall somewhere in the middle of the band gap. Consequently, unless the Fermi energy falls within a band or very near the band edge, for any electron or hole state at ordinary temperatures one can safely assume that $|E-E_F|>>kT$, and, consequently that the Fermi-Dirac electron and hole distribution functions can be approximated satisfactorily by ordinary Maxwell-Boltzmann forms:

$$f(E)=e^{-(E-E_F)/kT} \quad ; \quad f_h(-E)=e^{-(E_F-E)/kT}$$

These expressions once again reflect the approximate symmetry in the behavior of holes and electrons. In passing, one should observe that the Fermi energy does not necessarily correspond to the energy of any real electron or hole state. In the most elementary sense, E_F is merely a characteristic parameter of the Fermi-Dirac distribution function. Indeed, for semiconductors (and insulators) the Fermi energy falls within the band gap where, in principle, energy states are absent. However, it is often convenient to consider a hypothetical electron state with energy exactly equal to E_F. This is called the *Fermi level* and is usually represented as a flat line on an aggregate band diagram.

If E_C is defined as the energy at the bottom of the conduction band and E_V as the energy at the top of the valence band, then the difference, E_C-E_V, evidently corresponds to the band gap energy, E_g. Consequently, the concentration of electrons at the conduction band edge, n, can be expressed as follows:

$$n=N_Ce^{-(E_C-E_F)/kT}$$

Similarly, the concentration of holes at the valence band edge, p, is:

$$p=N_Ve^{(E_V-E_F)/kT}$$

The factors, N_C and N_V, respectively define effective electron and hole concentrations at the bottom of the conduction band and top of the valence band under conditions of full occupancy. These values are independent of Fermi energy and depend only on the density of states for a specific semiconductor material. If one multiplies these expressions together, it evidently follows that:

$$np=N_VN_Ce^{-(E_C-E_V)/kT}=N_VN_Ce^{-E_g/kT}$$

This expression is also independent of the Fermi energy and, hence, is independent of changes in carrier concentration. The band gap energy is a characteristic property of the semiconductor material. Thus, carrier concentrations in a semiconductor constitute a mass action equilibrium similar to mass action equilibria frequently encountered in classical chemical systems. Accordingly, the equilibrium constant corresponds to the *intrinsic carrier concentration, n_i*, defined such that:

$${n_i}^2=N_VN_Ce^{-E_g/kT}$$

Thus, n_i, is the concentration of electrons at the conduction band edge in a pure semiconductor. Likewise, n_i is also the concentration of holes at the valence band edge in a pure semiconductor. It follows then that n_i depends only on absolute temperature, T, and material constants of the semiconductor. Furthermore, as a matter of common usage, for description of the electrical properties of semiconductors and unless otherwise explicitly stated, the terms “electron” and “hole” specifically denote a conduction band electron and a valence band hole.

So far, it has been only assumed that the Fermi level must fall very near the center of the band gap. Moreover, it is not difficult to demonstrate that this is, indeed, the case. Thus, from the preceding expressions, one can write:

$$N_C e^{-(E_C - E_F)/kT} = N_V e^{(E_V - E_F)/kT}$$

Upon taking the natural logarithm of both sides, one can readily solve for the Fermi energy to obtain:

$$E_F = \frac{1}{2}(E_C + E_V) + \frac{1}{2} kT \ln \frac{N_V}{N_C}$$

Physically, the parameters, N_C and N_V, are of the same order of magnitude; hence, the natural logarithm is relatively small. Therefore, at ordinary temperatures, the second term of this expression is negligible in comparison to the first term. Clearly, the first term just defines the midgap energy; hence, the Fermi level falls very near the center of the band gap. The aggregate band structure for a pure semiconductor is shown in the following figure:

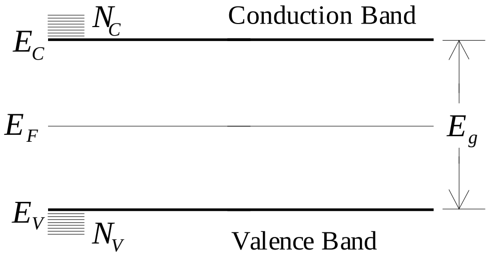


Fig. 4: Aggregate band diagram for an intrinsic semiconductor

As a practical matter, for intrinsic silicon at ordinary temperatures E_F lies slightly below midgap and differs from the midgap energy by only about one percent of E_g.

In passing, one might naively suppose that in analogy to molecular and crystal bonding, the band gap energy in a semiconductor crystal simply corresponds to the covalent binding energy of the solid. However, if this were true, a band gap should always exist within any crystalline solid since overall stability requires significant binding energy. However, in reality, covalent bonds within a crystal cannot be considered in isolation and, moreover, significant interaction between electrons is to be expected, which causes resulting bands to be broadened in energy. Therefore, although the difference of the average energies of conduction band and valence band states can be expected to correspond broadly to binding energy, the band gap energy is the difference of *minimum conduction band energy*, E_C, and *maximum valence band energy*, E_V, and, consequently, band gap energy is typically much smaller that overall binding energy and in some cases disappears altogether if the valence band and conduction band happen to overlap (as in the case of a semimetal). As observed previously, for an insulator, the band gap is quite large, but for a semiconductor the band gap is small enough so that electrons can be promoted from the valence band to the conduction band quite readily. (Again, in real materials the band structure is generally complicated; however, a simple picture assuming flat band edges is generally quite adequate to understand the electronic behavior of devices.)

Extrinsic Doping

Of course, in a pure or intrinsic semiconductor, promotion of an electron from the valence band to the conduction band results in the formation of a hole in the valence band. Alternatively, this process could be thought of as the promotion of a hole from the conduction band into the valence band leaving an electron behind. Within the context of a semiconductor crystal, holes and electrons appear on an equivalent footing. Upon initial consideration, this may seem somewhat strange. Indeed, one is naively tempted to consider electrons as somehow "more real" than holes. It is quite true that electrons can be physically separated, *i.e.*, removed into the vacuum, from a semiconductor crystal, and holes cannot. (As will be seen subsequently, this defines the work function for the semiconductor.) However, a vacuum electron cannot really be considered as physically equivalent to a mobile electron inside of a semiconductor crystal. To understand why this is so, one observes that within the crystal, a mobile electron is more accurately thought of as a local *increase* in electron density rather than a single distinct electron which can, in principle, be separated from the crystal. This picture quite naturally leads one to the view of a mobile hole as a corresponding local *decrease* in electron density. Such fluctuations in electron density are able to propagate throughout the lattice and behave as distinct particles with well-defined effective masses (or, again, more correctly, quasi-particles, since the density fluctuations themselves cannot be removed into the vacuum). Furthermore, the effective mass of a mobile electron within the crystal lattice is quite different (typically much smaller) than the rest mass of an electron in the vacuum. This, again, illustrates fundamental non-equivalence of a vacuum electron and a mobile electron within the lattice. Therefore, within this context, holes quite naturally appear as positively charged "particles" and electrons as negatively charged "particles". The well-known approximate symmetry of conduction band electrons and valence band holes follows directly as a consequence and, as expected, both electrons and holes behave as mobile carriers of opposite electrical charge. Of course, this idea is fundamental to any understanding of practical solid-state electronic devices.

Again, the condition of thermal equilibrium requires that at ordinary temperatures in an intrinsic semiconductor, a small number of electrons will be promoted to the conduction band solely by thermal excitation. To maintain charge neutrality, there clearly must be an equal number of holes in the valence band. Obviously, this concentration is just the intrinsic carrier concentration, n_i, as defined previously, and is a function only of the band gap energy, densities of states, and absolute temperature. However, n_i, is not a function of the Fermi energy. (It shall soon become evident why this assertion is so significant.) In addition, it is important to consider the effect of introduction of a small concentration of so-called *shallow level impurities* such as boron and phosphorus on the electronic structure of silicon. Clearly, the periodic chart implies that phosphorus has five valence electrons per atom rather than four as in the case of silicon. Accordingly, suppose that a phosphorus atom is substituted into the silicon lattice. The phosphorus atom forms the expected four covalent bonds with adjacent silicon atoms. Naturally, the four electrons associated with lattice bonding are simply incorporated into the valence band just as they would be for a silicon atom. However, the fifth electron ends up in a localized shallow electronic state just below the conduction band edge. These states are called *donor states*. It turns out that for temperatures greater than about 100°K, the silicon lattice has sufficient vibrational energy due to random thermal motion to promote this electron into the conduction band where it becomes a mobile carrier. Once in the conduction band, the electron is delocalized, thus, becoming spatially separated from the phosphorus atom. As a consequence, it is unlikely that an electron in a conduction band state will fall back into a donor state (even though it is of lower energy) and, thus, the phosphorus atom becomes a *positively ionized impurity*. Concomitantly, the density of conduction band states is several orders of magnitude larger than the density of donor states and since the Fermi-Dirac distribution implies that at ordinary temperatures the occupancy of any electronic state of energy near E_C is quite small, it follows that any electron of comparable energy is overwhelmingly likely to occupy a conduction band state, regardless of spatial characteristics, *i.e.*, localization or delocalization. In any case, because the energy required to ionize the phosphorus atom is small in comparison to the band gap and is of the same order of magnitude as thermal energy, *kT*, effectively all of the substituted phosphorus atoms are ionized. Thus, addition of a small amount of phosphorus to pure silicon causes the conductivity of the silicon to increase dramatically.

Conversely, suppose that instead of phosphorus, a boron atom is substituted into the silicon lattice. Boron has only three valence electrons instead of four. As a consequence, boron introduces localized empty electronic states just above the valence band edge. These states are called *acceptor states*. In analogy to donor states, for temperatures above about 100°K, the vibrational energy of the silicon lattice easily promotes electrons from the valence band into acceptor states. Alternatively, a boron atom could be thought of as having four valence electrons and a hole, *i.e.*, a net of three valence electrons. From this viewpoint, the ionization process can be considered as the promotion of a hole from an acceptor state into the valence band. In either view, mobile holes appear in the valence band and the boron atom becomes a *negatively ionized impurity*. As before, since holes are mobile it is unlikely that once an electron is in an acceptor state, it will fall back into the valence band. Again, the energy required to ionize the boron atom is small in comparison to the band gap and is of the same order of magnitude as *kT*, hence, just as for phosphorus, effectively all of the substituted boron atoms become ionized. In analogy to the introduction of mobile electrons into the conduction band by donor states, the appearance of mobile holes in the valence band also greatly increases the conductivity of the silicon crystal. (Of course, just as for the case of donor states, acceptor states are also described by Fermi-Dirac statistics, which implies that since they lie near the valence band edge, they are very likely occupied.) In general for silicon, Group VB elements act as *donor impurities* since they can contribute additional electrons to the conduction band. Likewise, Group IIIB elements act as *acceptor impurities* since they can trap electrons from the valence band and, thus, generate additional holes (or one could say that they "donate" holes to the valence band). This is the fundamental mechanism of *extrinsic doping*. (Shallow level impurities are generally called *dopants*.) Conventionally, a semiconductor crystal to which a controlled amount of a donor impurity has been added is said to be *n-type* since its electrical conductivity is the result of an excess concentration of negatively charged mobile carriers (electrons in the conduction band). In contrast, a semiconductor crystal to which a controlled amount of acceptor impurity has been added is said to be *p-type* since its electrical conductivity is the result of an excess concentration of positively charged mobile carriers (holes in the valence band).

In previous consideration of intrinsic semiconductors, it was asserted that the product of the concentration of electrons in the conduction band and holes in the valence band is equal to an equilibrium constant:

$$np = n_i^{\,2} = N_V N_C e^{-E_g/kT}$$

This result is also applicable in the case of extrinsically doped semiconductors. Furthermore, the semiconductor must remain electrically neutral overall. Therefore, if N_A and N_D are identified as concentrations of acceptor and donor impurities in the crystal, then assuming complete impurity ionization, one finds that:

$$0 = p - n + N_D - N_A$$

Combining these two expressions gives:

$$n_n = \frac{1}{2}\left[N_D - N_A + \sqrt{(N_D - N_A)^2 + 4n_i^{\,2}}\right] \quad ; \quad p_p = \frac{1}{2}\left[N_A - N_D + \sqrt{(N_A - N_D)^2 + 4n_i^{\,2}}\right]$$

Here, n_n and p_p are defined as *majority* carrier concentrations, *i.e.*, electron concentration in an *n*-type semiconductor and hole concentration in a *p*-type semiconductor. Therefore, the quantity, $N_A - N_D$ is the *net* impurity concentration. Obviously, if the net impurity concentration vanishes, *n* and *p* are equal to n_i even though acceptor and donor impurities are present. Clearly, such a semiconductor is "intrinsic" and doping is said to be wholly *compensated*. (This condition is not equivalent to a pure semiconductor as will become evident in subsequent discussion of mobility.) Partial compensation occurs if both donor and acceptor impurities are present in different amounts.

In practice, the net concentration of dopants is much larger than the intrinsic carrier concentration. Thus, the preceding expressions simplify as follows:

$$n_n = N_D - N_A \quad ; \quad p_p = N_A - N_D$$

Minority carrier concentrations, p_n and n_p, are easily found from the carrier equilibrium condition.

$$p_n n_n = p_p n_p = {n_i}^2$$

Clearly, the concentration of minority carriers is many orders of magnitude smaller than the concentration of majority carriers for net impurity concentrations above 10^{12} cm^{-3}. Typically, practical doping concentrations are substantially larger than this.

Obviously, since the occupancy of the electron states of the crystal is altered by doping, the Fermi level in a doped semiconductor can no longer be expected to be at midgap. If the doping concentration is not too large, as is usually the case, it is a simple matter to determine the Fermi energy within the context of Maxwell-Boltzmann statistics. One observes that the ratio of mobile carrier concentrations is trivially constructed from fundamental definitions:

$$\frac{p}{n} = \frac{N_V}{N_C} \frac{e^{(E_V - E_F)/kT}}{e^{-(E_C - E_F)/kT}}$$

For convenience, one takes the natural logarithm of both sides to obtain:

$$\frac{kT}{2} \ln \frac{p}{n} = \frac{E_C + E_V}{2} + \frac{kT}{2} \ln \frac{N_V}{N_C} - E_F$$

The first two terms on the right hand side are immediately recognizable as Fermi energy for an intrinsic semiconductor, *i.e.*, the so-called *intrinsic Fermi energy*, which is conventionally denoted as E_i, hence:

$$\frac{kT}{2} \ln \frac{p}{n} = E_i - E_F$$

Of course, as determined previously, E_i lies very close to midgap. Naturally, the majority carrier concentration is determined by the dopant concentration and the minority carrier concentration is then obtained from the carrier equilibrium. Therefore, in the case of an *n*-type semiconductor:

$$\frac{kT}{2} \ln \frac{4n_i^2}{[N_D - N_A + \sqrt{(N_D - N_A)^2 + 4n_i^2}\,]^2} = kT \ln \frac{2n_i}{N_D - N_A + \sqrt{(N_D - N_A)^2 + 4n_i^2}} = E_i - E_F$$

Similarly, in the case of a *p*-type semiconductor:

$$\frac{kT}{2} \ln \frac{[N_A - N_D + \sqrt{(N_A - N_D)^2 + 4n_i^2}\,]^2}{4n_i^2} = kT \ln \frac{N_A - N_D + \sqrt{(N_A - N_D)^2 + 4n_i^2}}{2n_i} = E_i - E_F$$

As observed previously, it is usual for the net impurity concentration to far exceed the intrinsic carrier concentration. In this case, the preceding expressions simplify respectively for *n*-type and *p*-type cases as follows:

$$-kT\ln\frac{N_D-N_A}{n_i}=E_i-E_F \quad ; \quad kT\ln\frac{N_A-N_D}{n_i}=E_i-E_F$$

It is immediately obvious that Fermi energy, E_F, is greater than the intrinsic energy, E_i, for an *n*-type semiconductor and less than E_i for *p*-type. This is illustrated by the following diagrams:

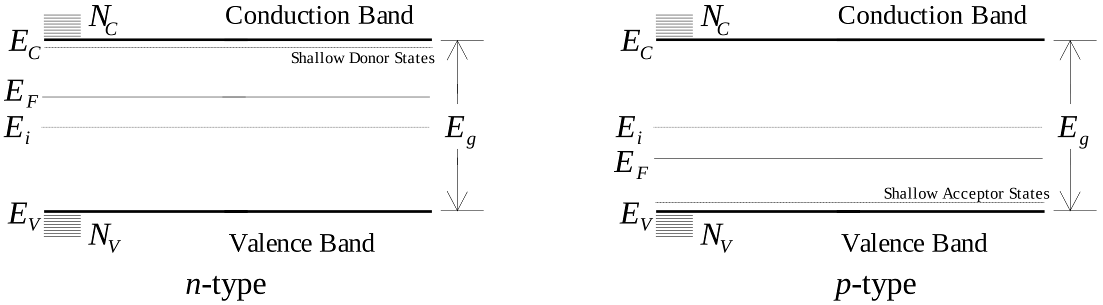


Fig. 5: Aggregate band diagrams for *n* and *p*-type extrinsic semiconductors

One recognizes immediately that this is consistent with the definition of the Fermi level as corresponding to the energy for which the occupation probability is one half. Clearly, if additional electrons are added to the conduction band by donors, then the Fermi level must shift to higher energy. Conversely, if electrons are removed from the valence band by acceptors, the Fermi level shifts to lower energy.

Of course, the Fermi level in extrinsic semiconductor must also depend on temperature as well as net impurity concentration as illustrated in the following figure:

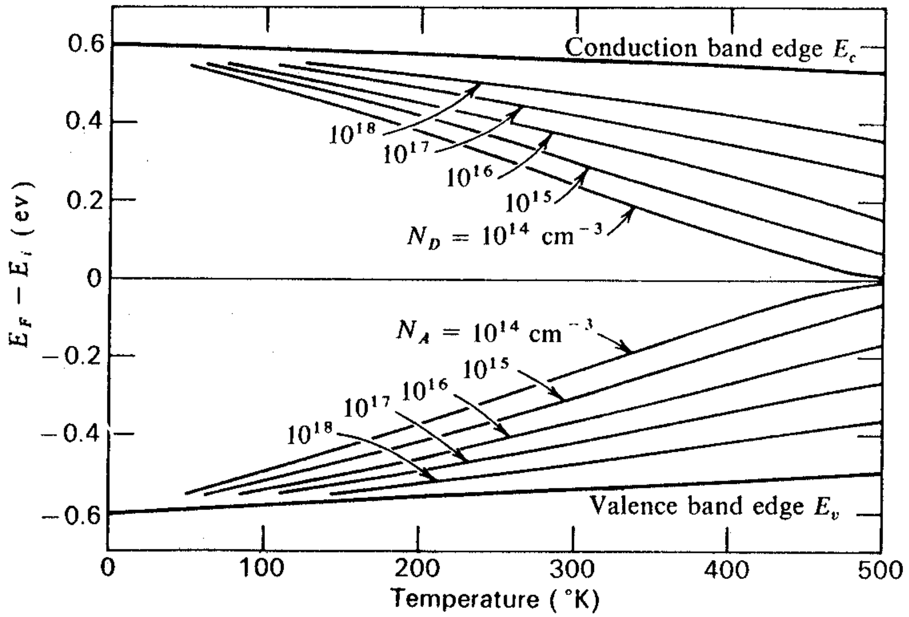


Fig. 6: Effect of temperature on Fermi level shift in extrinsic silicon

Clearly, as temperature increases, for a given impurity concentration, the magnitude of any shift in Fermi level due to extrinsic doping decreases correspondingly. This is easily understood if one observes that total carrier concentration must be the sum of extrinsic and intrinsic contributions. Obviously, at very low temperatures (nearly absolute zero) the carrier concentration is also very low. In this case, the thermal energy of the crystal is insufficient either to promote electrons from the valence band to the conduction band or to ionize shallow level impurities. Hence, mobile carriers are effectively "frozen out" and the crystal is effectively non-conductive. However at about 100°K in silicon, the thermal energy is sufficient for shallow level impurities to become ionized. In this case, the mobile carrier concentration becomes essentially identical to the net dopant impurity concentration. (Thus, the previous assumption of complete dopant impurity ionization in silicon at ordinary temperatures is evidently justified.) Therefore, over the so-called *extrinsic range*, majority carrier concentration is dominated entirely by net dopant concentration and remains essentially fixed until, in the case of silicon, ambient temperatures reach nearly 450°K. Of course, minority carrier concentration is determined by the value of the carrier equilibrium constant, *i.e.*, n_i, at any given ambient temperature. At high temperature, thermal excitation of intrinsic carriers becomes dominant and the majority carrier concentration begins to rise exponentially. Of course, there must also be a corresponding exponential increase in minority carrier concentration. Indeed, as temperature increases further, any extrinsic contribution to carrier concentrations becomes negligible and the semiconductor becomes effectively intrinsic. This behavior is illustrated by the following figure:

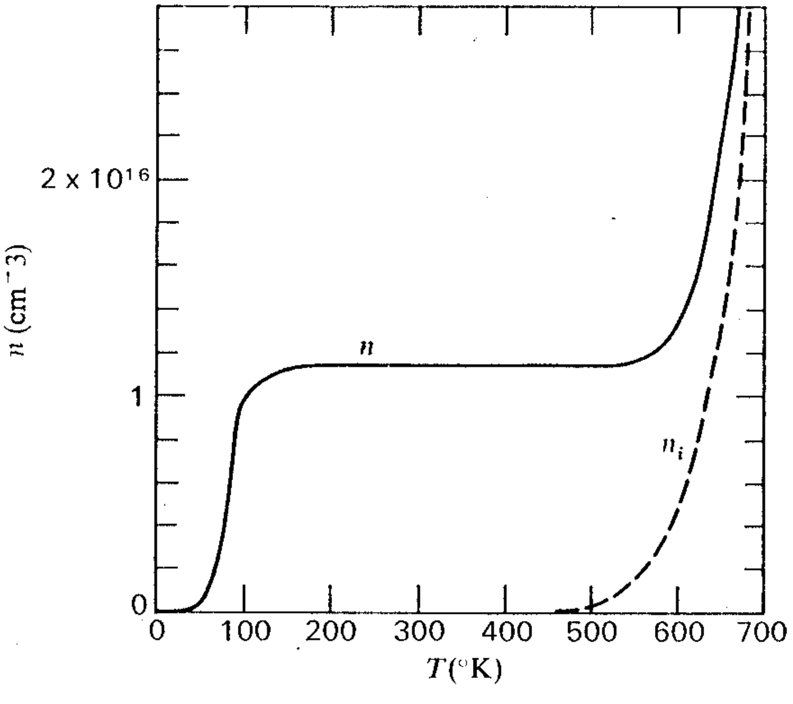


Fig. 7: Effect of temperature on mobile electron concentration in--*solid line*: *n*-type extrinsic silicon (N_D=~1.15(10^{16}) cm^{-3}) and--*dashed line*: intrinsic silicon

Although this figure only illustrates the behavior of *n*-type silicon, one expects that *p*-type silicon has entirely equivalent behavior.

In passing, the difference between the Fermi energy and intrinsic Fermi energy in an extrinsic semiconductor is often expressed as an equivalent electrical potential, *viz.*, the *Fermi potential*, φ_F, which is formally defined, thus:

$$\varphi_F = \frac{\left|E_i - E_F\right|}{q}$$

Here, of course, q is the fundamental unit of charge, $1.602(10^{-19})$ C. In addition, it is important to observe that if net dopant concentration becomes too high, *i.e.*, the Fermi level approaches a band edge, then an elementary analysis can no longer be applied. This is true primarily for two reasons: First of all, it is no longer possible to approximate Fermi-Dirac statistics with Maxwell-Boltzmann statistics for band edge states. In addition, interaction between carrier spins becomes important at very high mobile carrier concentration. (This is characteristic of a so-called degenerate fermion fluid, hence, a semiconductor doped at this level is said to be *degenerate*.) Second, shallow level impurities are no longer effectively all ionized. This is also a consequence of the high carrier concentration, which tends to repopulate ionized shallow level states. Indeed, if the Fermi level coincides exactly with shallow level states of either type (as is the case for very high dopant concentration), then one expects the occupancy of these states to be exactly one half. In practice, this condition is only realized for dopant concentrations in excess of 10^{18} cm^{-3}. Critical dopant concentrations in solid-state electronic devices are typically lower than this.

Carrier Mobility

Naturally, in order for a semiconductor crystal to conduct significant electrical current, it must have a substantial concentration of mobile carriers (*viz.*, electrons in the conduction band or holes in the valence band). Moreover, carriers within a crystalline conductor at room temperature are always in a state of random thermal motion and, analogous to the behavior of ordinary atmospheric gas molecules, as temperature increases the intensity of this motion also increases. Of course, in the absence of an external field net current through the bulk of the crystal must exactly vanish. That is to say that the flux of carriers is the same in all directions and that there is no net electrical current flow out of or into the crystal. However, if an external electric field generated by an external potential difference is applied, then carriers tend to move under the influence of the applied field resulting in net current flow and, consequently, mobile carriers must have some average drift velocity due to the field. (This is exactly analogous to the motion of molecules of a fluid under the influence of hydrodynamic force, *i.e.*, a pressure gradient.)

At this point an obvious question arises; how large is this drift velocity? In principle, it is possible to estimate its size by simple consideration of Newton's Second Law, *i.e.*, "force equals mass times acceleration":

$$F = ma$$

Clearly, the force "felt" by any carrier is $\pm qE$, *i.e.*, just the simple product of carrier charge with electric field strength. Therefore, it follows that:

$$\pm qE = m^* \frac{d\bar{v}}{dt}$$

Here, m^* is identified as *effective carrier mass*. Considering the case of electrons first, of course, m^* cannot be expected to be the identical to vacuum electron mass, but is, in fact, considerably smaller. Accordingly, since a mobile electron within a semiconductor crystal has less inertia, under the influence of an applied force it initially accelerates more rapidly than would a free electron in a vacuum. Physically, this is consistent with a picture of an electron in the conduction band as merely a density fluctuation, rather than a distinct individual particle. Indeed, just as waves on the ocean may move faster than the water itself, electron density fluctuations may propagate more rapidly than the background "fluid", *viz.*, valence electron density. As one might expect, it is also found that holes have an effective mass which is similar in size (usually somewhat greater, due to substantially larger inertia) as effective electron mass. Likewise, a hole in the valence band can also be identified as a fluctuation of background electron density. (Of course, in contrast to electrons, free particle mass for holes must remain undefined.) Furthermore, the difference in effective masses for holes and electrons is a consequence of the detailed band structure. Hence, within the aggregate two band picture of a semiconductor, effective carrier masses are conveniently treated as fundamental material parameters. Accordingly, if one assumes that applied electric field is constant (usually a good approximation on a size scale commensurate with crystal structure), then one can trivially integrate Newton's Second Law:

$$\frac{\pm qE}{m^*}\Delta t = \bar{v}$$

Here, $\bar{v}$ is defined as *carrier drift velocity*. Obviously, electrons and holes have drift velocities of opposite sign (*i.e.*, they drift in opposite directions) due to their opposite electric charges. Clearly, this formulation assumes that carriers do not accelerate indefinitely under the influence of an applied field, but rather, that they reach a limiting drift velocity due to scattering within the bulk of the crystal.

Concomitantly, it is not at all clear what value should be used for Δt in the preceding expression. Thus, to estimate Δt, one must return to the equilibrium picture of carrier motion in a solid. In this case, just as for an ordinary gas, it is possible to define a mean time between collisions, τ_{col}. However, for a semiconductor, collisions between carriers and the crystal lattice itself dominate rather than collisions between carriers only as would be analogous to the behavior of gas molecules. Consequently, scattering can be greatly increased by the presence of defects, impurities, *etc.* In any case, if one formally substitutes τ_{col} for Δt, then one can write:

$$\bar{v} = \frac{\pm qE}{m^*}\tau_{col}$$

Thus, one finds that carrier drift velocity is proportional to applied field and the associated constant of proportionality is called *carrier mobility*, μ:

$$\bar{v} = \pm\mu E$$

In general, hole and electron mobilities are conventionally considered as specific properties of the semiconductor itself and are treated as characteristic material parameters.

For completeness, it should be further emphasized that this primitive treatment of carrier motion is only applicable when external electric field strength is relatively low (<1000 V/cm) and, conversely, that at high values of the field, carrier drift velocity is no longer simply proportional to applied field, but tends toward a constant value called the *saturated drift velocity*. Indeed, it is quite easy to understand the cause of this behavior. At low field strength, random thermal motion is the dominant part of overall carrier motion and as a consequence, τ_{col} must be effectively independent of field strength. In contrast, at high field directed carrier motion dominates. In this case, τ_{col} becomes inversely related to the field strength and, consequently, tends to cancel out the field dependence of the carrier drift velocity. Alternatively, saturated drift velocity can be considered as broadly analogous to terminal velocity for an object "falling" through a constant field of force (such as a sky diver falling through the atmosphere before opening his or her parachute). Within this context, electron and hole mobilities in pure silicon are illustrated in the following figure as functions of electric field strength:

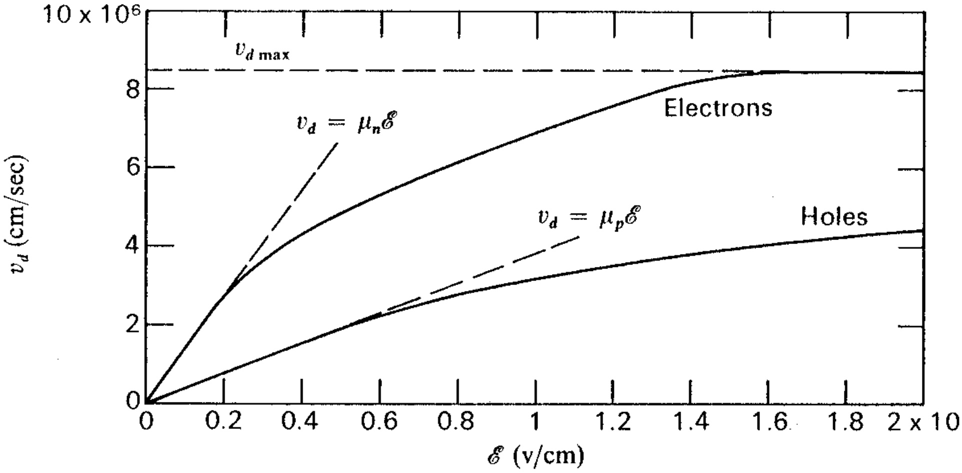


Fig. 8: Effect of applied electric field on carrier drift velocity (intrinsic silicon)

Clearly, the saturated drift velocity for electrons is approximately $8.5(10^6)$ centimeters per second and for holes about half that value.

Of course, mean time between collisions, τ_{col}, can be expected to be determined by various scattering mechanisms; however, a fundamental mechanism is found to be *lattice scattering*. This occurs because the crystal lattice always has finite vibrational energy characteristic of ambient temperature. Indeed, fundamental quantum mechanical principles require that this energy does not vanish even in the limit that the lattice temperature falls to absolute zero. However, if the vibrational energy characteristic of the lattice becomes sufficiently small (*i.e.*, temperature is very low), scattering becomes insignificant and superconductivity is observed. Of course, this occurs only at deep cryogenic temperatures. In any case, quantum mechanical treatment of lattice vibration results in the definition of fundamental vibrational quanta called *phonons*. Indeed, these are similar to the quanta of the electromagnetic field, *i.e.*, photons, and, as such, exhibit particle-like behavior within the crystal. Without going into great detail, lattice scattering can be thought as the result of collisions between mobile carriers (electrons and holes) and phonons.

A second important scattering mechanism is *ionized impurity scattering*. Essentially, this is a consequence of electrostatic, *i.e.*, Coulombic, fields surrounding ionized impurity atoms within the crystal. Of course, ionized impurities are introduced into the crystal lattice by extrinsic doping and, thus, one should expect that, in general, carrier mobilities will decrease as doping increases. This is, indeed, observed experimentally. Furthermore, an important point to be made is that although donors and acceptors become oppositely charged within the lattice, overall effects of electrostatic scattering do not depend on the sign of the charge. Therefore, both donor and acceptor atoms can be expected to have similar influence on carrier mobility. Clearly, this implies that in contrast to the effect of ionized impurities on carrier equilibrium for which donors and acceptors compensate each other, the effect of ionized impurities on carrier scattering is cumulative and, as such, compensation should not be expected. Accordingly, in contrast to carrier concentrations which depend on the difference of acceptor and donor concentrations, ionized impurity scattering depends on the sum of acceptor and donor concentrations. Therefore, although acceptor and donor impurities compensate each other in terms of extrinsic doping, they still invariably contribute to reduction of carrier mobility. Accordingly, carrier mobility in compensated intrinsic semiconductor will necessarily be lower than in pure semiconductor (which is also, of course, intrinsic). Within this context, the effect of impurity concentration on overall carrier mobilities in silicon (at a nominal ambient temperature of 300° K) is illustrated below:

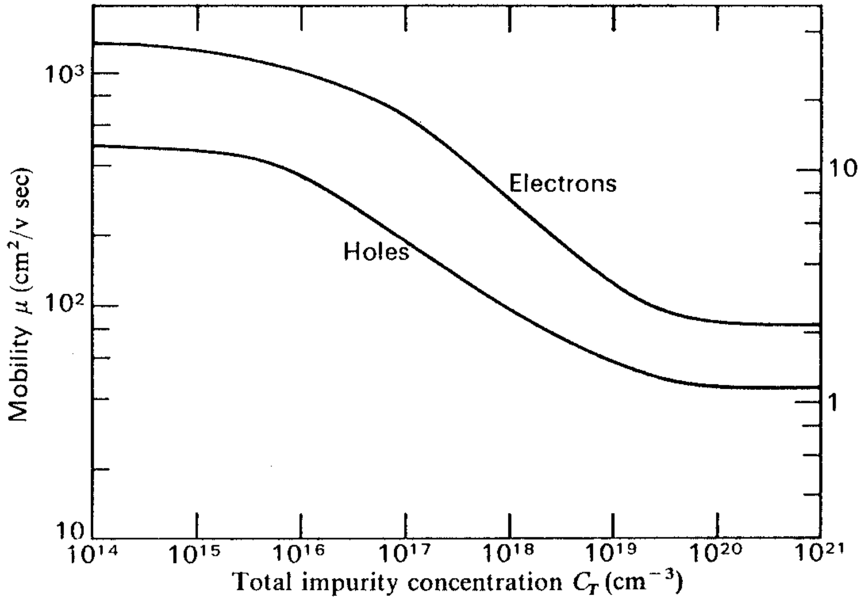


Fig. 9: Effect of total impurity concentration on carrier mobility (and diffusivity) in silicon

Moreover, the curves appearing in the figure can be satisfactorily represented by the empirical formula:

$$\mu_x = \mu'_x + \frac{\mu_x^0 - \mu'_x}{1 + \left(\dfrac{C_I}{\alpha_x}\right)^{\lambda}}$$

This kind of mobility model was first formulated in the late 1960's by Caughey and Thomas. By convention, x can denote either electrons or holes, *i.e.*, *e* or *h*, denoting electrons or holes, respectively, and μ_x^0 and μ'_x are evidently to be interpreted as intrinsic and compensated carrier mobilities, which, respectively, for electrons and holes have typical values and 1414 and 471 centimeters squared per volt second and 68.5 and 44.9 centimeters squared per volt second. Likewise, α_x is defined as a "reference" impurity concentration, which, again, for electrons and holes has typical values of 9.20(10^{16}) and 2.23(10^{17}) impurity atoms per cubic centimeter. In contrast, the exponent, λ, differs little between electrons and holes typically having values of 0.711 and 0.719, respectively.

For completeness, an obvious third scattering mechanism that one could corresponds to collisions between mobile carriers of opposite charge (*i.e.*, scattering of electrons by holes and vice versa). However, as observed previously, in doped semiconductors, one type of carrier generally predominates so that the effect of this type of scattering is negligible. Similarly, scattering between carriers of the same type is also possible; however, since total momentum and electric charge is rigorously conserved during such a collision, there is no net effect of this type of scattering on the mobility. Therefore, as stated at the outset, carrier-carrier scattering processes can be ignored in comparison to carrier-lattice scattering processes.

Within this context, one can consider the current, *I*, flowing in a hypothetical rectangular crystal of heavily doped *n*-type semiconductor:

$$I = -qn\bar{v}A = qn\mu_e \frac{\Delta V}{L} A$$

Here, μ_e is electron mobility, ΔV is an applied potential, *L* is the physical length of the crystal, and *A* is its cross sectional area. If one recalls *Ohm's Law*, resistance can be defined as follows:

$$R_n = \frac{1}{qn\mu_e}\frac{L}{A}$$

A similar expression is obtained for the flow of holes in a *p*-type rectangular semiconductor crystal:

$$R_p = \frac{1}{qp\mu_h}\frac{L}{A}$$

Here, of course, μ_h is hole mobility. Clearly, for an intrinsic or lightly doped semiconductor, flow of both electrons and holes must be considered. Accordingly, resistance is given by the general formula:

$$R = \frac{1}{q(n\mu_e + p\mu_h)}\frac{L}{A}$$

Obviously, this expression is applicable to any doping concentration since majority carriers dominate for heavy doping (thus, simplifying this expression to the two forms appearing above) and, moreover, evidently corresponds to the conventional combination of two parallel resistances. Therefore, *resistivity*, ρ, which is an intensive material property (*i.e.*, independent of size and shape of the material), is identified thus:

$$\rho = \frac{1}{q(n\mu_e + p\mu_h)}$$

Of course, by definition, resistivity is the reciprocal of *conductivity*.

Clearly, resistivity is generally a function of carrier concentrations and mobilities. Therefore, one expects that the resistivity of *p*-type silicon should be higher than the resistivity of *n*-type silicon having equivalent net dopant impurity concentration. Furthermore, since the dependence of mobility on impurity concentration is relatively weak, if net dopant impurity concentration is increased, then resistivity should decrease due to a corresponding increase of majority carrier concentration (even though ionized impurity scattering may further reduce mobility). This behavior in uncompensated extrinsic silicon is illustrated by the following figure:

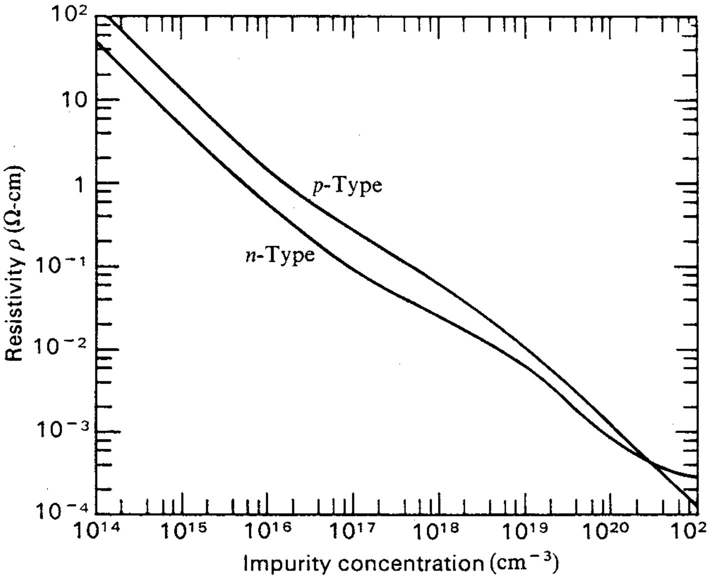


Fig. 10: Resistivity of bulk silicon as a function of impurity concentration

Clearly, determination of resistivity requires determination of both mobility and carrier concentration. Of course, carrier concentration depends on *net impurity concentration*, *i.e.*, the difference between donor and acceptor concentrations. However, it is *total impurity concentration*, *i.e.*, the sum of donor and acceptor concentrations that must be considered to determine mobility. Obviously, the behavior of resistivity as a function of impurity concentration is not simple and depends on the combined behavior of carrier concentrations and mobilities. (However, the "kink" in the plot of resistivity at high values of impurity concentration which is evident in the preceding figure can be attributed to reduction of carrier mobilities due to ionized impurity scattering.)

As might be expected, at ordinary temperatures it is found to good approximation that temperature dependence of the contribution of lattice scattering to carrier mobility, denoted as, μ_x^L, is described by an inverse power law. (As usual, x denotes either electrons or holes as is appropriate and T denotes absolute temperature.) Accordingly, μ_x^L can be represented as follows:

$$\mu_x^L(T) = \mu_x^L(T_0)\left(\frac{T_0}{T}\right)^{\nu}$$

By definition, T_0 is some predefined reference temperature (typically 300° K). Here, ν is a positive exponent, which from fundamental theoretical considerations can be expected to have a value of $^3\!/_2$, but which empirical measurements indicate has a value close to 2 for intrinsic silicon. In contrast, temperature dependence of the contribution of ionized impurity scattering to mobility can be described by a normal power law, thus:

$$\mu_x^I(T) = \left(\frac{\alpha_x(T_0)}{C_I}\right)^{\lambda}\left(\frac{T}{T_0}\right)^{\gamma}$$

In analogy, to the case of lattice scattering, γ is a positive exponent, which for moderate impurity concentration, again, has a theoretical value of $^3\!/_2$, but falls to zero for very high concentration. (Indeed, empirical measurements indicate that γ may even become negative for very high concentrations, thus, inverting the power law dependence.) Of course, C_I is total concentration (rather than net concentration) of both positively and negatively charged ionized impurities, *i.e.*, C_I=N_A+N_D, $\alpha_x(T_0)$ is a constant parameter (again, to be regarded as reference impurity concentration) defined at the reference temperature, T_0,, and the exponent, λ, is an adjustable parameter, which frequently in the case of small variation in doping concentration for simplicity can be taken as unity. (However, in practice, values of λ between $^2\!/_3$ and ¾ are generally more realistic.)

Naturally, total mobility, μ_x, for carrier type, x, can be constructed from partial contributions specific to different scattering mechanisms in analogy to the combination of parallel resistances. Therefore, $\mu_x(T)$ is immediately obtained as a sum of reciprocals, thus:

$$\frac{1}{\mu_x(T)} = \frac{1}{\mu_x^L(T)} + \frac{1}{\mu_x^I(T)}$$

In passing, it should be noted that this formulation describes actual behavior of carrier mobility reasonably well for impurity concentrations below ~10^{18} per cubic centimeter, but much more poorly for higher concentrations approaching dopant solubility limits in which case carrier mobilities are severely underestimated. Even so, this expression is useful for explicit consideration of the temperature dependence of electron and hole mobilities in moderately doped silicon. Furthermore, an explicit expression for the behavior of resistivity as a function of temperature is readily constructed. Again, for extrinsic silicon if one considers majority carriers only, then one can write:

$$\rho(T) = \frac{1}{qn_x\mu_x^L(T_0)}\left(\frac{T}{T_0}\right)^{\nu} + \frac{1}{qn_x}\left(\frac{C_I}{\alpha_x(T_0)}\right)^{\lambda}\left(\frac{T_0}{T}\right)^{\gamma}$$

Here, n_x is n if x is e and n_x is p if x is h. Clearly, if C_I vanishes, ρ is a monotonic increasing function of T. However, if C_I is non-zero, then ρ may reach a minimum at some finite temperature. To determine this temperature, one differentiates the above expression as follows:

$$\frac{d}{dT}\rho(T) = \frac{\nu T^{\nu-1}}{qn_x\mu_x^L(T_0)}\left(\frac{1}{T_0}\right)^{\nu} - \frac{\gamma T_0{}^{\gamma}}{qn_x T^{\gamma+1}}\left(\frac{C_I}{\alpha_x(T_0)}\right)^{\lambda}$$

If one sets the derivative equal to zero and solves for the temperature, one obtains the result:

$$T_{min} = \left(\frac{\gamma}{\nu}\mu_x^L(T_0)\left(\frac{C_I}{\alpha_x(T_0)}\right)^{\lambda}\right)^{\frac{1}{\gamma+\nu}} T_0$$

Clearly, at low temperature, if C_I is sufficiently large so that T_{min} is >~100° K, resistivity is dominated by ionized impurity scattering. In this case, ρ decreases with temperature and the semiconductor is said to have a negative *thermal coefficient of resistance* (TCR). In contrast, at higher temperature and/or lower dopant concentration, lattice scattering dominates resistivity. In this case, ρ increases with temperature and the semiconductor is said to have a positive (or normal) TCR.

Crystal Structure

It is useful to begin any discussion of crystal structure with an elementary definition of some crystallographic terms. First of all, crystals are made up of identical, repeating arrangements of atoms called *unit cells*. By definition, a unit cell is the smallest volume of a crystalline solid that exhibits the symmetry properties of the whole crystal. Just seven basic crystal systems are known. These are, cubic, hexagonal, tetragonal, orthorhombic, trigonal, monoclinic, and triclinic. These may be further classified into fourteen Bravais lattice types:

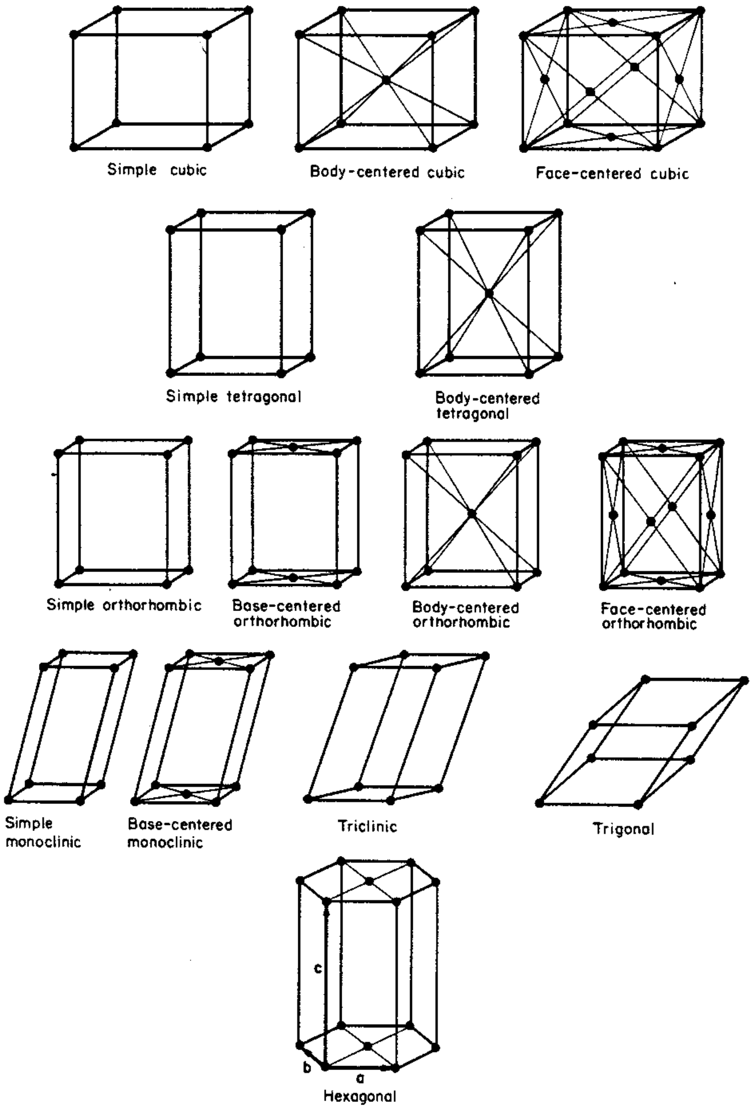


Fig. 11: The fourteen Bravais lattices

Of these types, for semiconductors the face centered cubic (FCC) system is of most importance. Fortunately, cubic lattices are also the easiest Bravais lattices to visualize and understand.

It is clear from the elementary structure of the Bravais lattices that each unit cell has several *lattice points*. In terms of the actual physical structure of a solid material, each lattice point is associated with a *basis group*. Thus, the lattice basis group for a particular crystal is a definite group of atoms associated with each lattice point. In the simplest case (which, for example, occurs in the case of some elemental metals) the basis group consists of just a single atom, in which case one atom occupies each lattice point. Of course, the lattice basis group must be identical for all lattice points. Obviously, for compound materials the basis group must consist of more than one atom since it cannot be defined as one kind of atom at one lattice point and another kind of atom at a different lattice point. Thus, one finds that in general, the basis group of a lattice consists of a definite repeating group of atoms. In some cases, the basis group may be identified with an actual covalent molecule that maintains its identity even when the lattice breaks up during melting or sublimation, (*e.g.*, as in the case of ice, water, and water vapor). However, in many cases, the basis group also breaks up with the lattice itself during change of phase (*e.g.*, as in the case of metallic or ionic solids). Finally, basis groups and unit cells should not be confused. Both are repeating groups of atoms, however, the basis group does not exhibit all of the symmetry properties of the whole crystal.

It is found that even for some elemental materials, the basis group consists of more than one atom. This is precisely the case for elemental silicon for which the lattice basis group consists of two atoms. To understand why this is so, one observes that the Bravais lattice for silicon is easily identified from powder x-ray diffraction patterns as FCC. However, one also recalls from previous consideration of electronic structure, that silicon has tetrahedral coordination due to the tetrahedral geometry of the sp^3 hybrid orbitals. Clearly, if a single silicon atom is inserted at each point of an FCC lattice, the resulting atomic coordination is not tetrahedral. However, if a two-atom basis group is inserted in the FCC lattice, tetrahedral coordination can be realized. The result is the *diamond cubic* structure, which can be thought of as two interpenetrating FCC lattices offset one quarter of the unit cell dimension in each direction:

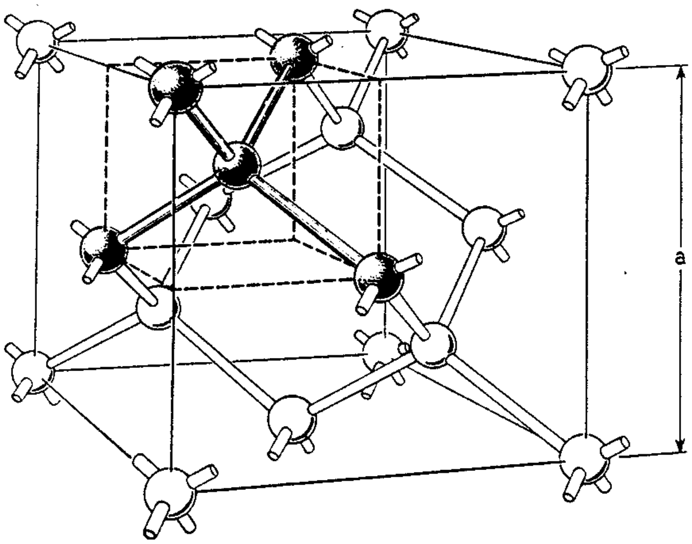


Fig. 12: Diamond cubic crystal structure

Obviously, the archetype of this structure is elemental carbon, *viz.*, diamond. It is well known that the diamond cubic structure has a high degree of symmetry including several mirror planes and two, three, and fourfold rotation axes. This is a direct consequence of the high degree of symmetry associated with tetrahedral coordination. In general, the group of symmetry properties serves as a unique specification of any crystal structure. In passing, one should note that the compound semiconductor, gallium arsenide, GaAs, has essentially the same structure as elemental silicon or germanium. However, for GaAs, the lattice basis group must consist of two different types of atoms, *viz.*, one Ga and one As, instead of two identical atoms, *i.e.*, either Si or Ge. This also corresponds to a cubic crystal structure, similar to the diamond cubic structure, called *zincblende* (after the naturally occurring mineral form of zinc sulfide, ZnS).

Another important term one often encounters in connection with semiconductor materials is *crystal orientation*. To see what is meant by crystal orientation, one returns to the basic unit cell which in the most general case, *i.e.*, a triclinic system, must be described by six independent *lattice parameters* (three lengths and three angles). However, again because of its high symmetry, a cubic system requires only a single lattice parameter, *a*, namely, the unit cell length. (Clearly, for a cubic system, the unit cell length is the same in all three directions and all the angles are 90°.) Thus, one naturally defines a coordinate system for the solid with the lattice parameter as the basic unit of length. Thus, in this "direct" space, any point within the crystal having all integer coordinates corresponds to a corner lattice point, *i.e.*, vertex, of a unit cell.

Crystal orientation is generally specified by *Miller indices*, (h,k,l). The exact definition of Miller indices as coordinates in "reciprocal" space is quite technical in nature and is beyond the scope of the present course; however a simple definition may suffice in terms of the three direct space coordinate axis intercepts. Since three points uniquely determine a geometric plane, the three axis intercepts define a *crystallographic plane*. Perhaps, the simplest example is provided by a plane that intersects each axis at a distance of one lattice parameter from the origin, *i.e.*, the plane has unit intercepts on each coordinate axis. This is primitively identified as the [111] crystallographic plane. In contrast, if one considers a plane that is parallel to the y and z axes, but has unit intercept on the x axis, how should this plane be designated? Clearly, from a strictly geometric point of view, the corresponding y and z axis intercepts for such a non-intersecting plane have "receded to infinity". It is found that a consistent designation results if the Miller indices, h, k, and l, are identified as reciprocals of the x, y, and z axis intercepts instead of the intercepts themselves. Thus, the reciprocal of ∞ formally corresponds to zero and the plane in question is consequently designated as [100]. Using this scheme, a plane that is parallel to the z axis, but has unit intercepts on the x and y axes is designated as [110]. Clearly, the case of [111] remains unchanged since unity is self-inverse. Indeed, it turns out that the definition of Miller indices as reciprocals of coordinate axis intercepts quite naturally follows from a general description of crystal diffraction phenomena in terms of three dimensional Fourier transforms. A further important observation to be made at this point is that since the coordinate origin in direct space is always translatable by an integral number of lattice parameters, the designation [hkl] really defines a *family* of planes rather than a single unique plane. (These families of planes take on particular importance for description of crystal diffraction phenomena.) Furthermore, although crystallographic planes can be arbitrarily specified, they are generally only useful when they correspond to a plane of atoms having a regular pattern, *i.e.*, a two dimensional lattice.

Digressing briefly, it is important to note that within the general context of semiconductor processing, the designation of crystallographic planes (and crystal orientation) is generic in nature. Of course, in a strict sense Miller indices, [100], [010], and [001] must denote different families of crystallographic planes; however, all of these planes correspond to a face of the unit cell. Accordingly, again, as a consequence of the high degree of symmetry characteristic of the diamond cubic structure, all of these planes are electrically and structurally equivalent. Thus, any crystallographic plane corresponding to a face of the unit cell is generically called a “[100]” plane. Likewise, any crystallographic plane that “cuts” a unit cell face diagonally and is also parallel to an edge of the unit cell is called a “[110]” plane. Similarly, any crystallographic plane that intersects three non-adjacent vertices of the unit cell is called a “[111]” plane. These generic planes are illustrated below:

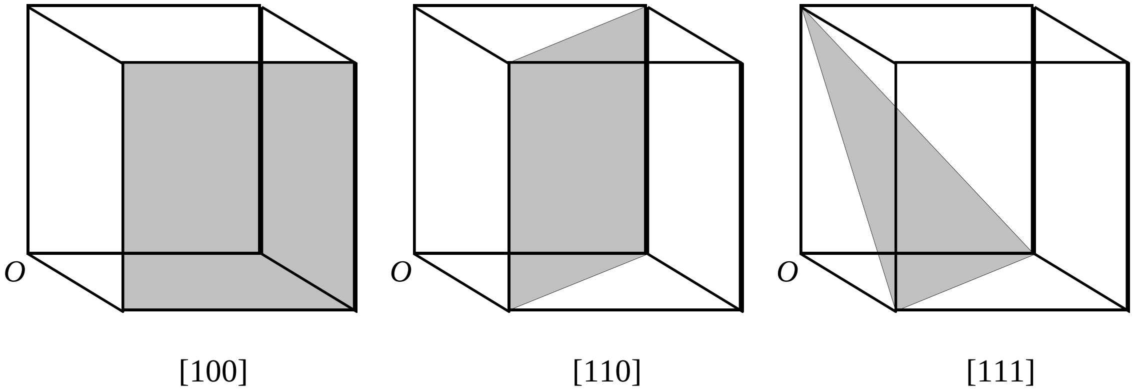


Fig. 13: Representative crystallographic planes

Clearly, in each unit cell, there must be three different, *i.e.*, non-parallel, [100] planes, six different [110] planes, and four different [111] planes.

In practice, physical crystals are generally designated by the orientation of their surfaces. In the case of semiconductor crystals, *i.e.*, silicon substrates, this is the crystallographic plane parallel to the surface used for device fabrication. Thus, a silicon crystal, *i.e.*, a substrate, which has a surface parallel to the unit cell face, is designated [100]. Similarly, if the surface can be thought of as intersecting three opposite corners of the unit cell, then the crystal is designated [111]. These two orientations are essentially the only ones ever used for device fabrication with [100] being much, much more common since it is used exclusively for fabrication of CMOS devices. A small amount of [111] material is still used for fabrication of bipolar devices for which very shallow doping is desirable. In principle, [110] substrates could be manufactured; however such material is of no real practical use and, as such, is very rare.

Clearly, the atomic arrangement for each of these surfaces, *i.e.*, [100], [111], and [110], is different. Directly related to this is the density of *dangling bonds* that are left behind if one breaks or “cleaves” a crystal parallel to a particular crystallographic plane. Naturally, this density is directly related to the surface energy of a crystal of a particular orientation. Therefore, one can primitively estimate relative surface energy by simply counting the number of bonds broken per unit cell if one cleaves the crystal along a given plane and then dividing by the surface area of that plane within the unit cell. (In reality, this simple picture is generally complicated by the occurrence of *surface reconstruction*, which lowers surface energy by overlapping neighboring dangling bonds.) One finds that in the diamond cubic structure, for a [100] plane, four bonds per unit cell must be broken in order to cleave the crystal. Obviously, the unit cell area associated with this cleavage just equals the square of the lattice parameter, a^2. Similarly, for cleavage parallel to a [111] plane, only three bonds per unit cell must be broken. Obviously, a [111] plane forms an equilateral triangle within the unit cell with each side having a length of $a\sqrt{2}$. The height of the equilateral triangle is $a\sqrt{3/2}$, thus, the corresponding area is $a^2\sqrt{2\times 3/2}/2 = a^2\sqrt{3}/2$. Finally, for [110] cleavage, again only three bonds must be broken. The corresponding area of a [110] plane within the unit cell is, of course, $a^2\sqrt{2}$. One, therefore, can determine primitive dangling bond densities, n_{100}, n_{111}, and n_{110} as follows:

$$n_{100} = 4/a^2 \qquad n_{111} = 2\sqrt{3}/a^2 \cong 3.8/a^2 \qquad n_{110} = 3\sqrt{2}/2a^2 \cong 2.1/a^2$$

Clearly, the dangling bond density for a [110] plane is much lower than either for a [100] or a [111] plane. Therefore, it follows that the crystal binding energy across a [110] surface is particularly low. Hence, a diamond cubic crystal should be much more easily broken parallel to a [110] surface than parallel to either a [100] or [111] surface. It turns out that [110] surfaces do indeed correspond to the natural *cleavage planes* of a silicon crystal. This can be demonstrated rather dramatically by breaking silicon substrates of different orientation from the center using a diamond-tipped scribe. One finds that [100] substrates naturally separate into quarters and [111] substrates into sixths.

Crystal Growth

Considering more practical matters, integrated circuits are fabricated on single crystal silicon substrates which are mirror-like, polished circular disks called *wafers* (4”/100 mm dia. 525 µm thick, 6”/150 mm dia. 675 µm thick, 8”/200 mm dia., and 12”/300 mm dia., *etc.*). At present, the largest wafers commonly used in integrated circuit manufacturing are 300 mm although there is still substantial usage of smaller sizes, *viz.*, 200 mm. In addition, substrates of 450 mm (18”) are under development by wafer manufacturers and are currently being used by a few manufacturers. Wafers are cut as slices from large single crystal ingots of silicon called *boules*. Of course, this silicon ultimately comes from quartzite sand, which is a naturally occurring mineral form of silicon dioxide, SiO_2. Typically, the raw oxide is reduced to *metallurgical grade silicon* in an electric furnace using carbon as the reducing agent:

$$\mathrm{SiO_2 + 2C \rightarrow Si + 2CO}$$

This material still contains substantial impurity and to obtain *electronic grade silicon*, metallurgical grade silicon is reacted with hydrogen chloride or chlorine:

$$\mathrm{Si + 3HCl \rightarrow SiHCl_3 + H_2}$$

$$\mathrm{Si + 2Cl_2 \rightarrow SiCl_4}$$

This produces volatile chlorides, *i.e.*, trichlorosilane ($SiHCl_3$) or tetrachlorosilane ($SiCl_4$), which are then carefully distilled and reduced again to silicon by pyrolysis in pure hydrogen (or some other high quality reducing agent). The resulting electronic grade polycrystalline material is quite pure and has as little as 0.05 ppb (50 ppt) of residual boron as the most common impurity.

There are two major crystal growth techniques. These are the *Czochralski* or CZ process and the *float zone* or FZ process. In the CZ process, electronic grade polycrystalline silicon is placed in a quartz crucible surrounded by graphite heat shielding and then heated to the molten state in an inert atmosphere by electrical heating elements. A seed rod having proper orientation is dipped into the melt and then controllably withdrawn. Naturally, the seed is much smaller in diameter than the desired crystal and, thus, the initial stage of CZ growth requires solidification outward from the seed to establish the desired diameter. Once diameter is established, the boule is slowly and controllably withdrawn such that molten silicon solidifies with the desired orientation and a large crystal is built up or “pulled”. During the growth process, both the boule and the crucible are rotated to enhance uniformity. Of course, all of this requires very precise measurement and control of temperature and heat flux, which generally can be achieved for large diameter crystals only by sophisticated computer feedback and control. Typically, a CZ apparatus can hold many kilograms of molten silicon. (This, of course, depends on the size of the wafers being produced and has greatly increased over the last forty years.) The resulting boule is substantially free of crystal defects; however, it does contain oxygen contamination arising from the quartz crucible and carbon contamination from elsewhere (graphite shields, susceptor, *etc.*). In contrast, in the FZ method, a solid electronic grade polycrystalline silicon rod is recrystallized and then refined using a "needles eye" furnace. Thus, the FZ method is essentially a classical *zone refining* technique. It produces the highest purity silicon available; however FZ silicon is generally quite brittle. Indeed, oxygen and carbon contamination in CZ silicon both enhances mechanical strength and allows for the application of internal gettering methods. For these reasons, the vast majority of silicon wafers used in integrated circuit fabrication are manufactured using the CZ process.

The Czochralski Process

From the point of view of integrated circuit manufacturing, it is desirable for the starting silicon wafers not to be intrinsic, but rather to be uniformly doped with some shallow level impurity, *i.e.*, B, P, As, Sb, *etc.* Thus, as well as crystal orientation, background doping (hence, majority carrier type and resistivity) is generally specified when wafers are purchased. Therefore, it is usual for the melt, hence, the grown ingot to be intentionally "contaminated" with a known quantity of shallow level dopant impurity. Therefore, control of dopant concentration and distribution during a CZ growth process is of fundamental importance.

As an initial description of Czochralski growth, it is usual to assume *rapid stirring* conditions, which implies that any excess impurity that might exist in the immediate vicinity of the growth interface is quickly dispersed into the melt. Consequently, such conditions imply that the melt is thoroughly mixed and, accordingly, concentration of any impurity is uniform throughout the melt, *i.e.*, right up to the freezing interface. Physically, this corresponds to very slow growth of the crystal such that the rate that impurity is dispersed throughout the melt is large in comparison to the rate that impurity is incorporated into the freezing crystal. Furthermore, one can safely assume that impurity diffusion within the solid crystal itself is unimportant since the diffusion coefficient of impurity in the liquid is many orders of magnitude larger than the corresponding solid diffusion coefficient. The CZ growth process can be represented pictorially as follows:

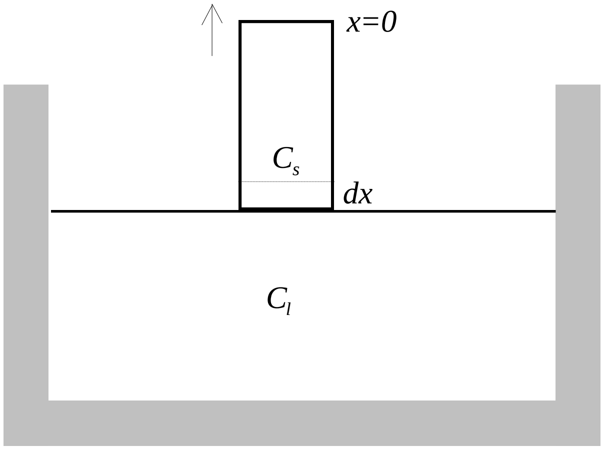


Fig. 14: Schematic diagram of the Czochralski (CZ) process

The arrow denotes the pulling direction. The dimension, x, is the length of crystal pulled from the melt, hence, the mass of crystal solidified, W, is just ρAx, such that A is the cross sectional area of the ingot (assumed to be uniform) and ρ is the density of silicon. Naturally, C_l is the volume concentration of solute, *i.e.*, dopant impurity, atoms in the melt and C_s is the volume concentration of solute atoms in the solid crystal. Clearly, at any given point during crystal growth and under the assumption of rapid stirring, C_l is uniform throughout the melt; however, C_s is a function of the position along the boule, *i.e.*, a function of x. Accordingly, if S is defined as the total number of solute atoms within the melt, then the differential number of solute atoms lost from the melt due to freezing a differential length of crystal is:

$$dS = -C_s A dx = -\frac{C_s}{\rho} dW$$

Here, the negative sign is formally included to indicate that atoms are lost from the melt to the growing crystal. Clearly, for a differential change in the length of the boule, *i.e.*, differential pull distance, dx, a corresponding differential mass of solid, dW, is added to the crystal. Obviously, these differential quantities are related simply as follows:

$$dW = \rho A dx$$

At an arbitrary point in the crystal growth for which a crystal of length, x, and mass, W, has been solidified, the mass of the remaining melt is just the difference, $W_o - W$, such that W_o is the initial mass of the melt before any crystal has been solidified, *i.e.*, pulled. Thus, it follows that the concentration of solute in the melt is given by:

$$C_l = \frac{S\rho}{W_o - W}$$

One formally solves this expression for S and combines it with the differential expression to obtain:

$$\frac{dS}{S} = -\frac{C_s dW}{\rho\left(\frac{C_l}{\rho}(W_o - W)\right)} = -\frac{C_s}{C_l}\frac{dW}{W_o - W}$$

This is a differential expression that relates the concentration of impurity atoms in the melt to the mass of the crystal at any stage during the growth process.

At this point, one might naively assume that the concentrations of impurity in the melt and in the solid are exactly equal; however, this is not the case. At a definite temperature such situations are described by a thermodynamic distribution equilibrium characterized by a constant coefficient, K, which is formally defined in the present case as the concentration ratio, C_s/C_l. In physical terms, one can regard this equilibrium as consequence of the fact that impurity atoms do not “fit” into the crystalline silicon lattice as well as silicon atoms. Therefore, one expects that impurity atoms will be incorporated into the freezing solid at a lower intrinsic rate than are silicon atoms themselves and, thus, impurity atoms will tend to be "rejected" back into the melt. Accordingly, *K* can be expected to be generally less than one. As is established by experimental observations and shown in the following table, this is generally found to be the case:

Dopant	*K*
B	0.72
P	0.32
As	0.27
Sb	0.020
Ga	0.0072
Al	0.0018
In	0.00036

Table 1: Distribution Coefficients for shallow level impurities in silicon

Therefore, the preceding expression takes the form:

$$\frac{dS}{S} = -K\frac{dW}{W_o - W}$$

This expression is readily integrated directly as follows:

$$\int_{S_o}^{S}\frac{dS'}{S'} = -K\int_{0}^{W}\frac{dW'}{W_o - W'}$$

As with other parameters, S_o is defined as the initial number of solute atoms in the melt, which, of course, is determined by the simple formula:

$$S_o = \frac{C_o W_o}{\rho}$$

Here, C_o is just the initial impurity concentration in the melt. For convenience, the integration variable W' is formally replaced with a new variable, *w*, defined as $W_o - W'$, hence:

$$\int_{S_o}^{S}\frac{dS'}{S'} = K\int_{W_o}^{W_o - W}\frac{dw}{w}$$

Clearly, the indicated integration is elementary, thus:

$$\ln S - \ln S_o = K(\ln(W_o - W) - \ln W_o)$$

Moreover, it follows immediately from the elementary properties of logarithms that:

$$\ln\left(\frac{S}{S_o}\right) = K\ln\left(\frac{W_o - W}{W_o}\right)$$

Obviously, one inverts the logarithm on each side to obtain:

$$\frac{S}{S_o} = \left(\frac{W_o - W}{W_o}\right)^K$$

This is more conveniently expressed in terms of initial melt weight and impurity concentration, which are generally known before crystal growth starts, hence:

$$S = \frac{C_o W_o}{\rho}\left(\frac{W_o - W}{W_o}\right)^K$$

Furthermore, *S* has been previously related to melt concentration, C_l:

$$\frac{C_l}{\rho}(W_o - W) = \frac{C_o W_o}{\rho}\left(\frac{W_o - W}{W_o}\right)^K$$

Therefore, it immediately follows that:

$$C_l = C_o\left(\frac{W_o - W}{W_o}\right)^{K-1}$$

Of course, it is concentration of dopant in the solid crystal that is really of interest, but this is trivially obtained from the distribution equilibrium:

$$C_s = KC_o\left(\frac{W_o - W}{W_o}\right)^{K-1}$$

Thus, the impurity concentration in the solid and the melt is determined at all stages of crystal growth under rapid stirring conditions. This expression may be recast as a function of ingot length as follows:

$$C_s = KC_o\left(1 - \frac{\rho Ax}{W_o}\right)^{K-1}$$

Physically, this equation describes variation of impurity concentration along the length of a CZ grown ingot under rapid stirring conditions. Clearly, at the "seed end", *i.e.*, $x = 0$, of the boule, impurity concentration just corresponds to the simple distribution equilibrium. However, as the growth process proceeds, impurity is rejected from the growing crystal, and the concentration of the melt increases. This causes the concentration of impurity to increase in the boule as a function of distance from the seed end. This is shown in the following figure for several values of K (the individual curves are labeled by the corresponding value of K):

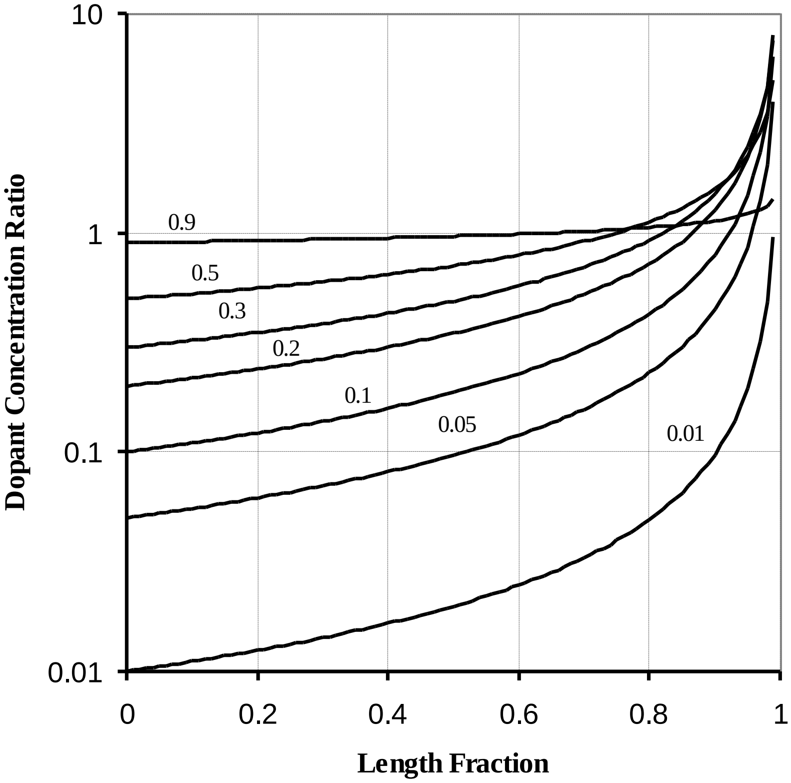


Fig. 15: Doping profile of a CZ crystal assuming rapid stirring

By definition, the seed end of the boule corresponds to a length fraction of zero. In contrast, the "butt end" is opposite of the seed end (*i.e.*, a length fraction nearly unity) and contains the last material solidified. Of course, dopant concentration ratio as indicated in the figure is just C_s/C_o. Clearly, the closer that K is to unity, the more uniform the doping profile obtained.

As might be expected, for realistic values of mixing, growth, and transport rates, the rejection rate of impurity at the growing crystal interface exceeds the rate at which impurities can be transported back into the bulk of the melt. Hence, the rapid stirring condition is not satisfied and breaks down. Consequently, in the vicinity of the growth interface impurity concentration builds up above the value observed in the bulk melt. Accordingly, this causes the crystal to be doped more heavily than would be expected under rapid stirring conditions. This behavior can be understood by assuming that a stagnant boundary layer of thickness, δ, exists between the solid/melt interface and the bulk of the melt. Therefore, if *D* represents the impurity diffusion coefficient within the melt and if *R* is the "instantaneous" pull rate, *i.e.*, *d*Δ*x*/*dt*, then assuming nearly steady-state conditions, the impurity concentration in the stagnant region near the growth interface may be described by the expression:

$$D\frac{d^2C}{d\Delta x^2} + R\frac{dC}{d\Delta x} = 0$$

Here, *C* is local solute concentration within the melt and Δ*x* is distance from the growth interface. Clearly, *C* must converge to the uniform concentration, C_l, if Δ*x* becomes large. Physically, the first term in the preceding expression accounts for diffusion of impurity atoms away from the growth interface back into the melt, and is a direct consequence of Fick's Second Law. Likewise, the second term accounts for removal of impurity from the melt due to freezing. Mathematically, this term has the appearance of a "net drift" due to some external "potential", but instead physically describes net motion of material due to pulling the crystal. Indeed, there must indeed be a net motion of impurity atoms with respect to the growth interface simply due to solidification, *i.e.*, in principle, all impurity atoms eventually must "pass through" the growth interface. This amounts to an overall relative motion of impurity atoms that can be regarded equivalently as being due to movement of the growth interface with respect to a stationary melt, or to movement of the melt with respect to a stationary growth interface, or as some combination of the two. (Terms describing convective flow of the melt are absent since the boundary layer is assumed to be stagnant.) Naturally, under conditions of nearly steady state, diffusion and freezing terms must effectively balance each other. For simplicity, *C*′ is identified with the derivative, *dC/d*Δ*x*, thus, the preceding equation becomes:

$$\frac{dC'}{d\Delta x} = -\frac{R}{D}C'$$

Clearly, this expression can be trivially integrated; however, instead of a definite integral it is convenient to construct an indefinite form as follows:

$$C' = (C'_0 - C'_\infty)e^{-R\Delta x/D} + C'_\infty$$

Here, C'_0 and C'_∞ are defined, respectively, as solute concentration gradients at the growth interface ($\Delta x = 0$) and far away from the growth interface ($\Delta x \to \infty$). Naturally, numerical values of C'_0 and C'_∞ correspond to imposition of suitable boundary conditions. Accordingly, as asserted previously, far away from the growth interface concentration of dopant impurity can be expected to be uniform, which implies that C'_∞ vanishes. Thus, the preceding formula can be further simplified:

$$C' = C_0' e^{-R\Delta x/D}$$

To determine C_0' diffusion and rejection fluxes of impurity atoms at the growth interface are assumed to be equal. Of course, this is consistent with the original differential equation describing the effects of diffusion and freezing in the stagnant boundary layer, and implies that the solute concentration profile across the growth interface and boundary layer remains essentially in a steady state (or at least only very slowly varying one). Thus, for a unit area of growth interface, one can write:

$$-DC_0' = R(\overline{C}_l - C_s)$$

Clearly, the term on the left just comes from Fick's First Law, which relates diffusion flux linearly to concentration gradient. Obviously, the number of impurity atoms per unit volume rejected back into the melt must just be the difference of impurity concentrations in the melt and in the solid exactly at the growth interface. (By definition, $\overline{C}_l$ is the impurity concentration exactly at the growth interface.) Therefore, the term on the right is the rejection flux and is just the product of the interfacial concentration difference and the instantaneous pull rate, *R*; hence, it follows that:

$$C' = \frac{dC}{d\Delta x} = -\frac{R}{D}(\overline{C}_l - C_s)e^{-R\Delta x/D}$$

It is a simple matter to integrate this expression across the entire boundary layer to obtain the expression:

$$C_l - \overline{C}_l = -\frac{R}{D}(\overline{C}_l - C_s)\int_0^\delta d\Delta x \exp\left(-\frac{R}{D}\Delta x\right) = (\overline{C}_l - C_s)\left(\exp\left(-\frac{R}{D}\delta\right) - 1\right)$$

Here, the solute concentration at the boundary layer edge, *i.e.*, $\Delta x = \delta$, has been assumed to be the uniform bulk concentration, C_l. Of course, this condition is not strictly realized unless $\Delta x \to \infty$; however, it is reasonable to assume that δ is sufficiently large so that the concentration for $\Delta x = \delta$ is only negligibly different from C_l. This expression is easily rearranged into a more convenient form:

$$\frac{C_l - C_s}{\overline{C}_l - C_s} = e^{-R\delta/D}$$

At this point, one redefines the ratio, C_s / C_l, as an effective segregation coefficient, K_e, since an enriched boundary layer lies between the bulk and the growth interface. (Of course, the ratio, $C_s / \overline{C}_l$, must equal the actual segregation coefficient, *K*, since $\overline{C}_l$ is defined as the concentration at the actual growth interface.) Therefore, the preceding expression becomes:

$$\frac{\frac{1}{K_e}-1}{\frac{1}{K}-1}=e^{-R\delta/D}$$

Obviously, one solves for K_e as follows:

$$\frac{1}{K_e}=\left(\frac{1}{K}-1\right)e^{-R\delta/D}+1$$

Upon formally taking the reciprocal the desired expression is obtained, thus:

$$K_e=\frac{K}{K+(1-K)e^{-R\delta/D}}$$

This result is applied to an actual CZ growth process by just substituting K_e into expressions obtained previously for rapid stirring conditions. The parameter, $R\delta/D$, is called *normalized growth parameter* and in practice is determined empirically for a given crystal growing apparatus. Clearly, if the normalized growth parameter is made sufficiently large (for example, by increasing the pull rate), then K_e tends toward unity, and this will result in a more uniform distribution of dopant along the length of the boule. Conversely, if the pull rate and, therefore, the normalized growth parameter becomes small, then K_e tends toward K. Obviously, this just corresponds to a return to rapid stirring conditions.

Both D and R can be adjusted by a judicious choice of process conditions, *e.g.*, temperature and pull rate. What about δ? It turns out that δ is a function of the rotation rate of the boule for which an empirical relationship has been determined experimentally:

$$\delta=1.8D^{1/3}\nu^{1/6}\omega^{-1/2}$$

Here, ν is the viscosity of the melt and ω is rotation rate. The pull rate, R, is closely related to the actual growth rate of the crystal, however, the instantaneous growth rate may differ from R because of thermal fluctuations, supercooling, *etc.* Indeed, if the pull rate is relatively small, the instantaneous growth rate may become negative (this is called *re-melting*). This can adversely affect both defect structure and doping distribution on a microscopic scale. In particular, if re-melting is not strongly suppressed by the use of a sufficiently large pull rate, the crystal may exhibit defect "swirl patterns" and dopant striations. Moreover, convective transport within the melt may also redistribute impurities non-uniformly. This is especially significant in the case of oxygen which is dissolved from the quartz crucible at the melt periphery. Indeed, because of these kinds of variations it is difficult to produce large diameter, lightly doped CZ wafers, *i.e.*, with a resistivity exceeding 100 Ω cm. Recently, immersion of the crucible during crystal growth in a strong magnetic field (the *magnetic Czochralski* or MCZ process) has been found to allow control of convective transport, which improves uniformity of large diameter crystals (>300 mm). As wafer sizes increase this may be expected to become an industry standard. Additionally, if the pull rate is low, the solidified crystal may be held for quite a long time above 950°C. This may allow sufficient time for thermally generated "microdefects" to form. The formation of such microdefects is effectively quenched if the pull rate exceeds a rate of roughly 2 mm/min.

Clearly, the normalized growth parameter and, hence, the effective segregation coefficient can be modified by changing rotation and pull rates during crystal growth. Indeed, it is common commercial practice to program growth parameters so as to obtain a uniform impurity concentration over a large fraction of an ingot. Accordingly, crystal growth proceeds in distinct phases: First, during an initial growth phase, as asserted previously, crystal diameter is built up to the desired dimension. Next, programmed pull and rotation rates are applied. This results in a crystal of constant impurity concentration over a large fraction of its length. Of course, at some point the melt is substantially exhausted and it becomes impossible to sustain a uniform composition. This defines a third growth phase during which the crystal is rapidly pulled out of the melt.

Zone Refining

As noted at the outset, in CZ growth there can be considerable carbon and oxygen contamination that comes from the quartz and graphite components of the process equipment. In general, this contamination causes no problem and perhaps may be beneficial since, as observed previously, ultrapure silicon is actually mechanically more fragile than ordinary CZ silicon and, also, as will be made evident subsequently, oxygen contamination may be used to good effect to set up an internal gettering scheme. However, there are some cases for which ultrahigh purity is desired, *viz.*, 1-10 KΩ cm). This material is most conveniently fabricated by zone refining. Again, to reiterate, within the industry substrates fabricated this way are called float zone (FZ) wafers.

To understand zone refining, suppose that just as in CZ growth, some impurity dissolved in molten silicon is in equilibrium with solid crystal. Once again, the distribution (or segregation) coefficient, *K*, is defined:

$$K = \frac{C_s}{C_l}$$

Of course, C_s is the impurity concentration within the solid and C_l is the impurity concentration in the liquid. Moreover, again, just as for Czochralski growth it is possible to maintain the system in a steady state, but not in rigorous thermal equilibrium, such that an effective distribution coefficient, K_e, is applicable instead of absolute distribution coefficient, *K*.

In practice, the basic technique of zone refining is to pass a solid piece of material, *e.g.*, an ingot of silicon, through a circular heating element, *viz.*, "needles eye". This creates a molten zone that slowly moves from one end of the ingot to the other. (Of course, migration of the molten zone along the ingot can be accomplished either by moving the ingot through a fixed heating element or by moving the heating element holding the ingot in a fixed position.) In any case, as the molten zone migrates and as a consequence of the distribution equilibrium, impurities are collected in the molten material and "swept" preferentially to one end of the ingot. Consequently, the rate that the molten zone passes through the ingot is analogous to the pull rate in the Czochralski process. Thus, the effective distribution coefficient should be essentially determined by the zone migration rate. Within this context, one expects that K_e should be significantly larger than K (or even, perhaps, approach unity) if the molten zone is moved very rapidly through the ingot and, in contrast, should approach or essentially equal K if the molten zone is moved very slowly. For clarity, it is instructive to consider a single pass zone refining process, which may be represented pictorially as follows:

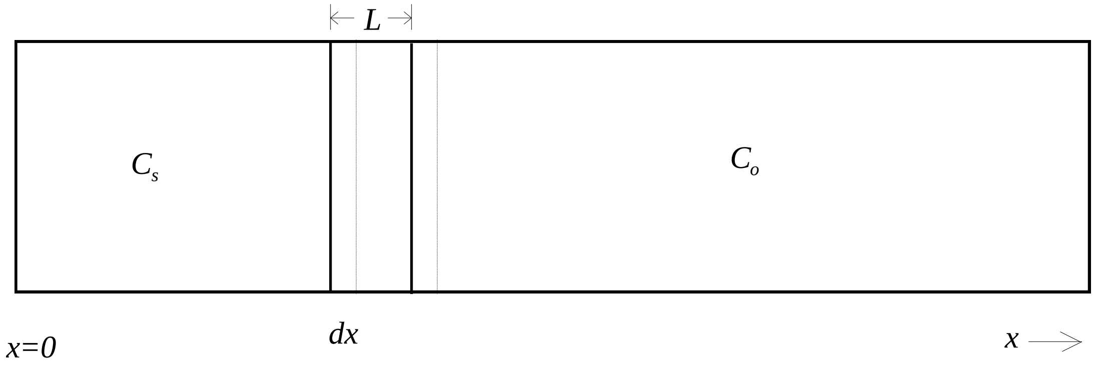


Fig. 16: Schematic diagram of the Float Zone (FZ) process

Here, L is the length of the molten zone, C_s is redefined as the impurity concentration in the refined section of the ingot, and C_o is the impurity concentration in the unrefined section. The variable, x, represents linear distance along the ingot (with the zone refining process initiated conventionally at $x=0$). If S is the number of impurity atoms in the molten zone and A is the cross sectional area of the ingot, then the differential change in S as the molten zone passes through the ingot is given by:

$$dS = C_o A dx - C_s A dx$$

This equation is just a formal expression of the difference in the rate that impurity atoms are added to the molten zone due to melting of unrefined material to the rate that they are lost to the melt at the freezing interface. Clearly, the impurity concentration in the liquid, C_l, is just S/AL. Hence, if one assumes that the distribution of impurity at the freezing interface is governed by the effective distribution coefficient, K_e, then it follows that:

$$\frac{dS}{dx} = C_o A - K_e \frac{S}{L}$$

(Obviously, the term, C_oA, is proportional to the rate that impurity atoms enter the liquid at the melting interface and, likewise, K_eS/L is proportional to the rate that impurity atoms are removed from the melt at the freezing interface.) If one assumes that the initial impurity concentration, C_o, is uniform throughout the unrefined section of the ingot, then, this differential equation is easily integrated by means of an exponential integrating factor. Accordingly, if one defines a new solute parameter, Q, as $S\exp(K_e/L)$, then one obtains:

$$\frac{dQ}{dx} = C_o A e^{K_e x/L}$$

Therefore, it follows trivially that:

$$Q(x) - Q(0) = \frac{C_o AL}{K_e}\left(e^{K_e x/L} - 1\right)$$

Recasting this expression in terms of S yields the result:

$$S(x)e^{K_e x/L} - S(0) = \frac{C_o AL}{K_e}\left(e^{K_e x/L} - 1\right)$$

The boundary condition for $S(0)$ must just be $C_o AL$ since melting an unrefined portion of the ingot of volume AL forms the initial molten zone. Upon substitution, one obtains:

$$S(x) = \frac{C_o AL}{K_e}\left(1 - (1 - K_e)e^{-K_e x/L}\right)$$

If one uses the definition of C_l and the distribution equilibrium, it follows that:

$$C_s(x) = C_o\left(1 - (1 - K_e)e^{-K_e x/L}\right)$$

This expression describes the concentration of initially uniformly distributed impurity after single pass zone refining.

Further purification can be achieved by additional zone refining. The equation describing the process is just the same as before, however, the initial impurity concentration is no longer uniform; hence, the original differential equation must be modified as follows:

$$\frac{dS}{dx} = C_o(x)A - K_e\frac{S}{L}$$

Here, $C_o(x)$ is an arbitrary (*i.e.*, non-uniform) initial impurity distribution. Again, this expression is recast in terms of Q:

$$\frac{dS}{dx} = \frac{dQ}{dx}e^{-K_e x/L} - K_e\frac{Q}{L}e^{-K_e x/L} = C_o(x)A - K_e\frac{Q}{L}e^{-K_e x/L}$$

Therefore, one obtains:

$$\frac{dQ}{dx} = C_o(x)Ae^{K_e x / L}$$

Naturally, this differential equation is easily integrated formally to give:

$$Q(x) - Q(0) = A\int_0^x dx' C_o(x')e^{K_e x' / L}$$

Following the result for constant C_o, one identifies $Q(0)$ as $C_o(0)AL$ and recasts this expression in terms of S:

$$S = Ae^{-K_e x / L}\int_0^x dx' C_o(x')e^{K_e x' / L} + C_o(0)ALe^{-K_e x / L}$$

It immediately follows from the distribution equilibrium at the freezing interface that the solute concentration in the refined ingot is given by the expression:

$$C_s(x) = \frac{K_e}{L}e^{-K_e x / L}\int_0^x dx' C_o(x')e^{K_e x' / L} + K_e C_o(0)e^{-K_e x / L}$$

Of course, to determine $C_s(x)$ numerically, prior knowledge of $C_o(x)$ is required. Naturally, $C_o(x)$ may be the result of a previous zone refining pass or may be a "grown-in" distribution due to a particular set of process parameters for a CZ growth process.

Within this context, if one substitutes the concentration profile obtained previously for single pass zone refining of an initially uniform ingot into the preceding expression, one obtains:

$$C_s(x) = \frac{K_e C_o}{L}e^{-K_e x / L}\int_0^x dx'\left(1-(1-K_e)e^{-K_e x' / L}\right)e^{K_e x' / L} + K_e^2 C_o e^{-K_e x / L}$$

Here, $C_s(x)$ is evidently the concentration profile of the ingot following two zone refining passes. Clearly, the integrals are all of elementary form and can be constructed explicitly to give the result:

$$C_s(x) = C_o\left(1-\left(1-K_e^2+\frac{K_e}{L}(1-K_e)x\right)e^{-K_e x / L}\right)$$

Of course, C_o retains the usual definition as initial uniform impurity concentration in the ingot. Naturally, one can compare this expression for two passes to the single pass result and also to the initial uniform impurity concentration. Clearly, at the starting, *i.e.*, seed, end of the ingot, the impurity concentration was, of course, just C_o prior to zone refining. After a single pass, this is reduced to K_eC_o. After two passes this is further reduced to $K_e^2C_o$. Clearly, since K_e is smaller than one, this shows that the seed end of the ingot is progressively purified. Concentration profiles for one and two zone refining passes with various effective distribution coefficients are shown in the following figure:

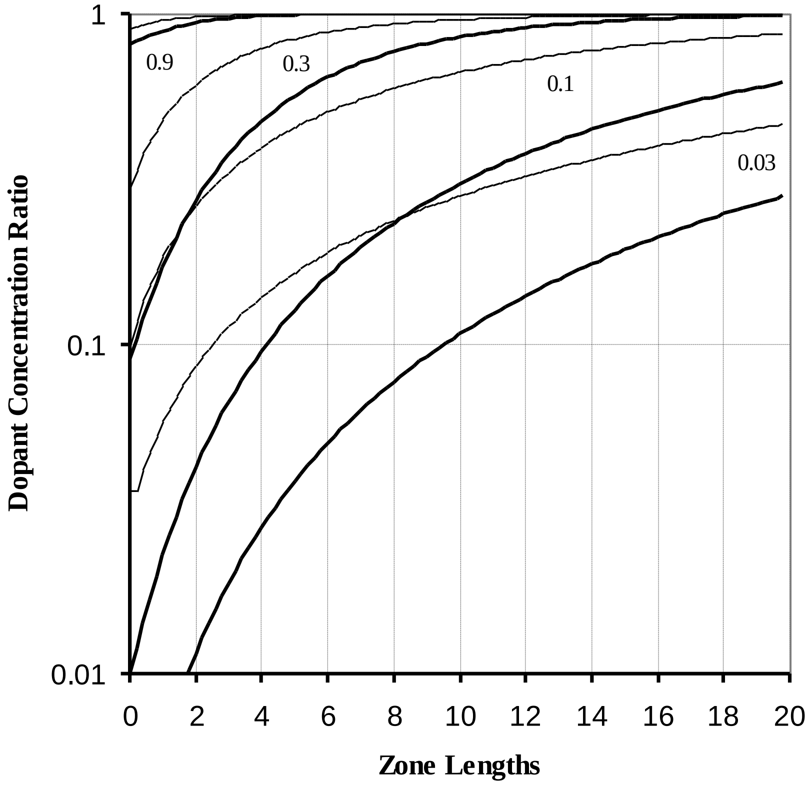


Fig. 17: Doping profile of an FZ crystal following one (*narrow line)* and two (wide line) zone refining passes

Of course, the dopant concentration ratio is, again, just C_s/C_o. The horizontal axis is the number of zone lengths refined. (Hence, by definition, the seed end of the ingot corresponds to zero zone lengths.) One observes that in contrast to the Czochralski process for which an effective distribution coefficient near unity is desired, in zone refining better results are obtained the smaller the value of K_e. (Of course, K_e can never become smaller than K itself.) Clearly, this implies that a slow rate of migration of the molten zone through the ingot is desirable.

Physically, zone refining literally sweeps impurity atoms toward one end of the ingot, *i.e.*, the end toward which the molten zone moves. Since, no impurity is physically removed from the ingot this end actually becomes more impure. However, zone refining process parameters can be set up in such a way that the impure end is a reasonably small part of the total volume of the ingot. Clearly, if zone refining is repeated many times, the bulk of an ingot can be refined to ultrahigh purity and the impure end can simply be removed resulting in a large amount of highly refined material. Of particular importance is the production of wafers with low oxygen content since these cannot be produced using the CZ process. Of course, zone refining does not leave impurity uniformly distributed. If this is a requirement, then once the impure end has been removed, the ingot can be heat treated to redistribute impurity more uniformly. To reiterate, the main advantage of the float-zone process is the very low impurity concentration in the silicon crystal. In particular the oxygen and carbon concentrations are much lower as compared to the CZ process, since the melt does not come into contact with a quartz crucible, and no hot graphite container is used. As a practical matter, FZ ingots are produced from an initial polycrystalline silicon ingot, which is seeded with a crystal of desired orientation at the start of the process. Alternatively, a monocrystalline ingot can be further purified by zone refining. Even so, the FZ process is more expensive than then CZ process, and, at present, crystal diameter is limited to 200 mm.

Intrinsic Defects in Semiconductors

In all previous consideration of crystal structure and crystal growth, for simplicity it has been assumed that the silicon crystal lattice is entirely free of defects. Of course, in reality, this cannot be true since at any temperature greater than absolute zero, no crystal of finite size can be absolutely perfect. Indeed, there are a number of different types of defects that can exist within the crystal lattice of any pure material. In general, such *intrinsic lattice defects* can be broadly classified in terms of dimensionality, *viz.*, *point*, *line*, *plane*, and *spatial* or *volume defects*. Moreover, any foreign species present within the crystal lattice may obviously also be regarded as a kind of defect. As a matter of semantic terminology, such impurities are to be regarded as *extrinsic lattice defects*; however, as will become evident subsequently, these may actually initiate the appearance of intrinsic defects. In any case, it is useful to limit discussion (at least temporarily) to the various types of intrinsic defects, *i.e.*, defects that do not require the presence of foreign atoms.

Point Defects

Naturally, point defects are the simplest kinds of defects that can exist within a crystal lattice of which the most elementary example is a *vacancy* (also called a *Schottky defect*). As a conceptual matter, a vacancy can be regarded as the result of a hypothetical process in which an atom is removed from a *lattice site* within the bulk of the crystal and transferred to the crystal surface. (Within this context, the term "lattice site" refers to an actual atomic site within the crystal and, therefore, is not the generally the same as a lattice point associated with the corresponding Bravais lattice.) As might be expected, formation of a vacancy requires a net energy input into the silicon lattice, which is approximately 2.3 eV. Physically, this energy corresponds to breaking bonds within the bulk and reforming bonds on the surface. In addition, a small portion is associated with reorganization or restructuring of the lattice. That the energy is positive is to be expected since binding energy of an atom within the bulk is greater in magnitude (*i.e.*, more negative) than that of an atom on the surface. Of course, 2.3 eV is large compared to the mean thermal energy at room temperature; hence, at 300°K normal thermal fluctuations can produce only a very small concentration of vacancies within an ordinary silicon crystal.

A second type of intrinsic point defect is a *self-interstitial*. This kind of defect can be thought of as the "inverse" of a vacancy for which an atom of the crystal is hypothetically transferred from the surface into the interior. However, since no unoccupied lattice site is generally available at some arbitrary location within the crystal lattice, the excess atom must "squeeze" into an interstitial site existing within the lattice. Typically, within a close packed solid, interstitial sites are small and formation of an interstitial defect requires an even larger energy input than formation of a vacancy. Obviously, this implies that self-interstitial defects (or just interstitials) should be very rare within such materials (as is, indeed, the case). In contrast, silicon is characterized by the relatively open diamond cubic structure, for which there are five interstitial sites per unit cell. Moreover, these are reasonably large; hence, in silicon, formation energy of a self-interstitial is commensurate to that of a vacancy. Formation of both a vacancy and self-interstitial is illustrated pictorially as follows:

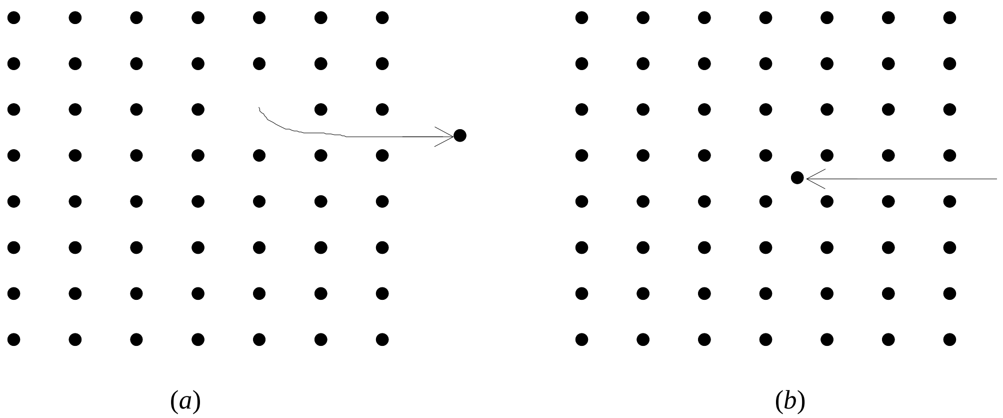


Fig. 18: Intrinsic point defect formation: (*a*) vacancy; (*b*) interstitial

For clarity, the locations of interstitial sites within the diamond cubic structure are shown in the following figure:

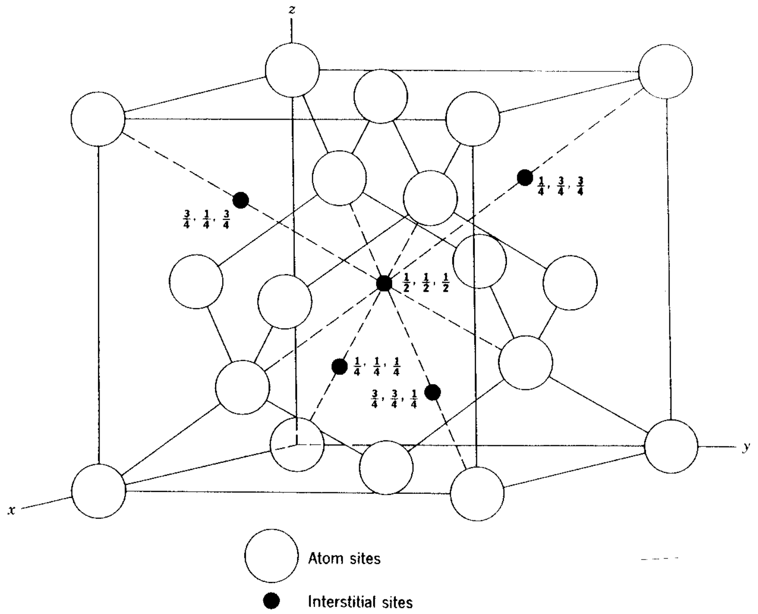


Fig. 19: Interstitial sites in the diamond cubic structure

Naturally, a vacancy and interstitial can form simultaneously if an atom is displaced from a lattice site into a nearby interstitial site. Since, the newly formed interstitial and vacancy are in close proximity, the strain introduced into the lattice is less than in the case of isolated vacancies and interstitials. Hence, the formation energy of a vacancy-interstitial pair is reduced in comparison to the total formation energy required to produce an isolated vacancy and isolated interstitial. Accordingly, an associated vacancy-interstitial pair is regarded as a particular type of defect and; hence, is called a *Frenkel defect.*

In passing, instructive analogies can be made with other dynamic equilibria. In particular, vacancies and interstitials can be considered in a generalized sense as "solutes" in a "solution" for which the solid silicon lattice is regarded as "solvent". Furthermore, once formed, in analogy to ordinary solutes in aqueous solutions, vacancies and interstitials do not remain stationary, but, can migrate, *i.e.*, diffuse, due to random thermal motion within the "solvent medium", *i.e.*, the crystal lattice. In addition, thermal generation of vacancies and interstitials, *i.e.*, Frenkel defects, can be expected to be governed by a mass action equilibrium analogous to the mobile carrier equilibrium or the autodissociation equilibrium of water (which defines the well-known pH scale). In this sense, vacancies and interstitials have a relationship to an undefected silicon crystal that is analogous to the relationship of hydroxide and hydronium ions to pure water. Therefore, one expects that vacancies and interstitials must satisfy an equilibrium expression of the form:

$$[V][I] = K_{eq}$$

Here, the left hand side is the product of vacancy and interstitial concentrations (denoted as $[V]$ and $[I]$, respectively). Of course, K_{eq} is a thermodynamic equilibrium constant and, as such, depends on temperature. This process can be represented schematically as a kind of "chemical" equilibrium:

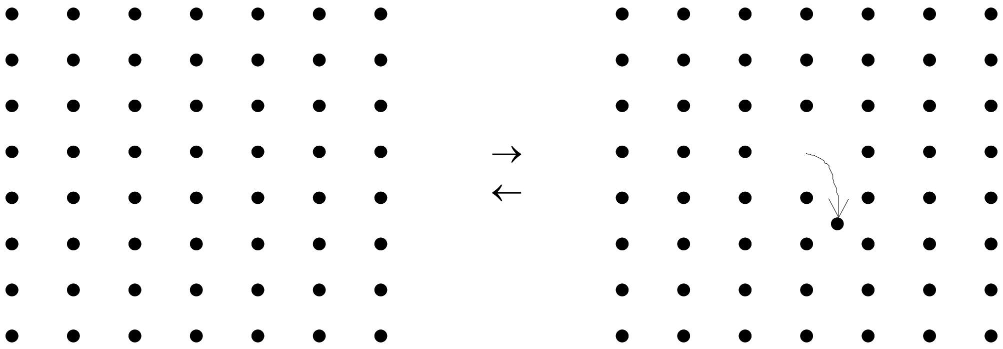


Fig. 20: Vacancy-interstitial "chemical" equilibrium

Clearly, the figure on the left represents an undefected lattice. The figure on the right represents "dissociation" of the lattice into a vacancy and an interstitial. It is further worthwhile to observe that vacancies and interstitials strongly interact and can also recombine, resulting in a significant release of energy. (Physically, one can regard a vacancy and an interstitial as exerting an attractive force toward each other.)

Of course, various other combinations of point defects can also occur. For example, the formation of a single vacancy requires the breakage of four crystal bonds, but the formation of a *di-vacancy* requires the breakage of only six bonds and not eight. Consequently, the formation of a di-vacancy requires less energy per defect than the formation of an isolated vacancy. (This is similar to the formation of a Frenkel defect.) It turns out that di-vacancies are commonly present within a silicon lattice. Conversely, *di-interstitials, i.e.*, atoms in adjacent interstitial sites, are formed with difficulty (if at all) since this requires introduction of a large amount of strain energy into the silicon lattice.

Within the present context, a di-vacancy could be viewed as a "bound state" of two vacancies. Thus, similar to a vacancy-interstitial pair, self-interaction of two vacancies can be considered to be the result of an attractive force (though weaker than the attraction between vacancies and interstitials). Conversely, consistent with the high lattice strain energy associated with the formation of di-interstitial defects, one expects interstitials to exert a mutually repulsive force.

Line Defects

A line defect is called a *dislocation*. In general, two ideal types of dislocations exist, *viz.*, *edge* and *screw*. Ideal edge and screw dislocations are illustrated by the following figure:

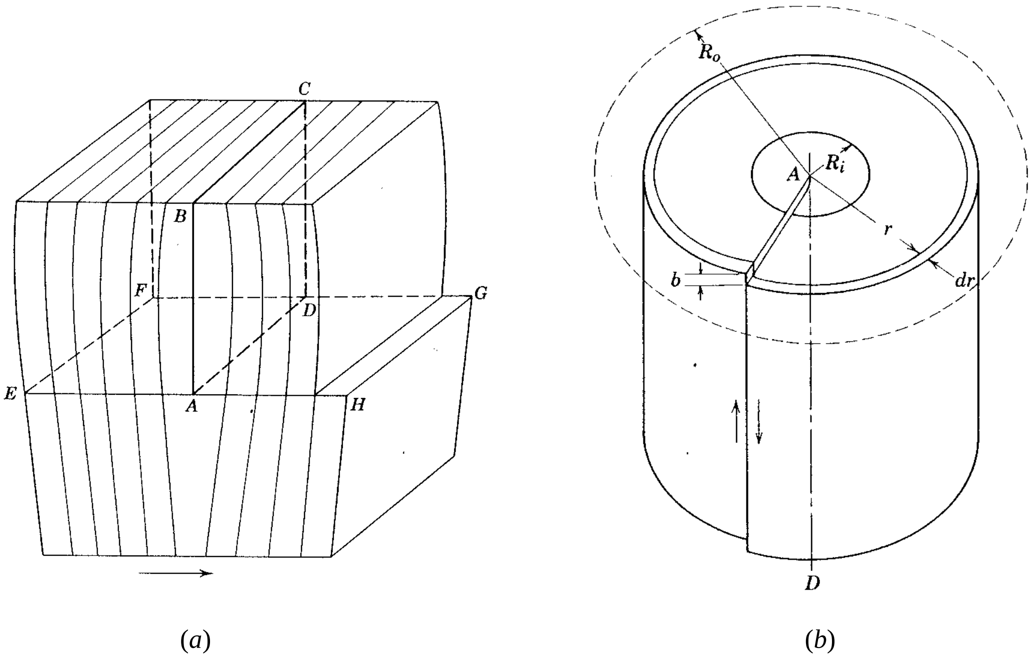


Fig. 21: Ideal (*a*) edge dislocation; (*b*) screw dislocation

Of these two types, edge dislocations are the easiest to visualize and, conceptually, an ideal edge dislocation can be considered as the result of hypothetical insertion of an extraneous "half-plane" of atoms along one of the crystallographic directions into an otherwise non-defective crystal lattice. By definition, the edge of the inserted half-plane corresponds to the *dislocation line* or *axis*. (In the (*a*) figure above, the dislocation axis associated with an edge dislocation lies along line segment *AD*.) Clearly, the crystal lattice must be disrupted in close proximity to the dislocation; however, far from the dislocation the crystal lattice remains relatively "perfect". Obviously, there must also be a localized increase in strain energy corresponding to the existence of a dislocation within the crystal lattice. Concomitantly, an ideal screw dislocation is more difficult to visualize, but is formed if the crystal is "sheared" parallel to the axis of the dislocation. For a pure screw dislocation, no extraneous plane of atoms is required; however atomic planes within the lattice are displaced into an arrangement resembling a “spiral staircase”. Again, the corresponding disruption of the crystal structure in proximity to the dislocation axis results in a local increase in crystal strain energy. (In the (*b*) figure, the line segment *AD* again coincides with the dislocation axis of an ideal screw dislocation.) Of course, the presence of defects of any dimensionality within an otherwise perfect crystal lattice can be expected to be associated with a localized increase in potential, *i.e.*, strain, energy. This is easily understood if one recalls that binding energy is maximized within a perfect, undefected lattice. Since potential energy of any bound state is by definition, a negative quantity, any localized weakening of crystal bonding must correspond to a localized increase in potential energy. Therefore, it is obvious that any disturbance in crystal bonding can be expected to be fundamentally associated with the presence of defects. Therefore, it is not surprising that under the influence of an externally applied stress, dislocations of both types can move with relative ease along a corresponding *slip plane*. (In the case of point defects, such motion corresponds to stress enhanced diffusion.) Generally, the dislocation axis lies in the slip plane. (Obviously, the planar section *EFGH* in the preceding (*a*) figure coincides with a slip plane.) Moreover, although the dislocation moves, the atoms themselves do not move significantly. Indeed, all that is necessary for a dislocation to move is a localized rearrangement in crystal bonding.

Line defects, *i.e.*, dislocations, can also interact with point defects naturally present within the lattice. This is most easily understood for the simple case of a pure edge dislocation. Suppose that due to the random thermal motion of the lattice, a vacancy migrates to the dislocation axis. Clearly, this is equivalent to the removal of an atom from the edge of the extraneous half-plane that defines the dislocation axis. This process is called *vacancy capture*. Of course, in this case the atom lost from the half-plane ends up in a nearby lattice site. Conversely, suppose that an atom migrates away from the dislocation axis to form an interstitial. Again, this is equivalent to removal of an atom from the extraneous half plane, but in this case, an interstitial defect has been formed. Not surprisingly, this process is called *interstitial generation*. Accordingly, both processes, *i.e.*, interstitial generation or vacancy capture, cause the dislocation axis to “climb” out of its associated slip plane. Clearly, these processes are essentially identical to formation of vacancy-interstitial pairs (Frenkel defects) or recombination of a vacancy and an interstitial; however, here they occur in close proximity to a dislocation. Obviously, one expects that rates and, perhaps other characteristics of these processes will be substantially affected by localized strain energy associated with the dislocation.

Dislocations can be characterized quantitatively in terms of the *Burgers vector*, **b**. To define **b**, one first considers a path taken within the crystal that, by definition, would return exactly to its starting point if no line defects are present, *viz.*, *Burgers circuit*. Clearly, the path is closed and its length corresponds to some integral number of lattice parameters. If instead of a perfect crystal, the Burgers circuit encloses the axis of a dislocation, it is no longer closed and the starting and ending points are now separated by a small displacement. This displacement determines the Burgers vector. (In the diagram of a pure screw dislocation appearing previously, the length of the Burgers vector is represented by the parameter, *b*.) Edge and screw dislocations are characterized by the orientation of the Burgers vector with respect to the dislocation axis. Clearly, for a pure screw dislocation, **b** is parallel to the dislocation axis. In contrast for a pure edge dislocation **b** is perpendicular to the dislocation axis. Of course, in real crystals the situation is rarely ideal and dislocations occur in loops and tangles and, thus, are generally not perfectly straight lines. Accordingly, they are neither purely edge or screw, but are of “mixed” character. Indeed, if one “follows” a dislocation through a crystal, it can appear as edge or screw or some intermediate mixture at different locations. Therefore, for a real dislocation in a real crystal, **b** and its orientation with respect to the dislocation axis varies from point to point.

Plane Defects

As asserted within the context of line defects, within a crystalline solid dislocations generally form closed *dislocation loops*. (Clearly, an unclosed dislocation must terminate somewhere on the surface of the crystal.) Of course, by definition, an edge dislocation corresponds to the edge of a partial atomic plane. Thus, if an edge dislocation forms a closed loop, there must be a corresponding partial atomic plane either present or absent within the crystal lattice. In this way a pure edge dislocation loop defines the boundary of an ideal planar defect called a *stacking fault*. If the dislocation loop corresponds to the absence of part of an atomic plane, the corresponding stacking fault is said to be *intrinsic*. Conversely, an *extrinsic* stacking fault is formed if a partial atomic plane is inserted into the crystal. Intrinsic and extrinsic stacking faults are illustrated pictorially by the following figures:

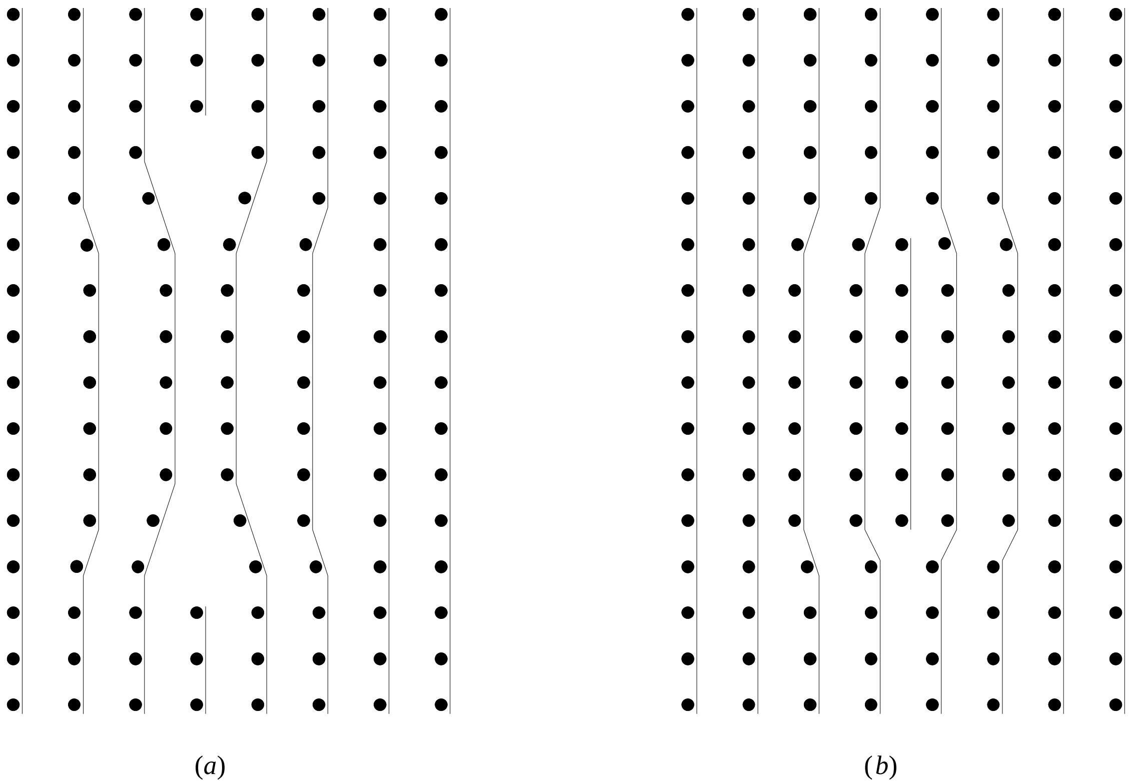


Fig. 22: Ideal (*a*) intrinsic stacking fault; (*b*) extrinsic stacking fault

Naturally, just as for dislocations, real stacking faults can be expected to be more complicated than just these idealized types.

A third kind of ideal planar defect is a *twin* or *growth fault*. This fault occurs if the stacking order of crystalline layers is inverted symmetrically with respect to some plane within the crystal. Thus, a twin fault is not bounded by a dislocation loop. In particular, for a diamond cubic lattice, twins are formed by reversal of the atomic stacking order about a [111] plane. Moreover, one observes that if a twin fault extends through the entire body of an otherwise perfect crystalline solid, it is more natural to regard the whole solid as consisting of two separate perfect crystals joined at the twin plane. Indeed, if twin faults are present to any great degree, one expects the regularity of the overall "crystal" to be severely disrupted. In this case, such material is more properly regarded as a *polycrystalline solid* with individual crystalline regions separated by disordered regions called *grain boundaries*. In practice, twin faults should never be present in substrates used for semiconductor device fabrication.

Spatial Defects

Obviously, spatial (or volume or bulk) defects can be formed by concentration of defects of lower dimensionality. For example, vacancies can coalesce to form bulk voids. Growth and stacking faults can concentrate to form grain boundaries. (In this case, the single crystal character of the lattice is disrupted and the material, thus, becomes polycrystalline.) Indeed, defect density may become so large that all crystal structure is effectively lost and the material is essentially *amorphous*, *i.e.*, without any long-range order. Indeed, from the point of view of processing, any significant quantity of dislocations, stacking faults, or bulk defects in electrically active surface layers of a wafer generally cause poor device performance and low yield. Accordingly, such defects are technologically unacceptable and, usually, are not present in the starting material, but may be caused by subsequent process conditions during device fabrication, *e.g.*, thermal shock, mechanical damage, *etc.* (Some of the causes of these kinds of defects will be considered subsequently in connection with ion implantation, *etc.*) In any case, spatial defects need not be considered further since they are catastrophic to device performance and must be rigorously eliminated from any practical fabrication process.

Thermodynamics of Intrinsic Point Defects

As asserted previously, formation of intrinsic point defects within a silicon lattice is generally caused by random thermal motion of the atoms within the lattice itself. At room temperature, thermal energy is small in comparison to the binding energy of the lattice, thus, very few defects are formed; however, this number is not zero, therefore, spontaneous defect generation can be described by thermodynamics. Moreover, before proceeding further, it is important to define some basic thermodynamic terms. In particular, there are four classical thermodynamic state functions. These are the potential energies, *E*, *internal energy*, and, *H*, *enthalpy*, and free energies, *A* and *G*, called *Helmholtz* and *Gibbs free energies*, respectively. As a matter of generality, *E* and *A* are applicable to thermodynamic systems for which volume is constant. Likewise, *H* and *G* are applicable to thermodynamic systems for which pressure is constant. For systems including only condensed phases, *e.g.*, crystalline solids, this distinction is irrelevant and *E* and *H* can be considered identical as also can *A* and *G*. Therefore, when considering the behavior of crystalline solids, one can refer to potential or internal energy and free energy without ambiguity. In addition to *E*, *H*, *G*, and *A*, two additional thermodynamic quantities are important. These are absolute or *thermodynamic temperature*, *T*, and *entropy*, *S*. Temperature is, of course, a familiar concept, however entropy is much less familiar. Within a broad context, entropy is a measure of disorder or randomness characteristic of a physical system. For example, entropy increases during melting of a solid material even though temperature remains constant.

How does one determine these quantities for a crystalline material? As might be expected, internal energy can be identified with the total binding energy of the crystal. However, the identity of free energy is not as obvious. By definition, free energy is an amount of energy associated with a thermodynamic system which is available to "do work", that is to say, to drive some physical process. Physically, the product of temperature and entropy, *TS*, relates internal energy and free energy. Specifically, *TS* must be subtracted from internal energy to obtain free energy.

$$A = E - TS$$

Thus, *TS* is identified as just that part of the internal energy which corresponds to random thermal motion and, therefore, is not externally available. Furthermore, before continuing with a specific discussion of point defects, it is important to note that for most thermodynamic systems, absolute values of thermodynamic functions are not available. However, changes in thermodynamic functions relative to some reference state will serve just as well. Therefore, instead of absolute values of *E*, *H*, *G*, *A*, and *S*, relative values denoted as ΔE, ΔH, ΔG, ΔA, and ΔS, are used, thus:

$$\Delta A = \Delta E - T\Delta S$$

This expression is readily applied to the generation of point defects within a silicon crystal. (For a solid, a convenient thermodynamic reference state is a defect free crystal.)

Beginning with consideration of vacancy generation, one defines *N* as the number of atomic lattice sites and *M* as the number of vacancies existing in some unit volume of the crystal. Clearly, N is easily determined by inspection of the diamond cubic crystal structure. Therefore, it is desirable to specify M as a function of N and T. Thus, if ΔE_v is the energy of formation of a single vacancy (approximately 2.3 eV), then, considering a unit volume of crystal, the free energy change for the formation of M vacancies corresponds to the expression:

$$\Delta A_{Mv} = M\Delta E_v - T\Delta S_{Mv}$$

Here, ΔA_{Mv} is the free energy of formation of M vacancies and ΔS_{Mv} is the associated entropy change. Physically, the entropy change can be formally separated into two parts, ΔS^C_{Mv}, "configurational" entropy and, ΔS^X_{Mv}, "excess" entropy. Configurational entropy arises from an increase in disorder associated with an introduction of M vacancies into a perfect crystal lattice. To determine configurational entropy, one recalls Boltzmann's famous relation that fundamentally defines entropy:

$$S = k \ln \mathrm{W}$$

Here, entropy, S, in an absolute sense, is related to the natural logarithm of the number of equivalent, but distinguishable microscopic arrangements, W, associated with a particular physical system. The constant of proportionality is Boltzmann's constant, k. (Indeed, it is Boltzmann's relation that provides the fundamental definition of k.) One observes from elementary probability theory that the number of possible distinguishable arrangements of M vacancies in N lattice sites, W_{Mv}, simply corresponds to the binomial coefficient:

$$\mathrm{W}_{Mv} = \frac{N!}{(N-M)!\,M!}$$

Clearly, configurational entropy for a perfect crystal, *i.e.*, a crystal with zero vacancies, vanishes since there is only one distinguishable arrangement, *i.e.*, the one with every lattice site occupied. Therefore, Boltzmann's relation and the preceding formula can be combined to determine the configurational entropy change, ΔS^C_{Mv}, as follows:

$$\Delta S^C_{Mv} = k \ln \mathrm{W}_{Mv} = k \ln\left(\frac{N!}{(N-M)!M!}\right) = k \ln N! - k \ln M! - k \ln(N-M)!$$

For simplicity, one can ignore excess entropy, ΔS^X_{Mv}, which may be thought of as caused by change in the number available vibrational states of the crystal due to introduction of M vacancies. (ΔS^X_{Mv} is generally small.) Thus, the free energy change is given by:

$$\Delta A_{Mv} = M\Delta E_v - kT \ln N! + kT \ln M! + kT \ln(N-M)!$$

This expression can be further modified using Stirling's approximation for large factorials:

$$\ln N! \cong N \ln N - N$$

Hence, it follows that:

$$\Delta A_{Mv} = M \Delta E_v - NkT \ln N + MkT \ln M + (N - M)kT \ln(N - M)$$

Thus, the free energy of formation of M vacancies is a function of temperature, energy of formation of a single vacancy, number of lattice sites, and number of vacancies.

Physically, for some definite temperature thermodynamic processes for which the free energy change is large and negative spontaneously occur. Conversely, those for which the free energy change is large and positive are non-spontaneous and do not occur, *i.e.*, the reverse process is spontaneous. If the free energy change exactly vanishes, *i.e.*, forward and reverse processes have the same tendency to occur, then the process is in a state of equilibrium. Clearly, as expressed above, ΔA_{Mv} corresponds to formation of M vacancies in a perfect crystal. The number of vacancies will be stable, *i.e.*, in equilibrium, if the free energy change is positive either for the formation of additional vacancies or the loss of existing vacancies. This means that addition of one more vacancy or removal of a vacancy does not change free energy. Mathematically, this implies that ΔA_{Mv} is at an extremum; hence, one considers the partial derivative of ΔA_{Mv} taken with respect to the number of vacancies, M:

$$\frac{\partial}{\partial M} \Delta A_{Mv} = \Delta E_v + kT \ln M - kT \ln(N - M)$$

Clearly, the condition of equilibrium requires that the value of ΔA_{Mv} is at a minimum with respect to M. Thus, the derivative appearing on the left hand side above must vanish; hence one finds that:

$$\ln\left(\frac{M}{N - M}\right) = -\frac{\Delta E_v}{kT}$$

Generally, M is small in comparison to N. Therefore, one may replace $N-M$ with N and construct the exponential to obtain a final result:

$$M = N \exp\left(-\frac{\Delta E_v}{kT}\right)$$

As desired, this formula expresses the functional relationships for the number (or density) of vacancies in terms of N and T. It has the form of a product of an exponential factor which contains the temperature dependence (*i.e.*, a "Boltzmann factor") and a "pre-exponential" factor which is characteristic of the material (in this case, it is N, the number or density of atomic lattice sites). For completeness, if the excess entropy term had been included as a "correction", the preceding formula would be simply modified as follows:

$$M = N\exp\left(-\frac{\Delta E_v - T\Delta S_{Mv}^X}{kT}\right)$$

It is commonly the case for thermally activated processes to be described by expressions of this form.

Silicon self-interstitial defects can be treated analogously. Thus, the free energy change for the formation of M interstitials is as follows:

$$\Delta A_{Mi} = M\Delta E_i - T\Delta S_{Mi}$$

Here, ΔE_i is the formation energy of an interstitial. Obviously, ΔA_{Mi} is the free energy of formation of M interstitials and ΔS_{Mi} is the associated entropy change. Again, the entropy change can be divided into configurational and excess parts. As expected, the configurational part is of the form:

$$\Delta S_{Mi}^C = k\ln \mathrm{W}_{Mi}$$

However, to evaluate the configurational entropy change, one must consider the number of interstitial spaces per unit volume, N', rather than the number of lattice sites. Of course, N and N' are easily related by noting that there are eight lattice sites in a diamond cubic unit cell, but only five interstitial sites, thus:

$$N' = \frac{5}{8}N$$

Within this context, one can immediately write:

$$\mathrm{W}_{Mv} = \frac{N'!}{(N'-M)!M!}$$

The analysis proceeds just as in the case of vacancies, hence:

$$M = N'\exp\left(-\frac{\Delta E_i}{kT}\right) = \frac{5}{8}N\exp\left(-\frac{\Delta E_i}{kT}\right)$$

Obviously, excess entropy can again be treated as a correction. Naturally, the concentration of Frenkel defects can also be obtained by a similar analysis. Of course, the formation energy, ΔE_f, must be appropriate for Frenkel defects and a slight modification must be made to the entropy term; however, the result is essentially the same as obtained previously in the case of vacancies with ΔE_f replacing ΔE_v.

Within this context, a vacancy-interstitial thermodynamic equilibrium constant, K_{eq}, can be constructed directly from the preceding results:

$$K_{eq} = \frac{5}{8} N^2 \exp\left(-\frac{\Delta E_v + \Delta E_i}{kT} \right)$$

The similarity between the vacancy-interstitial equilibrium and hole-electron equilibrium is evidently apparent. Clearly, the energy required to create an isolated vacancy and an isolated interstitial is just $\Delta E_v + \Delta E_i$. This is analogous to the band gap energy in the case of mobile carriers. Furthermore, the product of lattice site density and interstitial site density, $5N^2/8$, plays exactly the same role as the product of effective densities of states. As expected, K_{eq} is a function of temperature, but not defect concentrations.

To conclude consideration of point defects, one observes that the presence of a vacancy theoretically results in four unsatisfied bonds that normally bind an atom in the vacant lattice site to its immediate neighbors. These “dangling” bonds can be viewed as half-filled sp^3 orbitals which are able to accept (theoretically, at least) as many as four extra electrons from the normal valence band. In this case, the vacancy becomes negatively charged leaving behind holes in the valence band. Depending on the energy of these localized states relative to the band gap, a vacancy can act much like a dopant atom. It is also possible for vacancies to donate electrons to the conduction band if the atomic configuration allows some or all of the dangling sp^3 orbitals to overlap. Indeed, since various atomic rearrangements can occur to reduce the energy of the vacancy, the situation can become quite complicated. Suffice it to say that vacancies can become electrically active and act like acceptor, donor, or deep level states. Furthermore, interstitial defects can also become electrically active since they also locally disturb the overall symmetry of the crystal. Interstitials typically become positively charged and exhibit donor-like behavior. (This behavior will be discussed in more detail in connection with diffusion mechanisms.)

Foreign Impurities

So far, only intrinsic defects have been considered, which, by definition, are the only kind of defects that can exist in a pure silicon crystal. However, if one allows for the existence of foreign impurity atoms within the crystal, other kinds of defects become possible. Indeed, substitution of electrically active dopant impurity atoms into crystal lattice sites normally occupied by silicon can be thought of as a kind of crystal point defect. Of course, such defects are desirable since they can be used intentionally to modify the electrical characteristics of the silicon crystal in an advantageous way. However, shallow level dopants, can also occupy interstitial sites, in which case, dopant atoms are no longer electrically active and, therefore, not beneficial. (Indeed, it is important to reduce the interstitial concentration of dopants to be as small as possible.) Indeed, foreign atoms can occupy either lattice, *i.e.*, substitutional, or interstitial sites, which with the exception of substitution of shallow level dopant atoms, generally has the undesirable effect of introducing electronic states near the middle of the band gap. Furthermore, in addition to point defects associated with impurities, foreign atoms can also agglomerate to form bulk defects called *precipitates*. (Often precipitates will "decorate" other crystal defects such as dislocations or stacking faults.) If such precipitates become large, they can disrupt the background crystal structure. Again, this is generally catastrophic; however, in contrast for the case of oxygen, precipitates can actually be manipulated to beneficial effect by allowing an *internal* (or *intrinsic*) *gettering* scheme to be realized.

Effects of Oxygen and Carbon

As asserted previously, oxygen and carbon are normally occurring contaminants in silicon produced by the CZ method. Indeed, oxygen is introduced into the silicon by dissolution of the quartz crucible itself. Typical concentrations are in the range of 10^{17} to 10^{18} cm^{-3}. Furthermore, oxygen concentration can be enhanced by the presence of other impurities such as boron. Typically, about 95% of all oxygen atoms occupy interstitial sites and, therefore, are truly dissolved in the crystal. The remaining 5% exist as "complexes" such as SiO_4. Even so, interstitial oxygen is found to increase the yield strength of silicon. This increase can be as much as 25% greater than pure silicon and significantly improves the mechanical characteristics of wafers. Typically, the effect increases with oxygen concentration until the solid solubility limit is exceeded and oxygen precipitates are formed. An additional effect of oxygen contamination is formation of donor states in the crystal. This is thought to be caused by SiO_4 complexes and or complexes formed with acceptor atoms. Clearly, in the second case, the presence of oxygen doubly compensates the acceptor impurity by essentially converting it to a donor. Of course, these effects must be closely controlled to maintain a stable resistivity.

In contrast to oxygen, carbon atoms are generally substituted into silicon lattice sites. (This is expected since carbon is a Group IVB element.) Carbon is neither electrically active, *i.e.*, it does not act as either a donor or acceptor, nor does it tend to form precipitates as does oxygen. Even so, its presence is generally undesirable because carbon tends to enhance precipitation of oxygen and formation of intrinsic point defects.

Internal Gettering

Internal gettering is an important process technique that has found wide use in various fabrication processes. In general, gettering methods are used to sequester harmful defects and impurities away from the electrically active areas of a device or circuit. The terminology of gettering actually derives from the days of vacuum tube electronics when a chemically active material or *getter*, *e.g.*, an alkali metal such as cesium or potassium, was placed inside the glass envelop before final evacuation and seal. After sealing, the tube was heated to activate the getter, thus, removing residual oxygen and nitrogen gases from the interior atmosphere. The situation for semiconductor electronics is similar except that instead of residual gases, it is desirable to remove metallic contaminants and associated defects. (Metallic contamination, *e.g.*, iron, nickel, chromium, *etc.*, is particularly destructive to device performance.)

It has long been known that a region of crystal damage captures contaminant atoms and defects. This occurs because lattice energy must be increased in order to “fit” foreign contaminant atoms either into lattice or interstitial sites. This additional energy is not required if pre-existing defects already disrupt the lattice. Hence, in the course of thermal processing contaminant atoms diffuse and tend to be collected by defects. Furthermore, defects themselves are not stationary, but also migrate during thermal processing. In general, increases in lattice energy associated with isolated defects tend to be reduced if defects congregate into a damaged region. To promote this process, it is useful to create a defected or damaged region intentionally somewhere on or within the wafer prior to thermal processing. If, for example, the damaged region is on the backside of a wafer, then contaminants and defects are effectively removed from the front side. Since, integrated circuit elements are customarily fabricated on the front of the wafer such a scheme can be quite beneficial. However, this does not mean that all defects and contamination can be rendered innocuous. Therefore, due care must still be taken to prevent defect formation and contamination. There are a number of ways to set up an effective gettering scheme. Early implementations required introduction of defects into the backside of the wafer by sand blasting or rapid oxidation in a phosphorus oxytrichloride, $POCl_3$, ambient. More recently, high dose argon ion implantation or polysilicon deposition on the wafer backside have been used for this same purpose. All of these methods are conceptually similar and can be called *extrinsic* or *external gettering* because they require external doping or damage. In all cases, the damage is created as late in the process as possible to minimize the possibility of the defects being annealed out by subsequent thermal processing, and thus, re-releasing deleterious impurities previously captured back into the bulk.

In contrast, internal gettering schemes, which manipulate primary impurities, *viz.*, oxygen, introduced during manufacturing of the substrates themselves (CZ process), have become attractive. A typical scheme begins by driving off oxygen from the wafer surface (in which active devices will be subsequently fabricated) by means of an initial high temperature anneal step in an inert ambient, *e.g.*, argon. At high temperature, oxygen is very mobile in silicon and is easily lost through the surface. This creates a *denuded zone* at the surface of the crystal. Conceptually, the formation of this denuded zone can be considered as a kind of inverse doping, in which impurity atoms are lost from the surface rather than added. Naturally, the thickness of the denuded zone depends on the temperature and length of heat treatment. Following the denuding step, wafers are then annealed at lower temperature to nucleate oxygen clusters within the bulk silicon. Of course, this occurs in the bulk below the denuded zone because the oxygen concentration exceeds the solid solubility limit at the lower temperature. (As indicated previously, it is thought that carbon also plays some role in the nucleation of oxygen clusters, *i.e.*, oxide precipitates.) When a sufficient degree of nucleation has been achieved, annealing temperature is then raised to induce a faster cluster growth rate. Once an oxide precipitate reaches a critical size, the resulting lattice strain causes formation of dislocation loops and stacking faults. These defects then act as active gettering sites. It is important to note that the temperature in the growth phase cannot be taken too high, otherwise the concentration will fall below the solid solubility limit and oxygen clusters will redissolve rather than grow. To understand this process more fully, it is worthwhile to consider the behavior of oxide precipitates in some detail.

Nucleation and growth of oxide precipitates can again be treated from a thermodynamic point of view. Thus, one can write down an expression for the free energy of formation of an oxide precipitate comprised of *N* stoichometric units, *e.g.*, moles, of silicon dioxide, SiO_2:

$$\Delta A = N\Delta E_{\mathrm{SiO_2}} - NT\Delta S_{\mathrm{SiO_2}} + A\sigma + gV$$

Here, *A* (do not confuse with the free energy change, ΔA) is the surface area of a single precipitate and *V* is its corresponding volume. The parameters, σ and *g*, are, respectively, free energy of formation of new silicon/silicon oxide interface per unit area, *i.e.*, solid surface tension, and the lattice strain energy per unit volume induced by an oxide precipitate. The thermodynamic quantities, $\Delta E_{\mathrm{SiO_2}}$ and $\Delta S_{\mathrm{SiO_2}}$, are energy and entropy of formation of one stoichometric unit of SiO_2 from one stoichometric unit of silicon atoms within the lattice and two stoichiometric units of oxygen atoms in interstitial sites. Extensive reference tables of standard heats (enthalpies) of formation and entropies have been compiled and this information can be used to estimate these quantities for oxide formation within the silicon lattice. (For a condensed phase, heat and energy of formation can, of course, be considered as the same.) In particular, the entropy change must include a configurational contribution obtained by an analysis very similar to the preceding treatment of vacancies and silicon self-interstitials. Similarly, the energy change must include contributions from various binding energies associated with the silicon crystal lattice, oxygen interstitials, and oxide precipitates. (For more details, refer to Appendix A.)

If, for simplicity, precipitates are regarded as spherical, then this equation can be modified as follows:

$$\Delta A = \frac{4\pi r^3}{3}(n\Delta E_{\mathrm{SiO_2}} - nT\Delta S_{\mathrm{SiO_2}} + g) + 4\pi r^2\sigma$$

Here, *n* is stoichometric density, which relates number of stoichometric units to volume. (Specifically, *n* can be specified in moles/cm^3 by dividing the ordinary mass density of SiO_2 by the formula weight.) The standard energy of formation of SiO_2 is negative since oxidation of silicon is exothermic. Similarly, the entropy change is also negative since SiO_2 is more ordered than dissolved oxygen atoms randomly distributed in silicon. However, due to the explicit negative sign, the entropy term must make a positive contribution to ΔA. Furthermore, both the strain, g, and surface energy, σ, terms make positive contributions to the free energy. Therefore, only if the formation energy term is sufficiently negative, is it possible for the cubic term to be negative and oxide precipitates to be thermodynamically stable. Clearly, if temperature is sufficiently high (as is characteristic of denuding), then formation energy is totally compensated by entropy and strain terms. Thus, at high temperature oxide precipitates of any size are unstable and dissolve into the silicon lattice. However, even at lower temperatures there are further complications. In particular, since the surface energy coefficient is positive, the quadratic term must dominate the cubic term as oxide precipitate radius tends toward zero. Thus, very small oxygen clusters can never be thermodynamically stable under any condition. Clearly, after denuding at high temperature (~1100°C), oxide precipitates are most likely absent, having been dissolved. One might ask then, how could oxide precipitates ever be reformed? What is required is some non-equilibrium nucleation process. To consider this question, it is useful to digress briefly and discuss the nature of thermodynamic equilibrium in general.

By definition, thermodynamic equilibrium defines a dynamic, not a static steady state. This means that both “forward” and “reverse” processes occur at the same rate which, of course, results in a net rate of zero, *i.e.*, a steady state. Thus, in the present case, small oxygen clusters are randomly forming and dissolving continuously within the bulk silicon crystal. Clearly, if the net process is shifted away from equilibrium (by changing temperature, for example), forward and reverse rates are no longer equal with the thermodynamically favored one, *i.e.*, the one with a negative free energy change, occurring at a higher rate. However, if the process is still relatively close to equilibrium, then the non-favored process, *i.e.*, the one with a positive free energy change, will still proceed to some degree. This condition will persist until equilibrium is re-established under new conditions in which case, both rates are once again equal. To apply this concept to oxide precipitate formation, suppose that after the denuding step, temperature is reduced rapidly. Obviously, at the lower temperature, oxide precipitates larger than some critical radius are stable. However, after denuding, essentially no oxygen clusters exist in the bulk. Therefore, the system is not at equilibrium since there are no oxygen clusters present to undergo the “reverse” process, *i.e.*, dissolution of oxygen clusters. Therefore, only the “forward” process, *i.e.*, formation of oxygen clusters, can proceed to any appreciable extent. Thus, if the annealing temperature is reduced after denuding and if the oxygen concentration in the wafer is sufficiently high, even though they are not strictly thermodynamically stable, some oxygen clusters will form spontaneously. Clearly, during nucleation, oxygen clusters are continuously and randomly nucleated and re-dissolved. However, it is clear from the preceding form given for the free energy of a precipitat, that if by chance an oxygen cluster grows larger than the critical radius, continued growth becomes more favorable than dissolution. Therefore, during the nucleation step, one expects some oxygen clusters to form and some fraction of these to grow larger than the critical radius instead of re-dissolving. Clearly, the temperature chosen for the nucleation process must be sufficiently low so that the critical radius is reasonably small.

The size of the critical radius can be determined by consideration of the partial derivative of free energy with respect to cluster radius:

$$\frac{\partial}{\partial r}\Delta A = 4\pi r^2(n\Delta E_{\mathrm{SiO_2}} - nT\Delta S_{\mathrm{SiO_2}} + g) + 8\pi r\sigma$$

Clearly, maximum free energy must correspond to the critical radius, because a cluster of this size has the same tendency to either grow larger or re-dissolve, *i.e.*, the free energy change for both processes is negative. Thus, one sets the partial derivative equal to zero and solves as follows:

$$r_{crit}(n\Delta E_{\mathrm{SiO_2}} - nT\Delta S_{\mathrm{SiO_2}} + g) + 2\sigma = 0$$

From this one immediately obtains:

$$r_{crit} = -\frac{2\sigma}{n\Delta E_{\mathrm{SiO_2}} - nT\Delta S_{\mathrm{SiO_2}} + g}$$

Clearly, the lower the temperature, the larger the magnitude of the denominator and, hence, the critical radius is reduced. As observed previously, a small critical radius is desirable since this reduces the range of instability or *nucleation gap* and results in more efficient formation of oxide precipitates. One might ask, why not nucleate oxygen clusters at room temperature (or even lower)? Thermodynamically, this might be favorable, however the rate of oxygen cluster formation becomes so low that such a process is so slow as to be completely impractical. Therefore, in practice it is found that annealing temperatures of a few hundred degrees are optimal for oxygen cluster nucleation.

As asserted previously, after sufficient oxygen cluster nucleation is achieved, it is desirable to raise the annealing temperature to promote further growth of oxide precipitates. Of course, this causes re-dissolution of smaller nuclei since the critical radius becomes larger at higher temperature. (Clearly, oxide precipitates, which are smaller than the new critical radius defined by the higher temperature, but which were stable at the lower temperature of the nucleation process, become unstable and redissolve.) However, precipitates of radius larger than the critical radius at the higher temperature remain stable and, indeed, tend to grow larger. In addition, the increased processing temperature results in faster precipitate growth (and less processing time). Finally, once oxide precipitates become sufficiently large, they induce defects (dislocations and stacking faults) in the surrounding silicon lattice. These become active gettering sites and due to the initial denuding step, as desired, defects are absent within a surface layer which is typically at least a few microns thick. Of course, it is precisely in this undefected surface layer that active integrated circuit elements are to be fabricated. Thus, internal gettering provides a particularly elegant scheme in which active gettering sites are naturally located in close proximity to, but do not interfere with critical circuit elements. To summarize, the general characteristics of an internal gettering process can be illustrated as follows:

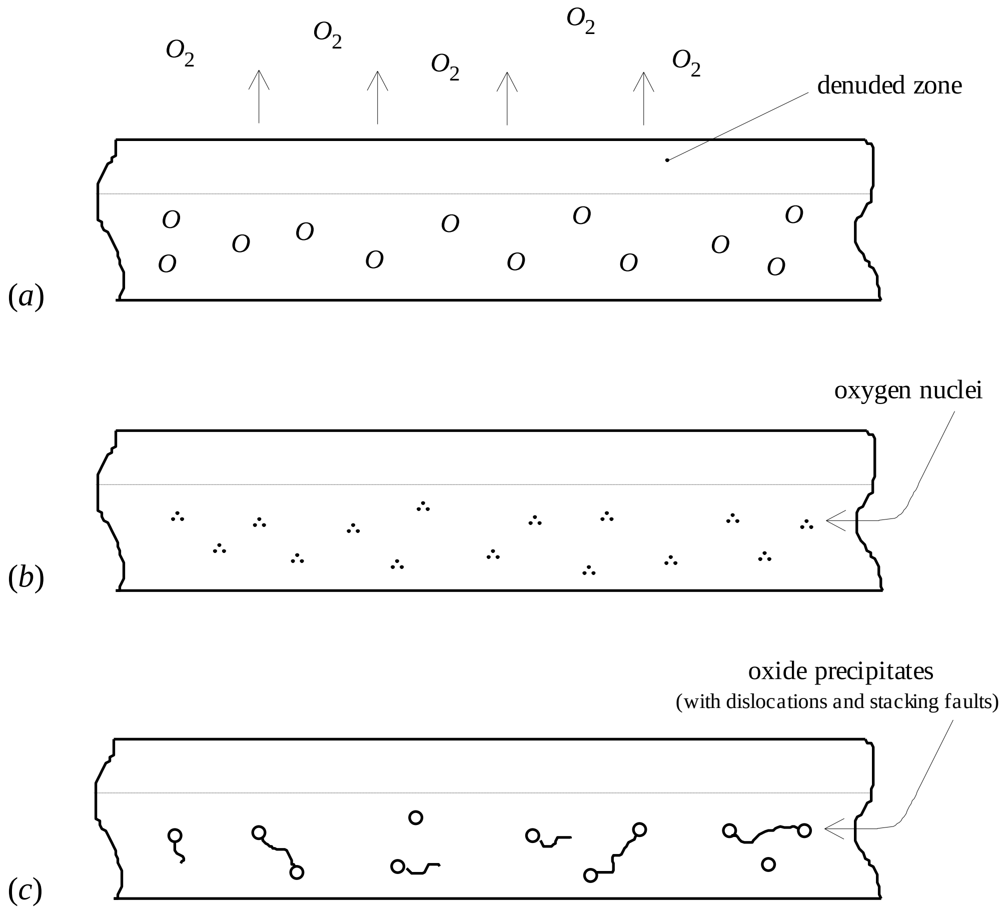


Fig. 23: Internal gettering process (*a*) denuding; (*b*) nucleation; (*c*) precipitate growth

Several factors serve to limit the size of oxide precipitates. First of all, it is obvious that once the available supply of dissolved oxygen is exhausted, precipitates can no longer grow larger. Second, very large precipitates result in high lattice strain. This exerts a very high pressure and associated large positive contribution to free energy, thus, retarding further growth and limiting precipitate size. In passing, it should be noted that thermal processing used for an internal gettering scheme need not be separate from thermal processing used for other purposes. This is attractive since fewer individual process steps are required in the whole integrated circuit fabrication process.

Ingot and Substrate Characterization

Classical methods for studying crystal defect structure are the metallographic methods. These require the use of selective etches which delineate the defect structure of the material. Various etches have been formulated for different kinds of defects in different kinds of materials and silicon is no exception. Selective etching can delineate many line, plane, and spatial defects. Therefore, once, the sample has been prepared, the delineated defect structure can be examined directly by optical microscopy. There are several common defect etches used for single crystal silicon. Virtually all of these contain hydrofluoric acid and a chemical oxidizing agent. (These go by a variety of names, *e.g.*, Secl etch, Wright etch, Yang etch, *etc.*) Each one is optimized for a particular defect structure or related use. The idea is the same in all cases; the etchant attacks the defected area because bonding in the lattice is disrupted allowing preferential attack by the etching chemistry. Typically, dislocations intersecting the crystal surface will result in a pyramidal shaped etch pit. Stacking faults will be revealed as linear features (planar features seen edge on) often with precipitates visible at the ends. Of course, defect etching is a destructive technique since it removes parts of the substrate surface. Nevertheless, metallographic techniques are still quite useful for process development and characterization.

X-ray Methods

It has been known for many years that atomic spacing characteristic of solid crystals is of just the right size to act as a diffraction grating for x-rays. Thus, x-ray diffraction is a powerful tool for the characterization of crystalline materials. The essential geometry of x-ray diffraction is illustrated below:

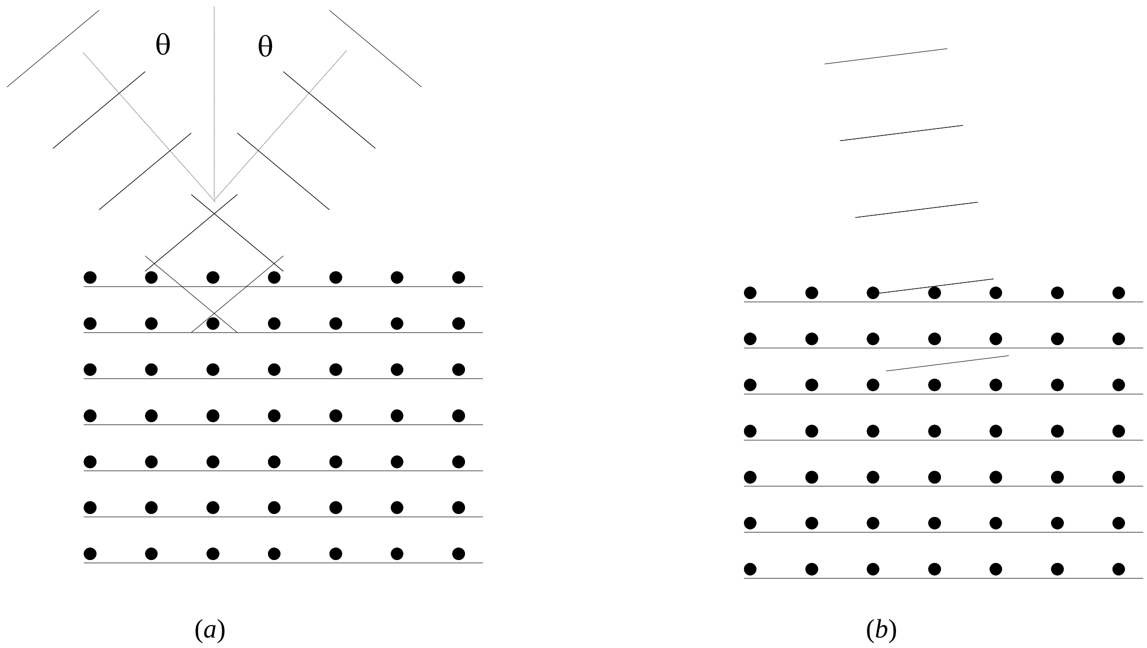


Fig. 24: X-ray diffraction (*a*) constructive interference; (*b*) destructive interference

Briefly, atomic crystal planes defined by Miller indices, *i.e.*, [*hkl*], act as specific reflectors for x-rays of a definite wavelength and incident angle, θ, (known as the *Bragg angle*). By varying the incident angle of monochromatic x-rays illuminating a single crystal, one can image a regular array of diffraction maxima. Such a pattern is called a *Laue pattern* and is characteristic of the material itself. Indeed, the Laue pattern can be used to construct a detailed picture of the atomic structure of a crystalline solid. (This is called *x-ray crystallography*.) It turns out that individual diffraction maxima of a Laue pattern correspond to points of a lattice defined in reciprocal space. (Reciprocal space was mentioned previously in connection with crystal orientation.) Each point of the *reciprocal lattice* corresponds to a reflection from a specific set of atomic planes, *i.e.*, a specific set of Miller indices. In general, the more sharply defined the diffraction maxima are, the better is the quality of the crystal. (A complete treatment of crystallography is far beyond the scope of the present course.) The closely related *Laue back-scatter method* is used to characterize large silicon crystals, *e.g.*, as-grown boules. This is because the crystal is usually too thick for classical transmission diffraction patterns to be obtained. Thus, for this method an unfiltered, broad wavelength band x-ray source is reflected from the surface of a boule and, thus, a Laue pattern is generated. However, the pattern is distorted since diffraction comes from various radiation wavelengths. Even so, this has the advantage of alleviating the need to move the sample to the exact Bragg angle as is necessary if a monochromatic x-ray source is used. Obviously, the lattice parameter is known a priori since the crystal is known to be silicon. Hence, the resulting back-scatter diffraction pattern allows precise determination of orientation and overall crystal quality. Accordingly, this method is used routinely by wafer manufacturers.

X-ray topography is another important imaging technique useful for the characterization of crystalline materials. Contrast is achieved through changes in the interplanar spacing existing within a crystal. (Changes in interplanar spacing change diffracted intensity if the crystal is oriented near a Bragg angle.) Homogenous strain and/or the defect structure of the crystal cause these changes. In x-ray topography a monochromatic, collimated source of x-rays is needed. Such an x-ray source can be realized either through use of apertures and filters or by use of a collimating crystal. In the second case, the crystal is oriented such that incident x-rays at the desired wavelength are reflected at a Bragg angle. (This also serves to produce a monochromatic beam since only one wavelength meets the Bragg criterion.) Dislocations, stacking faults, and precipitates all can be made visible using this technique. In addition, edge and screw dislocations can be distinguished. In the case of an edge dislocation, if the plane of reflection is perpendicular to the axis of the dislocation, *i.e.*, coincides with the slip plane, no contrast will be generated since the lattice spacing in this direction is minimally affected by the defect. A similar situation holds for a screw dislocation. Again, for a screw dislocation, no contrast is generated by reflection from the slip plane; however contrast is generated by reflections from planes perpendicular to the slip plane. Furthermore, edge and screw dislocations have characteristic intensity ratios for reflections parallel and perpendicular to the dislocation axis, which allows them to be easily distinguished. The double-crystal topographic arrangement also allows for measurement of strain in a crystal. In this case, the x-ray beam must be highly monochromatic and collimated. The sample is oriented at a Bragg angle and reflected intensity maximized. Following this “setup procedure”, the sample is then “rocked” through the maximum to generate an accurate diffraction lineshape. The width of the line is a direct measurement of the strain in the crystal. Such *rocking curves* are a direct indication of crystal quality since the existence of strain is often the result of defects.

Other Methods

Transmission electron microscopy (TEM) is another important material characterization technique. It is analogous to ordinary transmission optical microscopy except that the image is formed by electron waves rather than light waves. One disadvantage of TEM for the characterization of silicon substrates is that it requires a very thin section that is effectively transparent to electrons. This often requires tedious preparation using various chemical and physical techniques to thin a section of the sample. In practice, TEM is more useful for the characterization of process induced defects in the substrate rather than determination of starting material quality and, as such, is typically used for failure analysis. Indeed, very good images of dislocations, stacking faults, twins, precipitates, and volume defects can be obtained. In addition, electron diffraction patterns can also be obtained which are analogous to Laue x-ray diffraction patterns.

Fourier transform infrared spectroscopy (FTIR) is often used to determine the oxygen content of CZ substrates. Typical values of oxygen concentration in CZ wafers are in the $5(10^{17})$ cm^{-3} range. For intrinsic gettering, characterization of this concentration is highly important and is often specified by wafer fabricators.

Wafer Finishing

Silicon wafers are, of course, fabricated and finished from ingots (or boules), which are produced almost exclusively using the CZ process. As might be expected, many of the actual details of wafer finishing are proprietary; however, it is worthwhile to summarize generic processes. First of all, no as-grown ingot has a perfectly constant radius and typically has an uneven surface that typically appears rippled or wavy along the length of the ingot; therefore the ingot must be cut and ground to a specified shape. For integrated circuit manufacturing, this is a circular cross section of up to 450 mm in diameter (however, 200 and 300 mm diameters are still more common). In contrast, for solar cells this is usually a square cross section with rounded corners. Of course, due to the hardness of elemental silicon, diamond tooling is necessary for this operation. Once the desired shape has been fabricated, raw slices of specified thickness (usually from few hundred microns for small wafers to roughly a millimeter for large wafers) are then cut from the ingot using a sophisticated wire saw and diamond abrasive slurry. Slicing is illustrated schematically below:

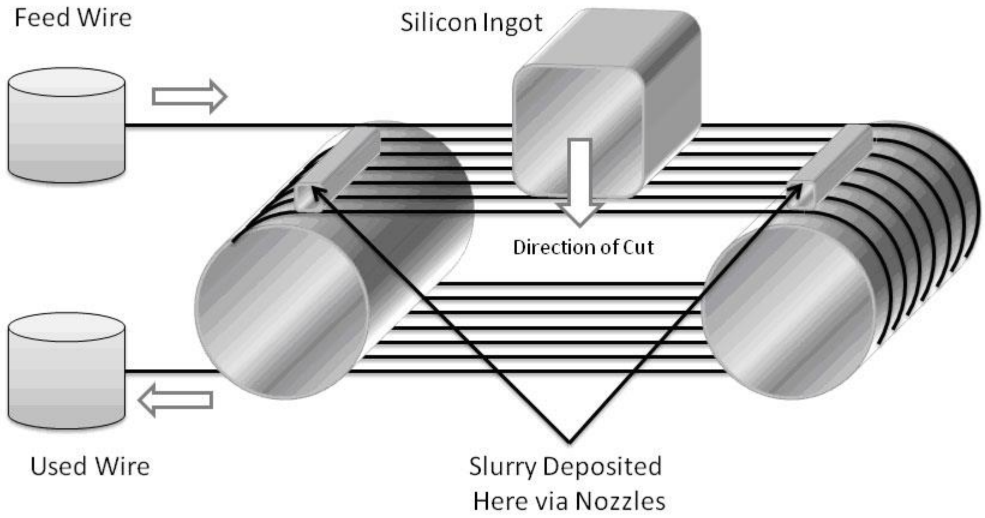


Fig. 25: Slicing of a silicon ingot (here shaped for solar cell fabrication)

In addition, wafer edges are shaped, *i.e.*, rounded, after slicing to prevent crack propagation and consequent fragility. Of course, the raw sawn surface is not to be expected to be suitable for device fabrication and must be polished.

Accordingly, chemical mechanical polishing (CMP) of wafers is done using a planar polishing machine and a chemically active slurry. Typically the slurry consists of fumed silica (SiO_2) dispersed in an alkaline solution (pH~12-14). Polishing pads are made of highly engineered composite textiles, typically of polyurethane or polyester. (In passing, it should be noted that this type of processing, long used to fabricate wafers, has more recently been introduced to integrated circuit manufacture as well.) A pictorial representation of CMP is shown in the following figure:

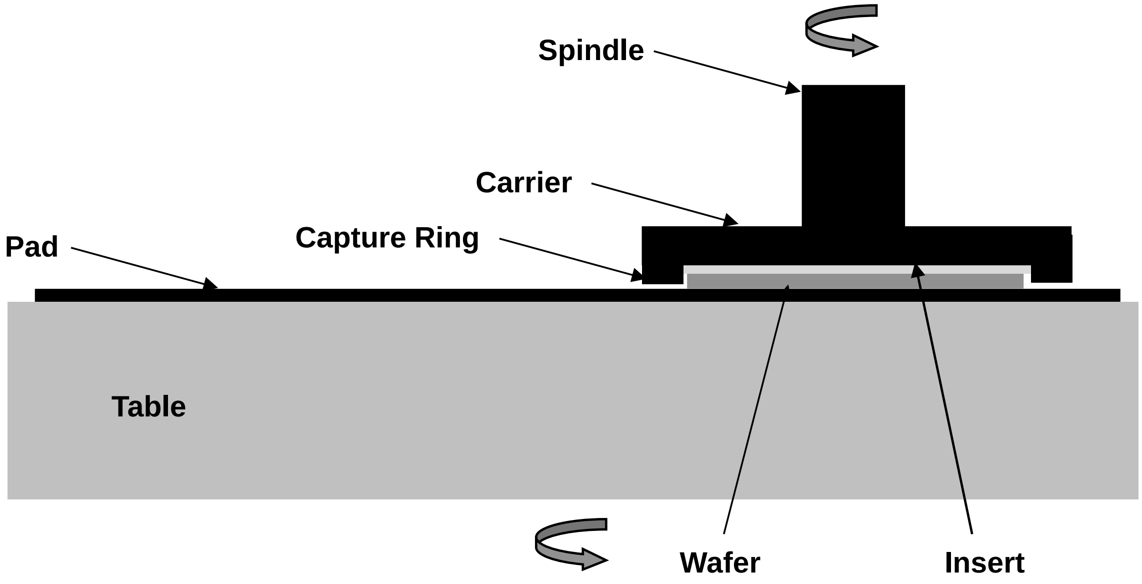


Fig. 26: Schematic of CMP machine

Here, slurry is introduced to the pad through a nozzle (not shown) and is entrained underneath the wafer by the rotation. For clarity, a single wafer configuration is illustrated; however, multiple wafers may be polished simultaneously on the same pad. Moreover, it might seem surprising that wafer and pad rotation is in the same direction; however, an elementary kinematic analysis readily demonstrates that the magnitude of relative surface velocity between the wafer and pad is more uniform in this configuration. (Indeed, relative surface velocity magnitude is exactly the same over the entire wafer surface if rotation rates of the pad and the wafer are exactly equal; however, this can result in “pattern coincidence; therefore, it is usual to rotate the pad and wafer at slightly different rates such that the ratio of the rates is irrational.)

Slurry residue is removed after polishing by specialized surface cleaning equipment while the wafers are still wet. Final chemical cleans follows (if necessary) and the finished wafers are packaged under ultraclean conditions. Within this context, wafer surfaces must approach atomic flatness, *i.e.*, any roughness must be on the nanometer scale or less. It might seem that this would be difficult to achieve; however, this is not the case.

Silicon Nanowires

Of course, over the whole history of modern semiconductor processing, wafers have represented (and continue to represent) the dominant physical form for semiconductor grade silicon used in microelectronic fabrication. Over time, the only significant change in this situation has been a continuing increase in wafer diameter (and coincident scaling of thickness) from less than 50 mm in the late 1950's to as large as 450 mm substrates at present. (Indeed, larger wafer sizes have been proposed, but it remains to be seen if these can be cost effective.) In any case, in analogy to structural steel it is likely that silicon wafers will remain an important item of commerce for many years to come. In contrast, whiskers of various materials have been known for more than fifty years. (The usage of the rubric "nanowires" is of relatively recent advent.) Indeed, a detailed description of silicon nanowire growth by researchers at Bell Labs appeared as early as 1964. Even so, for much of this time such structures remained at best merely laboratory curiosities and at worst appeared as troublesome defects in conventional manufacturing processes. It has only been in the last decade or so that "nanostructures" have become a specific object of research.

Vapor-liquid-solid (VLS) Growth Process

In contrast to growth of bulk silicon crystals, silicon nanowires are commonly grown using the *vapor-liquid-solid* or *VLS* process. This requires small, *i.e.*, nanometer-sized, particles of metal to be deposited on the surface of a larger substrate crystal. As ambient temperature is raised, the metal particles melt and absorb silicon from a gaseous precursor (usually silane, SiH_4) catalyzing silicon crystal growth at the liquid-solid interface. Clearly, as suggested by the term VLS, process temperature must be chosen such that the substrate remains solid, the catalyst is liquid, and the precursor vapor pressure is sufficient to supply silicon to the growth process at a sufficient rate. Accordingly, it is evident that nanowire growth requires establishment of favorable thermodynamic conditions across two heterogeneous phase boundaries, *viz.*, the vapor-liquid interface at the surface of the catalyst droplet and the liquid-solid interface between the catalyst and the growing nanowire. (Obviously, the liquid-solid interface has some similarity to the melt-ingot interface in conventional CZ and FZ crystal growth processes.) Within this context, it might seem that such conditions would be difficult to realize in practice; however, this is not the case. Indeed, a number of metals can serve as catalysts in the VLS process. Moreover, just as in conventional crystal growth, the orientation of the underlying substrate determines the orientation of the growing nanowire. However, in contrast to growth of bulk crystals not all nanowire orientations can be realized. The reason for this is due to basic thermodynamic constraints associated with the VLS process and for silicon nanowires only growth of the [111] orientation is found to be practical. Naturally, other kinds of nanowire materials can be expected to have different orientation dependence. In any case, typical characteristics of VLS growth are illustrated in the following figure:

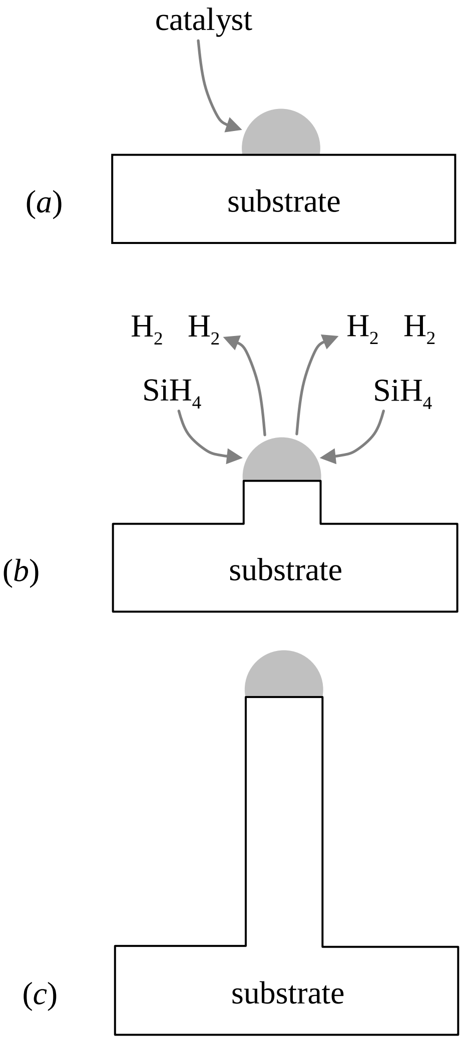


Fig. 27: Vapor-liquid-solid (VLS) process (*a*) catalyst on substrate; (*b*) growth; (*c*) finished nanowire

In general, nanowire length can be controlled by growth time; however, due to geometric as well as other effects considerable variation is to be expected.

Obviously, before silicon nanowires can be grown a suitable catalyst material must be identified. Clearly, such a catalyst should satisfy at least two fundamental requirements: First of all, it should have a reasonably low melting point with respect to silicon and, second, silicon and the catalyst material should form a well-defined *eutectic* alloy. Within this context, it turns out that metallic gold is the most common catalyst used for growth of silicon nanowires. At first glance this might seem unlikely since the melting point of pure gold is nominally, 1064°C; however, an alloy having atomic composition 18.6% silicon-81.4% gold, melts at only 363°C and, moreover, forms a eutectic mixture. This is illustrated by the well-known gold-silicon binary alloy phase diagram as shown in the following figure:

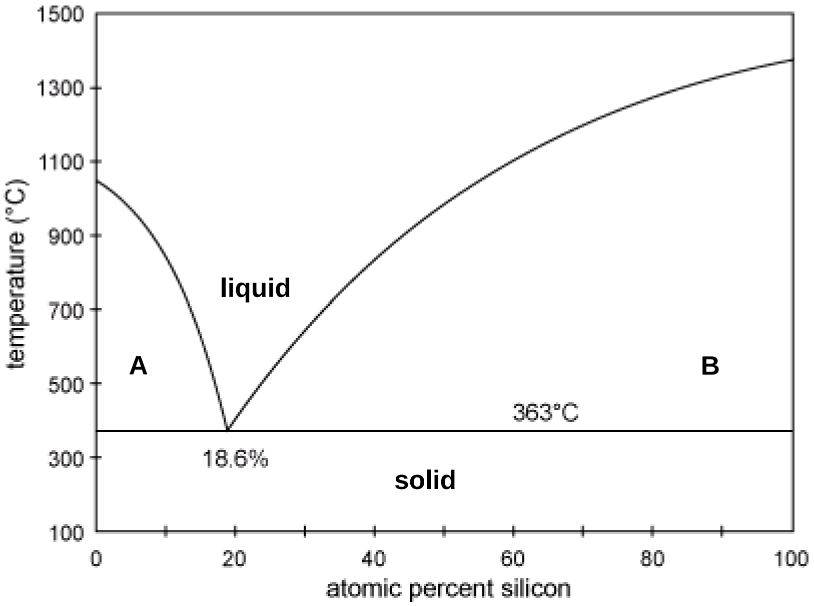


Fig. 28: Gold-silicon binary alloy phase diagram

Physically, a eutectic alloy corresponds to a binary mixture of materials, typically metals, having well-defined composition and minimum melting point. Below this temperature all mixtures irrespective of composition are solid. Thus, regions in the figure labeled "**A**" and "**B**" denote mixtures consisting of liquid eutectic and either solid gold or silicon, respectively. Accordingly, "liquidus" curves rise on either side of the "eutectic point" and define boundaries between the liquid phase and two-phase solid-liquid mixtures. (Likewise, "solidus" curves correspond to the horizontal line.) As might be expected, liquidus terminal points are identified with pure materials and, thus, in the gold-silicon phase diagram can be identified merely as standard melting points for gold or silicon. For completeness, it should be noted that other metals, *e.g.*, aluminum, copper, *etc.*, can also be used to grow silicon nanowires and, moreover, in analogy to gold form relatively low melting eutectic alloys.

Of course, once a catalyst material has been chosen, particles of this material must be controllably deposited or synthesized on the surface of the seed substrate. Accordingly, there are several different techniques for this, but perhaps the simplest method is to first deposit a catalyst film at low temperature (*e.g.*, near room temperature) by vacuum evaporation (or some other suitable technique). Upon subsequent heat treatment, if the deposited film is very thin (typically less than 10 nm) recrystallization causes a continuous film to break up into individual small crystallites. This phenomenon is called *agglomeration* and is particularly favored for thin films of noble (or semi-noble) metals such as gold. Of course, this produces a wide distribution in particle size which generally results in a similar variation in finished nanowire length. Nevertheless, this process is very simple and economical. Alternatively, pre-formed catalyst particles of controlled size may be deposited on the seed. Indeed, colloidal gold particles of various sizes are commercially available and can be readily used as catalysts for silicon nanowire growth. Of course, there is usually substantial cost associated with manufacture of the particles;

however, this may be offset with higher quality nanowires. Obviously, both of these techniques produce random distributions of silicon nanowires on the seed substrate surface. It comes as no surprise that a more regular distribution might be technologically desirable. Naturally, this requires fabrication of some kind of regular template. Within this context, various techniques using self-assembly have been suggested; however, the most reliable method is direct photolithographic transfer of a regular pattern to a layer of masking material covering the seed substrate surface. After processing the result is a regular array of openings to the underlying silicon seed, which then can be coated with catalyst and a regular array of nanowires grown. Moreover, since the geometry of the template can be precisely controlled a tight distribution of nanowire diameter and length is to be expected.

Clearly, once catalyst particles are in place, nanowire growth can begin. This is generally done at a temperature higher than the melting point of the catalyst-silicon eutectic in an atmosphere containing hydrogen and silane or chlorosilane. Accordingly, the silicon containing precursor gas is pyrolyzed on the surface of the catalyst droplet releasing silicon which dissolves in the catalyst. The concentration of silicon in the molten catalyst is controlled by a heterogeneous equilibrium between the gas phase and catalyst-silicon solution. In addition, a separate heterogeneous equilibrium exists between the catalyst droplet and the growing solid nanowire. In particular, once the concentration of silicon becomes sufficiently high within the molten eutectic, solid silicon crystallizes at the melt-solid interface. Moreover, this crystallization preserves crystal orientation of the underlying substrate. In principle, such a process can continue as long as precursor gas is supplied to support the growth process. Within this context, one might wonder why silicon nanowires grow only from the catalyst-wire interface. Indeed, direct epitaxial growth of silicon has been known for decades and, moreover, is widely used in commercial fabrication. The reason that direct growth does not occur at any appreciable rate during nanowire growth is, naturally, a consequence of the catalyst. Indeed, this is the fundamental function of any catalyst, which by definition does not change overall thermodynamics of a chemical reaction, but increases the rate due to a lowering of energetic barriers. In this case, catalyzed growth occurs at a much lower temperature, *e.g.*, 500-600°C, in comparison to direct growth, which requires temperatures of 1000 to 1100°C.

Nanowire Processing

Obviously, although single crystal silicon, nanowires have a much different physical form than wafers. This requires substantially different processing strategies to produce useful devices. (Indeed, no widespread commercial applications of silicon nanowires have as yet appeared, although there is extensive research directed toward applications such as chemical and bio-sensors, low temperature electronics, photovoltaics, *etc.*) First of all, nanowires generally must be "harvested" from the growth substrate and deposited on some other prefabricated substrate; therefore, they must be detached either by etching or by some mechanical release method. Concomitantly, it is evident from their size that nanowires cannot be handled individually, but are usually dispersed in a liquid carrier to form an "ink". In addition, after growth nanowires are generally not all the same length and, moreover, some may be defective, *e.g.*, branched or curved. Therefore, some filtering process must be applied to select desirable nanowires and reject defective ones. Again, this is an area of active research, but suffice it to say that it is not an easy problem and simple implementations of filters generally do not work due to clogging and other difficulties. Coincident with this are various requirements for accurate placement of nanowires. Naturally, this strongly depends on the application. In the case of photovoltaics a random deposition may be acceptable as long as density can be controlled to facilitate uniform light capture and good electrical connections. However, for more sophisticated applications of nanowires as electronic devices, precise placement is necessary. To accomplish this, it has long been known that non-spherical structures dispersed in a flowing liquid tend to become oriented with respect to the direction of the flow. In addition, electrostatic capture may be employed to deposit nanowires at precise locations and, moreover, to sort them (at least partially) with respect to length.

Amorphous Silicon Dioxide

So far, both electronic and material properties of single crystal silicon have been considered in some detail. In addition, effects of defects and impurities have also been considered. All of these properties are essential to modern solid-state electronics; however, if the characteristics of the semiconductor material itself were all that was important, silicon would actually present little (if any) advantage over germanium or gallium arsenide. (Indeed, some other semiconductor might very well be better suited from the point of view of carrier mobility, *etc.*) Accordingly, there is another material, quite different from single crystal silicon, which is also of essential importance, *viz.*, amorphous silicon dioxide. Within this context, it is worthwhile to compare the most obvious characteristics of single crystal silicon and amorphous silicon dioxide, *i.e.*, quartz glass: 1) silicon is crystalline, quartz glass is amorphous, 2) silicon conducts heat and electricity reasonably well, quartz glass is a poor conductor of both, 3) silicon is an opaque, metallic appearing material (although it is transparent at infrared wavelengths), quartz glass is very transparent well into the ultraviolet region of the spectrum.

Indeed, the success of silicon solid-state electronics is due, in no small part, to the fact that high quality amorphous silicon dioxide thin films are easily produced by direct oxidation of silicon. Therefore, even though germanium was commercialized earlier than silicon and, moreover, although it has higher intrinsic electron and hole mobilities, because a high quality, chemically stable germanium dioxide (GeO_2) layer cannot be formed on a germanium surface by direct oxidation (germanium monoxide (GeO) sublimes at 710ºC) represents a serious limitation. As a consequence, silicon is the material of choice for industrial production of the vast majority of solid-state electronic devices (although germanium and especially silicon-germanium alloys have undergone somewhat of a renaissance in recent years, but generally in combination with silicon). Similar issues also exist for compound semiconductors such as gallium arsenide (GaAs), silicon carbide (SiC), *etc.* Indeed, as a practical matter, a semiconductor material other than silicon will be used only if it has some unique property that silicon does not have. For example, because of higher carrier mobilities GaAs and more recently indium phosphide (InP) have found some commercial use for fabrication of high speed, high frequency devices, such as amplifiers for cell telephones and wireless information networks. In addition, GaAs and other III-V materials are direct band gap semiconductors and, thus, useful for optoelectronic devices such as lasers and light emitting diodes (LEDs), which are applications for which silicon is not well suited. Similarly, silicon carbide may be useful if high temperature operation is required since it has a much larger band gap than silicon. (Diamond also has similar semiconductor characteristics.) Other semiconductors, such as indium antimonide (InSb), cadmium selenide (CdSe), *etc.*, may find use as specialty optical detectors or emitters; however, the production volume remains small and integration level low. Consequently, silicon successfully competes with (*e.g.*, in device speed) or surpasses (*e.g.*, in integration level) all other semiconductor materials for all but a few specific applications. In any case, the silicon/silicon dioxide material system is dominant and is likely to remain so for the foreseeable future. This remains true regardless of any consideration that essentially without exception, all other semiconductor materials are much rarer than silicon and, consequently, inherently more expensive. (However, in practice the cost of the substrate is generally only a small part of the cost of a finished integrated circuit or other solid-state electronic device.)

Direct oxidation of the surface of a silicon wafer at high temperature in an oxidizing atmosphere is known conventionally as *thermal oxidation*. The resulting thin quartz glass film is known as *thermal oxide*. As observed at the outset, quartz glass is not crystalline, but is amorphous with an open random network structure. This is in distinct contrast with silicon, which, of course, has a very well-defined crystal structure. The fundamental unit of the network structure is the SiO_4 tetrahedron. A diagrammatic representation of an SiO_2 network is shown below:

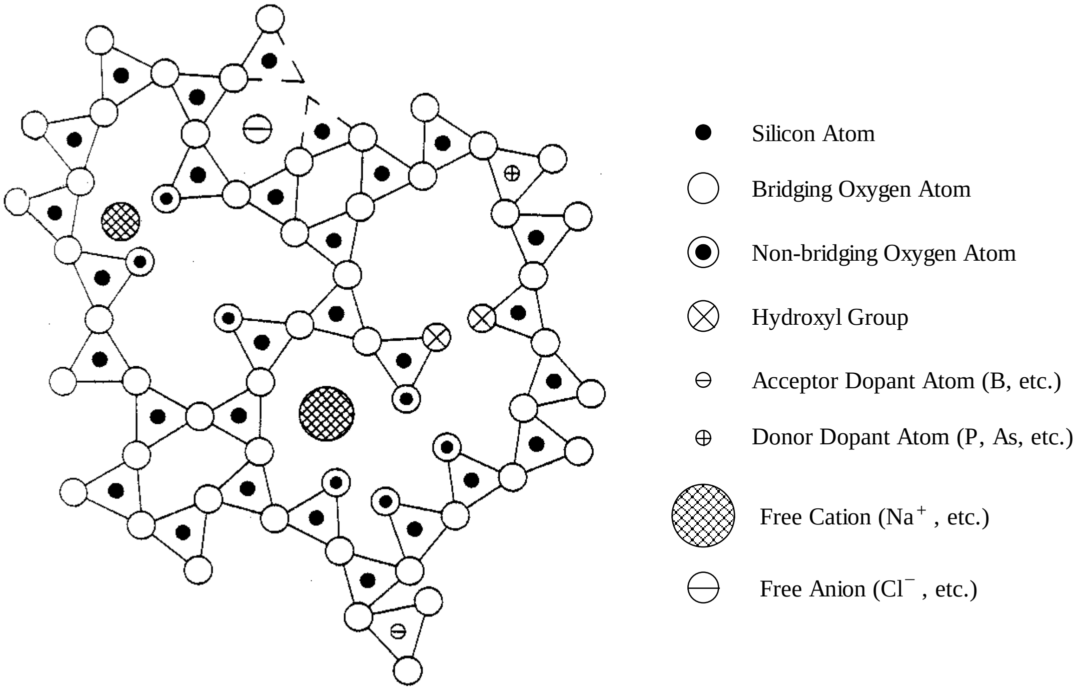

● Silicon Atom
○ Bridging Oxygen Atom
⊙ Non-bridging Oxygen Atom
⊗ Hydroxyl Group
⊖ Acceptor Dopant Atom (B, etc.)
⊕ Donor Dopant Atom (P, As, etc.)
Free Cation (Na$^+$, etc.)
Free Anion (Cl$^-$, etc.)

Fig. 29: Diagrammatic representation of quartz glass network structure

(Here, for convenience SiO_4 tetrahedrons are represented two dimensionally as triangles.) Indeed, thermal oxide has characteristics of both a liquid (*e.g.*, short-range order) and a solid (*e.g.*, rigidity and elasticity). Although the network structure of quartz glass is thermodynamically unstable below 1710ºC, the rate of *devitrification*, *i.e.*, crystallization, is negligible below 1000ºC. Therefore, once formed, thermal oxide is very stable under normal conditions.

Amorphous silicon dioxide has a well-defined refractive index of 1.46 and density of 2.27 g/cm^3. In a perfect structure, each SiO_4 tetrahedron is joined to four other tetrahedra, one at each apex. This implies that oxygen atoms must bridge between silicon atoms. Thus, in an ideal structure, each oxygen atom is bonded to two silicon atoms and each silicon atom is bonded to four oxygen atoms (hence, the stoichometric formula SiO_2). This results in a much less dense structure than single crystal silicon; therefore, the network structure of SiO_2 includes voids of various shape and size. Furthermore, the exact structure of these voids is generally process dependent. Additionally, some of the tetrahedra in the network may not be attached at all apexes. In this case, the oxygen atom must be bound to some other type of atom since two bonds are required. This is commonly hydrogen resulting in the incorporation of a hydroxyl (OH) group into the network structure. It is also possible for the silicon to become trigonally coordinated with only three oxygen atoms attached. Two of these are attached to other tetrahedra, the third one is unattached, *i.e.*, non-bridging, and is, in principle, doubly bonded to the silicon atom.

For the purposes of integrated circuit fabrication, thermal oxidation is a very effective process. It produces thin films of amorphous SiO_2 having a dense uniform network structure in comparison to other methods of thin film fabrication such as evaporation or chemical vapor deposition (CVD). The material properties of thermal oxide are quite uniform and invariant over time. Furthermore, even though it has an open structure, diffusion rates of many species in amorphous SiO_2 are quite low. Of particular importance are the usual shallow level impurities, B, P, As, Sb, Ga, *etc.* These species typically form oxides that themselves become strongly bound within the network (as illustrated in the preceding figure). Thus, SiO_2 is a very good mask for doping particular regions on the wafer surface. (This will be considered in more detail in later treatment of diffusion and ion implant processes.) Other species, which do not become bound in the network structure, diffuse quite rapidly in SiO_2. In particular, hydrogen diffuses quite readily as does oxygen, water, and a number of small inorganic anions and cations. All of these species diffuse through the voids in the network structure. (As will become evident, the fact that SiO_2 is permeable to H_2, O_2, and H_2O is of essential significance.)

Thermal Oxidation of Clean Silicon

As indicated previously, thermal oxidation of a clean silicon surface in an ambient oxidizing atmosphere is, perhaps, the most fundamental of all integrated circuit fabrication process. Physically, it is an example of a heterogeneous (gas-solid) chemical reaction. In conventional practice, either dry oxygen or pyrogenic steam is used as an oxidant. (Pyrogenic steam is produced by burning hydrogen and oxygen inside the oxidation furnace.) The two overall reactions are as follows:

$$\mathrm{Si} + \mathrm{O_2} \rightarrow \mathrm{SiO_2}$$

$$\mathrm{Si} + 2\mathrm{H_2O} \rightarrow \mathrm{SiO_2} + 2\mathrm{H_2}$$

Clearly, so-called *dry oxidation* in oxygen produces no gaseous products; however, *wet oxidation* in steam produces hydrogen as a byproduct.

The Deal-Grove Model of Thermal Oxidation

In general, an overall heterogeneous chemical reaction can be separated into several transport and reaction steps. First of all, the gaseous reactant must be transported from the bulk of the ambient gas atmosphere to the substrate surface. Accordingly, the flux of reactant to the substrate surface can be described by a simple mass transport equation:

$$F_1 = h_G(C_G - C_S)$$

Here, F_1 is oxidant flux to the substrate surface, C_G is bulk concentration of oxidant, C_S is the concentration of oxidant in proximity to the wafer surface, and h_G is a linear mass transport coefficient. This expression accounts for depletion effects in the gas phase due to consumption of oxidant by the reaction. Second, oxidant is dissolved in the surface of the thermal oxide film and diffuses to the Si/SiO$_2$ interface, hence:

$$F_2 = \frac{D}{x}(C_o - C_i)$$

Here, F_2 is oxidant flux diffusing through the growing thermal oxide film, C_o is dissolved oxidant concentration at the oxide surface, C_i is the dissolved oxidant concentration in the oxide at the Si/SiO$_2$ interface, D is the oxidant diffusivity in thermal oxide, and x is the thermal oxide layer thickness. Third, assuming first order kinetics, the oxidation reaction at the Si/SiO$_2$ interface corresponds to the expression:

$$F_3 = k_s C_i$$

In this case, F_3 is the oxidant flux (or, more correctly, a pseudo-flux) due to consumption of reactant by the oxidation reaction and k_s is a first order rate constant for the reaction.

Of course, oxidant concentrations, C_S and C_o, cannot be expected to be equal, but rather, to satisfy a heterogeneous distribution equilibrium across the gas-solid interface, *viz.*, Henry's Law:

$$H = \frac{C_o}{C_S}$$

Here, H is a distribution coefficient and is defined in analogy to distribution coefficients associated with crystal growth (except that the heterogeneous phases are gas and solid rather than liquid and solid). Clearly, H is closely related to the equilibrium solubility of the gaseous oxidant species in quartz glass and, naturally, is dependent on temperature and the microstructure of the glass. As has been noted previously, in the case of wet oxidation, a gaseous product, namely hydrogen, is formed. For generality, the diffusion flux of hydrogen back out of the oxide should also be considered since Le Chatelier's Principle implies that any local increase of hydrogen in proximity of the Si/SiO$_2$ interface must reduce the reaction rate, *i.e.*, favor the back reaction. However, since hydrogen is a small molecule and diffuses rapidly, it does not build up and its effects can be ignored. Clearly, for dry oxidation no gaseous products are formed and preceding expressions are entirely sufficient. Furthermore, assuming that wafer dimensions are much larger than film thickness, a one dimensional picture of thermal oxidation is satisfactory and is illustrated by the following figure:

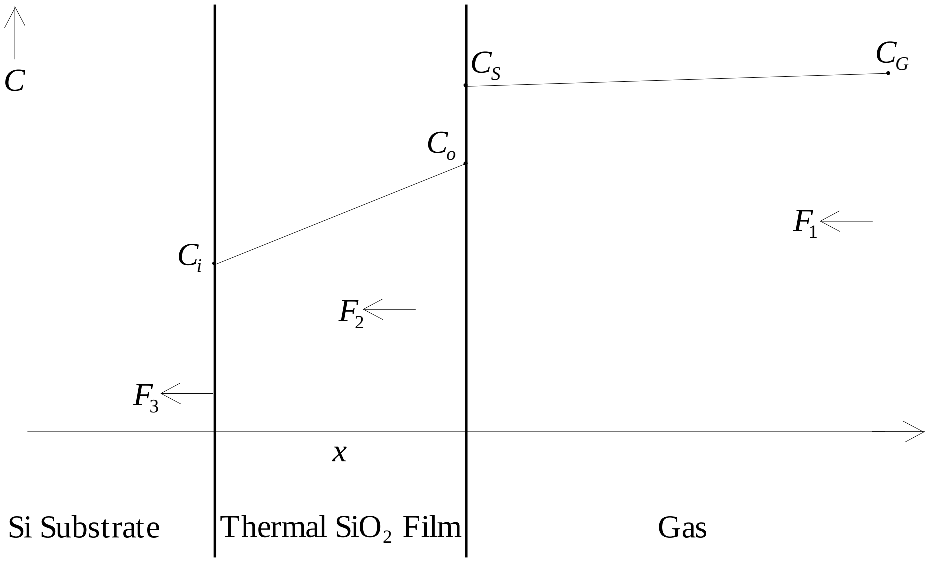


Fig. 30: Diagrammatic representation of the thermal oxidation of a clean silicon surface

Here, oxidant concentration, C, is plotted versus perpendicular dimension relative to the wafer surface. By definition, oxidant concentration within the silicon substrate is negligible.

Thus, assuming conditions of quasi-steady state, *i.e.*, assuming that any transients are small, all fluxes are taken to be equal. Accordingly, if one applies the distribution equilibrium and identifies F_1 as equal to F_3, one obtains:

$$k_s C_i = h_G\left(C_G - \frac{C_o}{H}\right)$$

Naturally, one solves this expression for C_G:

$$C_G = \frac{k_s C_i}{h_G} + \frac{C_o}{H}$$

Equivalently, one can identify F_2 as equal to F_3:

$$k_s C_i = \frac{D}{x}(C_o - C_i)$$

Consequently, this expression is solved for C_o:

$$\frac{D}{x}C_o = \left(k_s + \frac{D}{x}\right)C_i$$

$$C_o = \left(1 + \frac{k_s x}{D}\right)C_i$$

The two preceding expressions can be combined by substitution of this equation into the previous formula for C_G:

$$C_G = \frac{k_s C_i}{h_G} + \frac{1}{H}\left(1 + \frac{k_s x}{D}\right)C_i$$

$$C_G = \left[\frac{k_s}{h_G} + \frac{1}{H} + \frac{k_s x}{HD}\right]C_i$$

Inverting this equation to obtain an explicit form for C_i yields the desired result:

$$C_i = HC_G\left[\frac{Hk_s}{h_G} + \frac{k_s x}{D} + 1\right]^{-1}$$

Thus, the oxidant concentration at the Si/SiO$_2$ interface has been formally related to the concentration of oxidant in the gas phase. Of course, the concentration, C_G, is just fixed by gas pressure inside the furnace.

Naturally, the reaction flux, F_3, must be proportional to the thermal oxide growth rate; hence, one can write:

$$F_3 = N\frac{dx}{dt} = k_s HC_G\left[\frac{Hk_s}{h_G}+\frac{k_s x}{D}+1\right]^{-1}$$

Here, N is a proportionality constant relating the number of oxidant species arriving at the interface per unit area to the thickness of SiO_2 grown on that same area if all oxidant species react with the substrate. Of course, N is determined directly by consideration of the reaction stoichiometry and the density of the thermal oxide film. Clearly, this first order differential equation is easily integrated to give:

$$x^2 + 2D\left(\frac{1}{k_s}+\frac{H}{h_G}\right)x = \frac{2DHC_G}{N}(t+t_0)$$

Here, t_0 represents an initial condition, which in principle corresponds to some pre-existing thermal oxide layer thickness of x_0; therefore, one has:

$$t_0 = \frac{N}{2DHC_G}\left(x_0{}^2 + 2D\left(\frac{1}{k_s}+\frac{H}{h_G}\right)x_0\right)$$

Thus, t_0 is the time necessary to pre-grow a thermal oxide layer of thickness, x_0, under prevailing conditions, *i.e.*, growth conditions defined by current values of h_G, C_G, D, k_s, H, and N. Of course, in actual processing, the pre-existing oxide layer may be grown under different conditions; however, the properties of thermal oxide are sufficiently uniform so that only the thickness, x_0, is relevant to subsequent processing. Clearly, if x_0 equals 0, then t_0 equals 0.

Rather than expressing t as a function of x, it is desirable to express x as a function of t. This is easily accomplished by means of the quadratic formula:

$$x = D\left[\sqrt{\left(\frac{1}{k_s}+\frac{H}{h_G}\right)^2 + \frac{2HC_G}{ND}(t+t_0)} - \left(\frac{1}{k_s}+\frac{H}{h_G}\right)\right]$$

Thus, one obtains the general relationship between thermal oxide film thickness and growth time characteristic of the Deal-Grove model. From this formula, two important asymptotic expressions can be obtained. The first of these corresponds to the limit that t tends toward ∞. In this case, only the second term within the radical remains significant, hence:

$$x = \sqrt{\frac{2DHC_G}{N}t}$$

This defines the so-called *parabolic growth regime*. The second form is obtained if $t + t_0$ vanishes. In this case, one expands the radical as a Taylor series from which it follows that:

$$x = \frac{HC_G}{N\left(\frac{1}{k_s} + \frac{H}{h_G}\right)}(t + t_0)$$

This defines the so-called *linear growth regime*. Clearly, unless t_0 is small, *i.e.*, the initial oxide thickness is very small or absent, the linear growth regime cannot be realized.

Physically, the parabolic growth regime corresponds to the classical case of a *diffusion limited process* for which the rate limiting step is diffusion of oxidant through a relatively thick oxide film. Conversely, the linear growth regime corresponds to the case of a *reaction limited process* for which the rate limiting step is the interfacial reaction between oxidant species and the silicon substrate. In passing, one observes that another limiting regime, that of a *mass transport limited process*, is possible in principle. This situation would occur if oxidant became depleted in the gas phase in close proximity to the substrate surface. Clearly, this requires very rapid consumption of oxidant species by the oxidation process. However, in practice, oxidant transport in the gas phase is much more rapid than either diffusion of oxidant through the growing oxide film or the interfacial reaction itself. Therefore, a mass transport limited regime is never realized in conventional thermal oxidation processes, *i.e.*, for practical purposes, the coefficient, h_G, can be treated as indefinitely large.

In practice, one does not usually know (or care to know) all of the values of the various transport, equilibrium, and reaction rate coefficients. However, they can be collected into two aggregate rate constants, *A* and *B*, defined as follows:

$$A = 2D\left(\frac{1}{k_s} + \frac{H}{h_G}\right) \quad ; \quad B = \frac{2DHC_G}{N}$$

Therefore, in terms of *A* and *B*, the previous results can be recast as follows:

$$x^2 + Ax = B(t + t_0) \quad ; \quad t_0 = \frac{1}{B}(x_0^2 + Ax_0)$$

$$x = \frac{\sqrt{A^2 + 4B(t + t_0)} - A}{2}$$

Similarly, the parabolic and linear limiting forms are:

$$x = \sqrt{Bt} \quad ; \quad x = \frac{B}{A}(t + t_0)$$

By convention, *B* is known as the *parabolic rate constant* and *B*/*A* as *the linear rate constant*. Values for *A* and *B* (and or *B* and *B*/*A*) have been determined over a variety of conditions. Using these values, it is found that the Deal-Grove model describes thermal oxidation very well over a wide temperature range, *viz.*, 700°-1300°C. This is illustrated by the following figure:

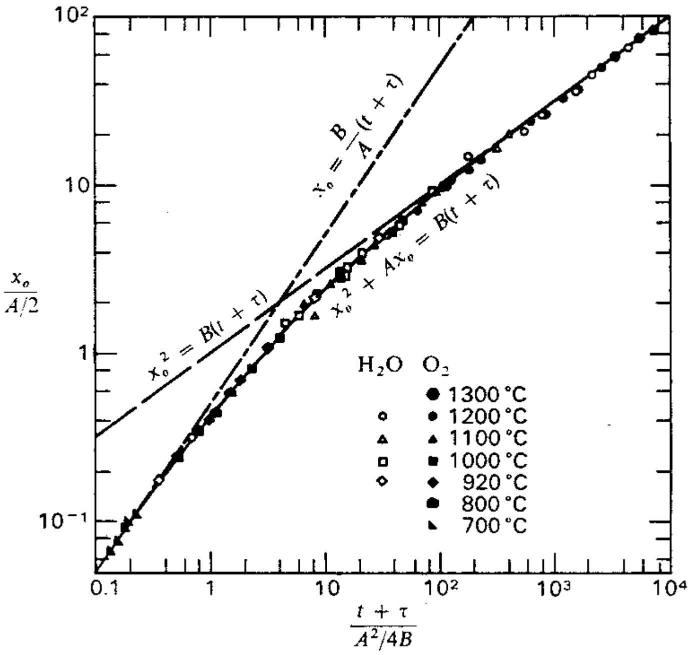


Fig. 31: Scaled thickness vs time for thermal oxidation (solid curve: Deal-Grove model; broken curves: linear and parabolic limits)

In practice, any conventional oxidation process used for integrated circuit fabrication will almost certainly be included within this temperature range.

Temperature Dependence of Oxidation Rate

Although, the Deal-Grove model is applicable over a wide range of temperatures, oxidation rate is strongly temperature dependent. As might be expected, this dependence has a classical *Arrhenius form*:

$$\kappa = \kappa_o \exp\left(-\frac{E_a}{kT}\right)$$

Here, κ can be identified as either the parabolic or linear rate constant. By definition, any reaction or process that is characterized by an Arrhenius form is said to be thermally activated and, accordingly, E_a, is identified as *activation energy*. To understand the precise meaning of E_a, one should think of any process (chemical reaction, diffusion, *etc.*) as a transition from some stable *reactant* state to a stable *product* state. In order, for both the reactant and the product state to be stable, the system must pass through some unstable "high energy" *transition state* (conventionally indicated by the symbol, ‡) during the process. Clearly, the transition state provides a "barrier" to free conversion of reactants into products. Such a "chemical" description of thermally activated processes can be represented pictorially as follows:

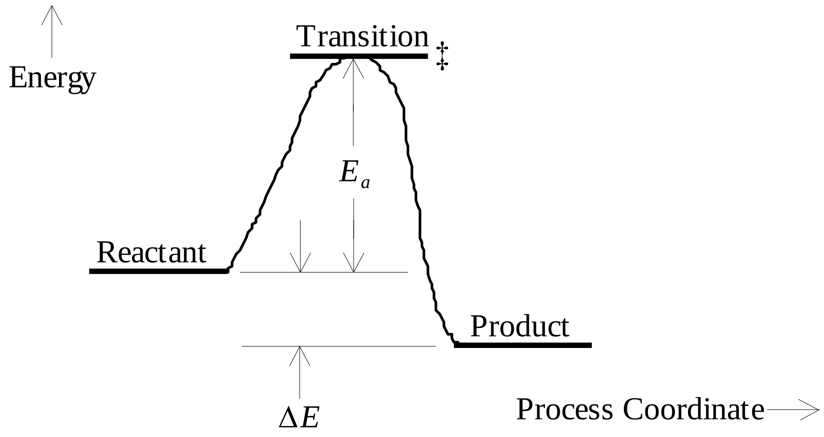


Fig. 32: Energetic relationship of transition, reactant, and product states

The horizontal dimension is defined as *process coordinate*, which is just a symbolic representation of the aggregate dynamics of the process. The vertical dimension represents thermodynamic internal energy. Thus, ΔE is the thermodynamic internal energy change for the overall process and E_a is the energy change taken between the reactant state and the transition state. (Thus, E_a can be thought of as a kind of formation energy for the transition state.) Clearly, because the energy of the transition state is higher than either the reactant or product states, it forms an *energy barrier* for the process and reactant and product states tend to be stable once they are formed. However, if thermal fluctuations randomly generate some of the transition state from the reactant state, then the product state is easily formed. Of course, the reverse process can also occur.

Digressing briefly, the case for which the product state has a lower internal energy than the reactant state is called an *exoenergetic* process since energy is released to the environment during the process. This situation is illustrated by the preceding figure. Conversely, if the product state has a higher internal energy than the reactant state, energy must be absorbed and the process is called *endoenergetic*. Obviously, the reverse of an exoenergetic process must be endoenergetic and vice-versa. (The terms exoenergetic and endoenergetic are analogous to the more common terms, *exothermic* and *endothermic*, except that they refer specifically to internal energy, rather than enthalpy.) Obviously, in the endoenergetic case, the activation energy must be larger than the internal energy change, ΔE, since it must be the sum of the activation energy for the reverse exoenergetic process, E_a, and ΔE. Of course, thermally activated processes for which reactant and product states are of equal internal energy, *e.g.*, diffusion, can be called *aenergetic*. However, the activation energy for an aenergetic process does not necessarily vanish and, clearly, is the same in both "forward" and "reverse" directions.

It is found that oxidation of clean silicon is a thermally activated, exoenergetic, *i.e.*, exothermic, process. Therefore, Arrhenius forms can be expected to represent temperature dependence of the linear and parabolic rate constants satisfactorily.

However, before providing specific values for activation energies and pre-exponential factors, it is important to note that oxidation rate is also found to depend on the orientation of the wafer surface, that is to say, that one finds that oxidation rates differ on [100] and [111] surfaces. Various models have been formulated to explain orientation dependence, however, in all of these it is attributed to differences in surface atom concentration and specific activation energy (derived from steric effects, *etc.*) For substrates commonly used in integrated circuit fabrication, one invariably finds that [111] wafers oxidize faster than [100] wafers under the same conditions. Since orientation is a property of the substrate only and does not affect the structure of the oxide once it is grown, *i.e.*, thermal oxide grown on [111] substrates is essentially identical to oxide grown on [100] substrates, one expects that orientation dependence enters the Deal-Grove model only through the specific rate constant, k_s. Therefore, it is to be expected that only the linear rate constant depends on orientation and that the parabolic rate constant is independent of orientation, as is, indeed, the case. Arrhenius forms for various process conditions and orientations appear in the following table:

Process	B/A for [100]	B/A for [111]	B
Dry Oxidation	$1.03(10^3)\, e^{-2.00/kT}$	$1.73(10^3)\, e^{-2.00/kT}$	$0.214\, e^{-1.23/kT}$
Steam Oxidation	$2.70(10^4)\, e^{-2.05/kT}$	$4.53(10^4)\, e^{-2.05/kT}$	$0.107\, e^{-0.79/kT}$

Note: Activation energies are in eV's, B/A is in μm/sec, B is in μm^2/sec

Table 2: Arrhenius forms for thermal oxidation rate constants

Clearly, steam oxidation is much faster than dry oxidation. Therefore, steam oxidation is advantageous for the growth of relatively thick oxide layers. These are typically field or isolation oxides, which surround devices and insulate the substrate from overlying wiring, *etc.* However, for oxides, usually thin, that are used as integral parts of devices, such as a gate insulator (or gate oxide), dry oxidation is generally used because it produces a higher quality Si/SiO$_2$ interface. The quality of this interface is critical for good electrical performance. (In the case of very thin oxides, this distinction breaks down. Indeed, the fabrication of ultrathin oxide layers is currently of great interest.)

Another process variable that is available to change oxidation rate is oxidant pressure. It is evident from the Deal-Grove model that B is proportional to oxidant concentration, C_G. Of course, C_G is just proportional to pressure (or partial pressure) through the usual gas laws. Therefore, both the linear and parabolic rate constants simply scale linearly with pressure. This provides several advantages. First of all, one can grow thick oxides much more rapidly at elevated pressure. However, "thermal budget", *i.e.*, the total exposure of the substrate to elevated temperature, rather than process time itself, is often a more important consideration in practical integrated circuit fabrication. Therefore, it is also advantageous to reduce the thermal budget without adversely affecting process time by lowering temperature and compensating the resulting lowered growth rate by increasing process pressure. An added benefit is that at lower temperature thermally activated defect generation is also reduced. Finally, for very thin oxides, the growth rate at normal atmospheric pressure may be too fast for adequate process control of oxide thickness. In this case, pressure (or partial pressure) can be reduced to a sub-atmospheric value to lower the growth rate and provide a more controllable process.

Deviations from the Deal-Grove Model

Before proceeding further, it is necessary to observe that there is one important deviation from the Deal-Grove model. In particular, the Deal-Grove model is unable to explain the kinetics associated with very thin oxide growth in dry oxygen. Specifically, a very rapid initial growth phase is observed. After this initial phase, the process follows the Deal-Grove model. Empirically, it is found that for dry oxide films of thickness greater than about 20 nm, the Deal-Grove model can be applied by assuming an initial fictitious oxide thickness of about this thickness, *i.e.*, one assumes that this initial film grows so rapidly that it can be taken as an initial condition for the Deal-Grove model. Of course, fabrication by dry oxidation of thin SiO_2 films having thickness on the order of 20 nm or less requires careful experimental characterization of growth kinetics in any initial growth regime. In contrast, a rapid initial growth phase is not observed for wet oxidation and the Deal-Grove model can be used to describe all stages of the process. (Wet oxidation is generally used for thick oxides anyway; however, recently there has been a renewed interest in using wet oxidation at very low temperature for fabrication of very thin oxide layers.)

The initial rapid growth phase in dry oxygen may be explained by observing that since no pre-existing oxide layer is present, the oxygen concentration is initially very high at the substrate surface. In this case, it is plausible that oxygen dissolves appreciably in the substrate itself to create a thin oxygen rich surface layer or *oxygen-diffused zone*. Of course, one expects the solubility of oxygen in silicon to be much less than in silicon dioxide, however, since little or no surface oxide is present, the concentration may still become significant. Thus, one can regard oxidation in the oxygen-diffused zone as more of a volume reaction than a surface reaction, *i.e.*, oxidation is occurring at an appreciable rate throughout the whole thickness of the oxygen-diffused zone. Therefore, since the whole thickness of the oxygen-diffused zone is rapidly converted to oxide, the apparent surface oxidation rate is “abnormally” high. Of course, once an initial oxide layer of sufficient thickness is formed, the concentration of oxygen at the interface falls and the oxygen-diffused zone disappears. Obviously, this must correspond to the onset of Deal-Grove kinetics. Alternatively, the initial rapid growth phase might be a consequence of deviation of the surface reaction kinetics from first order when the oxygen concentration is very high. (This is not necessarily inconsistent with the existence of an oxygen-diffused zone.)

For wet oxidation two observations can be made. First of all, water does not appear to dissolve or diffuse appreciably into silicon, hence, if any diffused zone is formed, it must be very thin. Secondly, the surface reaction in wet oxidation is inherently more rapid than in dry oxidation, which also serves to reduce the relative importance of any initial oxidation phase.

Oxidation Induced Defects

Under some conditions, thermal oxidation can produce *oxidation induced stacking faults* aligned with [111] planes. These stacking faults are typically extrinsic and, of course, are bounded by dislocations. Moreover, it is thought that oxidation induced stacking faults occur because thermal oxidation generates interstitial defects. Indeed, during normal oxidation, about one out of a thousand silicon atoms at the interface does not become incorporated into the growing oxide layer, but instead, diffuses back into the silicon lattice as an interstitial defect. Clearly, if the oxide growth rate is sufficiently high, these interstitials cannot come to equilibrium with vacancies, but rather “condense” as extrinsic stacking faults. (One recalls that an extrinsic stacking fault can be regarded as insertion of an extra plane of atoms.)

Within this context, it is found that stacking fault growth is thermally activated and is characterized by an Arrhenius form up to about ~1200°C. Above this temperature, stacking faults no longer grow larger, but shrink (a process called “retrogrowth”). This behavior can be understood if one recalls that the melting point of silicon is nominally 1414°C. Naturally, at a temperature near the melting point, one expects that lattice defects will be rapidly “annealed out” due to high atomic mobility. Furthermore, growth of oxidation induced stacking faults is found to be dependent on substrate orientation, majority carrier type, and defects. Accordingly, the growth rate of stacking faults is greater for [100] than for [111] substrates and stacking fault density is greater on *n*-type rather than *p*-type. Generally, the distribution of surface nucleated stacking fault lengths is very narrow. Furthermore, it is found that even for thick oxides, stacking fault growth is almost completely suppressed if oxidation temperature is reduced below 950°C. However, if it is desirable oxidize silicon substrates at higher temperature (perhaps to obtain a good Si/SiO_2 interface), subsequent high temperature annealing in an inert ambient can substantially reduce stacking faults.

Empirical observations indicate that for oxidation at the same temperature and time, the average length of oxidation induced stacking faults is greater for wet oxidation than for dry oxidation. This suggests that stacking fault length depends on oxidation rate as, indeed, is found to be the case. (However, if the same thickness of oxide is grown at a given temperature, wet oxidation will produce shorter stacking faults than dry oxidation since the oxidation time is much shorter.) Within this context, an empirical formula has been proposed to characterize dependence of stacking fault growth rate on oxidation rate as follows:

$$\frac{dl}{dt} = K_1 R_{ox}^n - K_2$$

Here, l is stacking fault length, R_{ox} is oxidation rate, and n, K_1, and K_2 are constant parameters. The exponent, n, is found to have a value of about 0.4. Therefore, dependence of stacking fault growth rate on oxidation rate is sub-linear and as indicated above, at some fixed temperature and oxide thickness, smaller stacking faults will be formed by a higher growth rate oxidation process. This suggests that high pressure oxidation should be useful for reduction of oxidation induced defects.

Kinetic Effects of Defects, Dopants, Chlorine, *etc.*

Defects in the silicon substrate are invariably associated with disruption in lattice bonding. Therefore, since lattice bonds are already broken, one expects that both wet and dry oxidation rates should be increased by the presence of defects. Although difficult to characterize quantitatively, this phenomenon is frequently observed. Furthermore, the effect of shallow level dopant concentration on oxidation rate is an important consideration for silicon integrated circuit fabrication. Indeed, it is well known that high dopant levels ($>10^{18}$ cm^{-3}) tend to accelerate both dry and wet thermal oxidation. The underlying cause of this is imperfectly understood; however, it may be a consequence of changes within the oxide structure itself due to the presence of dopants or enhanced defect generation within the substrate. Of course, any effect on oxide structure, hence, on oxidant diffusion coefficient, can be expected to change the parabolic rate constant. Accordingly, it has long been known that boron preferentially segregates into the oxide; therefore, since boron is trivalent rather than tetravalent, one may plausibly suppose that the oxide network structure should be weakened and oxidant diffusion enhanced. Conversely, defect generation within the substrate increases the surface reaction rate, but should not substantially affect oxide structure. Accordingly, the linear rate constant should be affected, but, not the parabolic rate. This evidently is the case for phosphorus and arsenic, which do not preferentially segregate into oxide. Clearly, oxidation of doped or defected silicon can be expected to deviate substantially from the Deal-Grove model. (Interaction between oxidation and dopant diffusion will be treated in more detail later.)

Chlorine (Cl_2) and chlorine containing species (*e.g.*, hydrogen chloride (HCl), trichloroethane (TCA), *etc.*) can be added to an oxidizing ambient with beneficial effects. Empirically, the presence of chlorine is found to improve the quality of the Si/SiO_2 interface. This may be partially a consequence of increased volatilization of metallic impurities. In addition, an increase in both linear and parabolic rate constants is also observed with the addition of chlorine or chlorine containing species to the oxidizing ambient. This may be due to two factors: First, enhanced vacancy generation at the Si/SiO_2 interface due to direct reaction of chlorine with silicon to produce volatile silicon chlorides, which allows more silicon migration to the surface or oxygen entrapment at the surface. Both of these effects should serve to enhance the rate. Second, chlorine incorporation into the oxide opens and expands the network structure resulting in an increase in the oxidant diffusion coefficient.

The Effect of Electric Field on the Semiconductor Surface

Before proceeding with detailed consideration of the Si/SiO$_2$ interface, capacitance-voltage analysis, device structures, *etc.*, it is necessary to consider the fundamental effect of an electric field on the surface of a semiconductor. To begin, one recalls, again, the physical meaning of the Fermi level (*i.e.*, Fermi energy) which, within Fermi-Dirac statistics is defined as the energy, E_F, for which electronic population probability is exactly one half. However, Fermi energy has further thermodynamic significance as the *free energy of mobile carriers*. Thus, for an amorphous or crystalline solid in thermal equilibrium, irrespective of whether it is a conductor, insulator, or semiconductor, E_F must have the same value everywhere within the solid. This is trivially obvious if the solid has uniform composition, however, at equilibrium, this condition must be satisfied even if the solid changes properties over some lateral dimension due to extrinsic doping or even gross changes in composition. Indeed, this behavior is fundamental to any understanding of contacts and junctions in semiconductors and metals.

Thus, in addition to band gap, crystallographic parameters, various thermodynamic equilibrium constants, *etc.*, another important basic material property is *work function*, which is defined as the energy (typically expressed in electron-volts) required to remove an electron from within a specific material to a state of rest in the vacuum, *i.e.*, to the *vacuum level*. Work functions are also commonly quoted in terms of equivalent electrical potential, *i.e.*, in volts. Of course, energy and potential are directly related by electrical charge, which for electrons and holes has magnitude of one fundamental unit, *i.e.*, nominally $\pm 1.602(10^{-19})$ C. Consequently, energy and electrical potential units are often carelessly treated as interchangeable; however, for consistency and to avoid confusion, work functions should be regarded as having units of energy. Physically, the work function is evidently a measure of aggregate electronic binding energy in the solid and is classically observed by applying a negative electrical potential, *i.e.*, a voltage, to some material (presumably having a reasonably clean surface) in a vacuum and measuring resulting current flow to an unbiased, *i.e.*, grounded, counter electrode also in the vacuum. (Alternatively, a positive potential can be applied to the counter electrode with the material grounded.) Accordingly, the observed current is not a slowly varying function of bias voltage, but exhibits a definite threshold which is characteristic of the work function of the material. (Indeed, this effect was first observed by Thomas Edison when he placed a second, biasable electrode inside a light bulb, which, as such, can be regarded as the world's first electronic device, the *vacuum diode*.) Typically, this threshold is found to be a few volts; hence, the work function is a few electron-volts. This is to be expected since this energy is of the same order of magnitude as electronic binding energies of electrons in atoms. Naturally, every solid material is, in principle, characterized by a different work function but, in practice, it may be very difficult to measure (as in the case of insulators).

Within this context, it is useful to consider the simple case of a surface contact between two dissimilar metals. In metals, the Fermi level generally does not fall in a band gap as it does in semiconductors or insulators. This situation can be realized physically two ways. In the most common case of a classical metal, the Fermi level falls within an occupied band, therefore, this band evidently can be only partially filled. Hence, electrons can easily make transitions between occupied and empty band states and are as a consequence, entirely delocalized, *i.e.*, mobile. In the case of a *semimetal*, an empty band overlaps a completely occupied band. Obviously, the Fermi level must fall at the top edge of the occupied band. Thus, a semimetal is analogous to a semiconductor having a zero or negative band gap. Again, electrons are delocalized and, hence, mobile. In any case, electronic states in metals are generally occupied right up to the Fermi level and, as a consequence, metals are characterized both by a large density of mobile, *i.e.*, *itinerant* or *conduction*, electrons (of the same order as the atomic density) and the absence of a band gap. (For completeness, it must be noted that mobile carriers in some metals appear to be positively charged and, hence, are more properly regarded as holes; however, this does not substantially change the basic picture of metallic conduction.)

As a "thought experiment", the following figure illustrates what happens when two dissimilar metals are brought into intimate contact. Here, E_{vac} is the energy of the vacuum level (conventionally taken to be zero), E_{F_1} is the Fermi level of "metal 1", and ϕ_1 is the associated work function. Similarly, E_{F_2} is the Fermi level of "metal 2" and ϕ_2 is the associated work function:

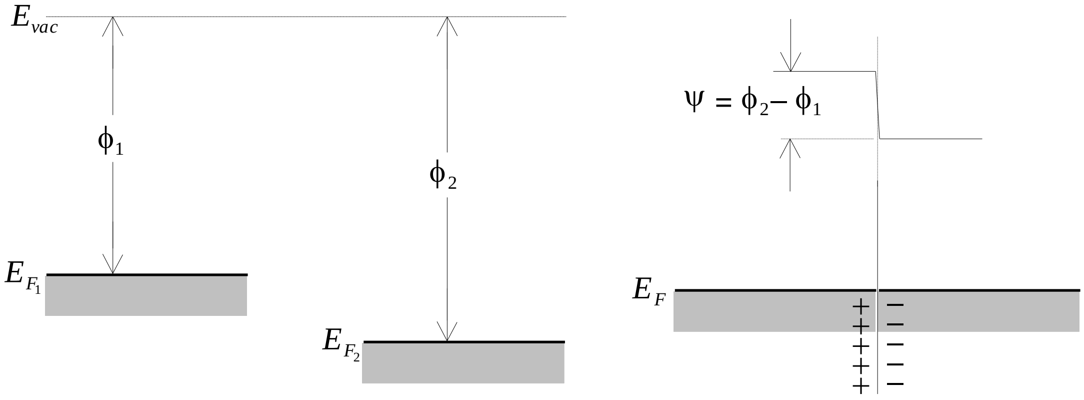


Fig. 33: Appearance of a contact potential at the interface of two dissimilar metals

As is indicated by the shaded regions in the figure, in a metal at low temperature, electrons occupy all available quasi-continuous band states up to the Fermi level. Of course, the Fermi levels, E_{F_1} and E_{F_2} must fall below E_{vac} since electrons are in bound states. If the two metals are widely separated in space (as indicated on the left side of the preceding figure), electronic equilibrium is not established and the Fermi levels do not necessarily coincide. Of course, in isolation, the Fermi levels differ from the vacuum level precisely by the work function; however, since the work functions are unequal, a free energy difference exists between electrons in metal 1 and metal 2. Therefore, if the two metals are brought into close proximity, *i.e.*, into contact, a spontaneous transient current flows. Physically, this current flow transfers electrons from the metal with the smaller work function (in this case, metal 1) to the metal with the larger work function (metal 2). Naturally, current continues to flow until thermodynamic equilibrium is established for mobile carriers. Of course, the condition of equilibrium requires that the Fermi level, E_F, must be the same in both metals. Thus, as a consequence of charge transfer, at equilibrium an electrical potential difference appears between the two metals.

This is called *contact potential* and simply corresponds to the quotient of difference of the work functions, ψ, and fundamental charge. Clearly, a contact potential exactly compensates initial disequilibrium arising from any difference in Fermi levels. However, due to the high density of mobile carriers within a metal, an electric field cannot exist within the bulk, hence, all of the contact potential difference must occur at the interface. (This is illustrated on the right side of the preceding figure.) Furthermore, since the width of the interface is, at most, on the order of a few atomic diameters, the required number of electrons transferred in order to produce a potential difference corresponding to ψ is, in fact, very small in comparison to the density of mobile carriers. As a result, for a metal-metal contact, it is extremely difficult if not impossible to measure the contact potential directly since any attempt to do so causes additional transient current flow which disturbs the equilibrium, *i.e.*, "shorts out" the potential. (In essence, any practical measuring equipment becomes part of the whole system and participates in the equilibrium.)

The situation for contact between a semiconductor and a metal is similar, with the added feature that, because of the existence of a band gap, moderate electric fields can exist within the bulk of a semiconductor. Typically, the work function of a metal, *e.g.*, aluminum, titanium, *etc.*, is less than that of an intrinsic semiconductor. For elemental materials, this supposition is easily rationalized from the periodic chart since metallic behavior is characterized by decreasing ionization potentials. (However, as will become evident subsequently, in the case of extrinsically doped semiconductors this situation may become inverted.) The following figure illustrates the case of a surface contact between a metal and intrinsic semiconductor, *viz.*, silicon:

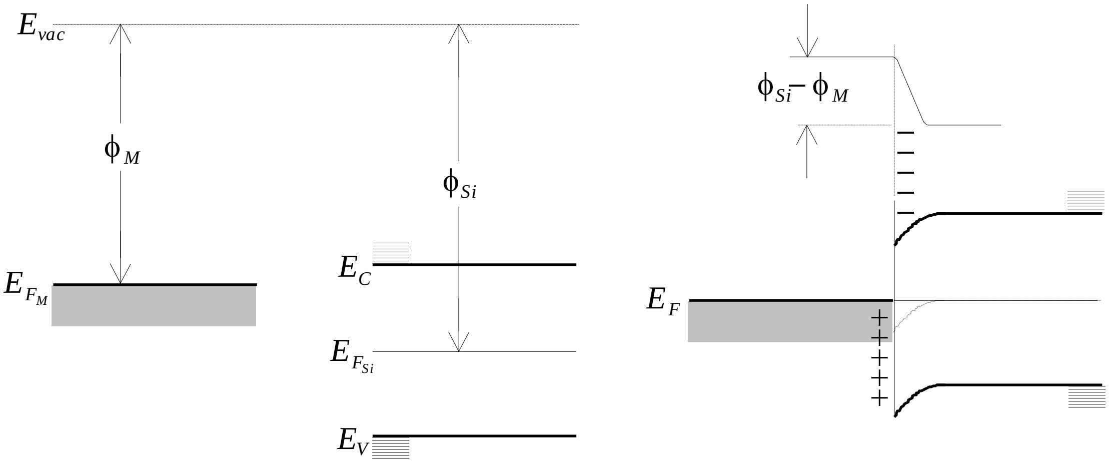


Fig. 34: Appearance of a contact potential at the interface of a metal and intrinsic semiconductor

Of course, E_{F_M} is the Fermi level of the metal and ϕ_M is the associated work function. Likewise, $E_{F_{Si}}$ is the Fermi level of the semiconductor and ϕ_{Si} is its work function. Just as in the case of two dissimilar metals, if one brings an intrinsic semiconductor and a metal into close proximity, the metal tends to lose electrons to the semiconductor simply because available energy states for electrons in the semiconductor are of lower energy, *i.e.*, because the Fermi level is lower in the semiconductor. Transient current flows until equilibrium is established and the Fermi level becomes the same in both the metal and the semiconductor. However, in contrast to the case of two metals, the transfer of electrons to the semiconductor results in an electric field that penetrates the surface and causes the band structure of the semiconductor to "bend". If, as has been assumed, the metal work function is smaller than the semiconductor work function, then the bands must bend "downward". This can be understood by observing that electrons transferred to the semiconductor from the metal occupy states in the conduction band. (Of course, carrier equilibrium implies that some electrons recombine with holes; however, this is only a small fraction of electrons transferred.) Since these excess electrons are supplied from an external source, holes do not appear in the valence band. Hence, the situation is very similar to the case arising from shallow donor levels and the surface of the semiconductor is no longer intrinsic, but becomes *n*-type. Accordingly, this implies that in proximity to the surface, the Fermi level must lie above the intrinsic level. However, the condition of equilibrium requires that the Fermi level is the same everywhere, thus, to satisfy this condition the energy of the bands themselves must shift. Of course, the electrostatic field arising from charge separation due to net electron transfer from the metal to the semiconductor tends to confine excess electrons in the semiconductor in close proximity to the surface. Therefore, at a distance sufficiently deep in the semiconductor bulk, the semiconductor remains intrinsic. This implies, as is shown on the right side of the figure that the bands bend smoothly in the region that the electric field penetrates into the semiconductor.

One can further consider the effect of extrinsic doping. In the case of a *p*-type semiconductor, the Fermi level in the bulk is shifted below the intrinsic level, E_i, by an amount corresponding to the Fermi potential, φ_F. Nevertheless, the effect of a metal contact remains unchanged and the bands bend as illustrated for lightly and heavily doped *p*-type substrates in the following figure:

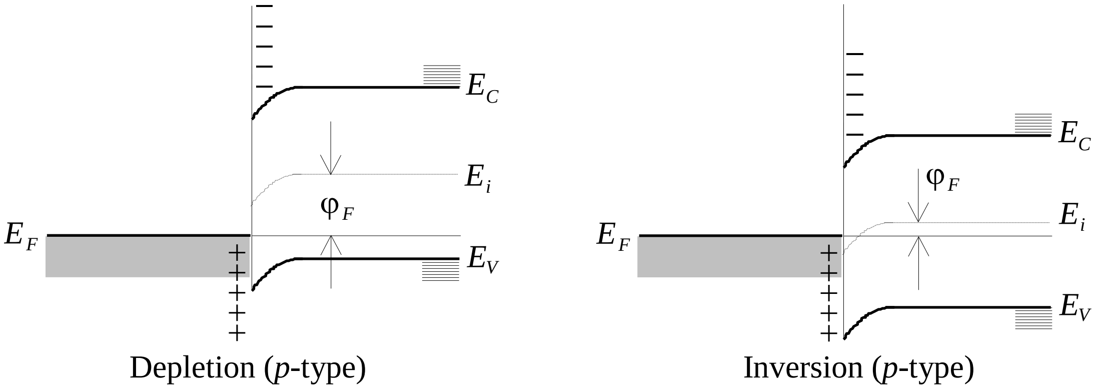
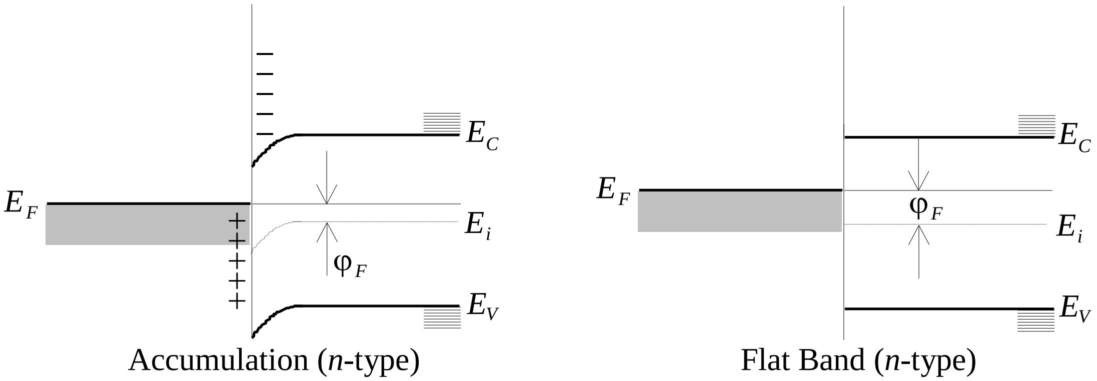


Fig. 35: Depletion and inversion of *p*-type extrinsic semiconductor due to a metal contact

(In these diagrams and those that follow, φ_F is to be understood as a measure of the energy magnitude, $|E_F-E_i|$.) Again, when the metal and semiconductor initially come into intimate contact, electrons are lost from the metal. Of course, in comparison to an intrinsic semiconductor, the effective work function for a *p*-type semiconductor is larger and downward band bending must be greater. However, in contrast to the intrinsic case, instead of occupying empty conduction band states, electrons recombine with excess holes in the valence band. If the net acceptor dopant concentration is sufficiently large (*i.e.*, heavy doping), an excess concentration of holes still remains at the surface once equilibrium is established. However, near the surface, the distance between the valence band edge and the Fermi level is increased, hence, the density of majority carriers (*i.e.*, holes) is decreased and, thus, effective acceptor doping is reduced. This condition is called *depletion* and corresponds to the left diagram in the preceding figure. Naturally, there is still net charge separation due to electron transfer and a corresponding surface electrostatic field.

Obviously, the intrinsic level bends along with the bands, therefore, in a doped semiconductor the possibility exists that the bands may bend far enough so that the intrinsic level and the Fermi level actually intersect. This typically happens if doping is light and in the case of a *p*-type semiconductor majority carriers change from holes to electrons such that at the surface the semiconductor changes from *p*-type to *n*-type. This condition is called *inversion* and is illustrated by the right diagram in the preceding figure. Inversion in a *p*-type semiconductor can be understood if one observes that upon initial contact, as before, electrons from the metal recombine with holes in the valence band. However, if acceptor doping is sufficiently low, all extrinsic holes are effectively consumed, *i.e.*, recombine, before equilibrium is established. At this point, the intrinsic and Fermi levels are exactly equal and the semiconductor surface becomes effectively intrinsic even though acceptor impurities are present. Consequently, additional electrons transferred from the metal must occupy empty conduction band states and, thus, the semiconductor surface becomes *n*-type. Of course, both depleted and inverted regions are necessarily confined to a layer of semiconductor near the surface associated with band bending.

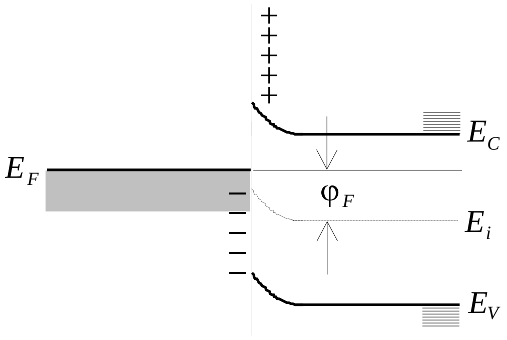
In contrast to the case of a *p*-type semiconductor, in an *n*-type semiconductor the Fermi level in the bulk is shifted above the intrinsic level, E_i, again, by an amount corresponding to the Fermi potential, φ_F. Therefore, in comparison to the intrinsic case, the effective work function for an *n*-type semiconductor is smaller. This allows several possibilities. First of all, if φ_F is not too large, then the effective work function of the semiconductor is still larger than for the metal and the bands still bend downward as in the *p*-type case, but the degree of bending is necessarily less. Again, electrons are transferred from the metal to the semiconductor; however, in this case rather than recombining with holes in the valence band, electrons immediately occupy the conduction band. Thus, the concentration of majority carriers (*i.e.*, electrons) is increased at the surface and effective doping is enhanced. This condition is called *accumulation* and is illustrated by the left diagram in the following figure. Moreover, it can happen that for a judicious choice of donor impurity concentration, the Fermi potential exactly offsets the work function difference between the metal and semiconductor. In this case, no charge transfer is required to establish equilibrium and the bands are not bent. This is called the *flat band* condition and is illustrated by the right diagram in the following figure:



Accumulation (*n*-type) Flat Band (*n*-type)

Fig. 36: Flat band condition and accumulation of *n*-type extrinsic semiconductor due to a metal contact

However, if the Fermi potential for an *n*-type semiconductor is sufficiently large, the effective semiconductor work function may actually be smaller than the metal work function. In this case, electrons are transferred from the semiconductor to the metal instead of from the metal to the semiconductor and bands are bent “upward” instead of downward. Obviously, the majority carrier concentration is reduced at the semiconductor surface; hence, this again corresponds to depletion. However, in this case depletion is a consequence of the direct loss of electrons from the semiconductor to the metal rather than recombination of excess electrons from the metal with holes in the semiconductor. In both cases, a space charge region appears in the surface layer of the semiconductor due to the presence of “uncovered” ionized impurity atoms. (Clearly, these charges are not mobile since they are fixed in the silicon lattice.) Of course, depletion of *n*-type silicon due to a metal contact is also represented by a band diagram, thus:



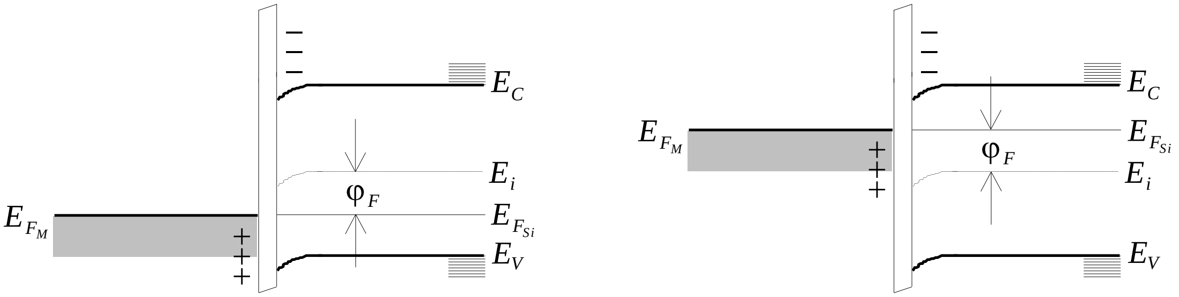
Depletion (*n*-type)

Fig. 37: Depletion of *n*-type extrinsic semiconductor due to a metal contact

All of these cases illustrate that penetration of the semiconductor surface by an electric field alters effective doping at the surface from the net extrinsic doping of the bulk. These are all examples of *field effect*. So far, these fields have been regarded as arising from charge transfer induced by work function differences; however, it should be apparent that field effects must arise any time an electric field penetrates a semiconductor crystal irrespective of the source of the field.

The MOS Capacitor

In the previous case of metal-semiconductor contacts, one could consider application of an external potential difference between the metal and semiconductor to attempt to either reduce or enhance field effects caused by work function differences alone. This is actually possible to a limited extent; however, if the magnitude of the applied potential becomes too great, a large amount of current will flow resulting in difficulties. (By no means, however, should metal-semiconductor contacts be considered useless, indeed, they form components of many useful solid-state electronic devices.) To remedy this situation, suppose that instead of just bringing metal and semiconductor into contact, that a thin layer of insulator is inserted between them. Typically, this is thermally grown silicon dioxide (but, other insulators can also be used with similar effect). The resulting structure is called a *metal-oxide-semiconductor capacitor* or just an MOS capacitor. Again, to establish equilibrium, charge transfer occurs in such a way as to bring the Fermi levels into correspondence. However, since the insulating layer is present, a sheet of charge of opposite polarity builds up at each insulator interface. The situation is similar to the simple metal-semiconductor case, except that part of the electric field penetrates the insulator. This condition is illustrated below:



Depletion (*p*-type)

Accumulation (*n*-type)

Fig. 38: Band diagrams for an unbiased MOS capacitor

Here, the oxide layer is represented by the parallelogram separating metal and semiconductor bands. (The width of the parallelogram corresponds to oxide thickness and the slope of the top and bottom sides to internal electric field strength.) For *p*-type silicon having a metal contact on the silicon dioxide layer, just as for a direct metal-semiconductor contact, the bands generally bend downward and the semiconductor becomes depleted or even inverted. Of course, the degree of depletion (or inversion) is dependent on substrate doping, but the amount of band bending is not as great as in the case of a metal-semiconductor contact since some of the electric field penetrates the oxide layer, *i.e.*, the potential due to the work function difference (contact potential) is distributed over both the oxide layer and a surface layer in the semiconductor. The situation remains similar for *n*-type silicon. Depending on the metal work function and substrate doping, the bands may bend upward or downward (corresponding respectively to accumulation or depletion and inversion), but again, not as much as in the case of a direct metal contact.

Within this context, since thermal oxide is an excellent insulator, it becomes possible to apply a much greater potential difference, *i.e.*, bias voltage, between the metal and semiconductor without an associated large current flow than is possible in the case of a simple metal-semiconductor contact. In passing, it should be noted that application of an external potential difference causes the Fermi levels in the metal and semiconductor in both simple direct metal contacts and MOS structures to become offset. The magnitude of the offset is, of course, exactly the energy gained by an electron "falling through" the applied potential difference. Therefore, the sign of the Fermi level offset is opposite the sign of the external potential difference since electrostatic potential is conventionally defined with respect to positive rather than negative charge. Thus, a negative potential difference between the metal and semiconductor causes a positive energy offset between the metal and semiconductor Fermi levels. Conversely, a positive potential difference between the metal and semiconductor causes a negative energy offset between the metal and semiconductor Fermi levels. (This situation is also characteristic of contact potentials in the absence of any external bias.)

Capacitance-Voltage Response of an MOS Capacitor

Application of an external bias voltage to an MOS capacitor, allows the surface layer of either *n*-type or *p*-type semiconductor to be accumulated, depleted, or inverted at will. Moreover, in contrast to the case of an ideal parallel plate capacitor, which has a constant capacitance for any value of applied voltage, the capacitance of an MOS structure changes as a function of the condition of the semiconductor surface, *i.e.*, the capacitance is different depending on whether the semiconductor surface is accumulated, depleted, or inverted. In addition, the capacitance-voltage (CV) response of an MOS capacitor depends both on the characteristics of the oxide layer and the semiconductor substrate. (As will become evident subsequently, the CV response of an MOS capacitor provides very useful information regarding the behavior and quality of a Si/SiO$_2$ interface.)

In practice, observation of CV response generally requires "sweeping" bias voltage and simultaneously measuring capacitance. In the *p*-type case, accumulation occurs if the applied bias is sufficiently negative. Clearly, this implies that a large concentration of majority carriers (*i.e.*, holes) is attracted to the semiconductor surface by the negative bias voltage. Accordingly, if the bias voltage is made more positive, hole concentration at the semiconductor surface must decrease. Moreover, a fixed bias can be found such that the surface hole concentration becomes just equal to the hole concentration due to bulk acceptor doping. In this case, the bands are flat (which formally specifies *flat band voltage*). Therefore, any further increase in applied bias must cause the semiconductor surface to become depleted. Obviously, increasing the bias voltage still further will cause the surface to become first intrinsic and then inverted. At very high positive bias voltage, the concentration of electrons in the inversion layer becomes large. Of course, the behavior of *n*-type semiconductor can be expected to be inverted with respect to bias voltage, but otherwise completely analogous. Naturally, accumulation for *n*-type semiconductor will occur at sufficiently high positive bias voltage (instead of negative). As bias is reduced, accumulation is followed by the flat band condition and then depletion as the majority carrier concentration (*i.e.*, electrons) falls below the bulk electron concentration due to net donor doping. Further reduction of bias voltage to negative values results first in an intrinsic surface and then in an inversion layer as holes become majority carriers. These six applied bias conditions of an MOS capacitor are illustrated in the following figure:

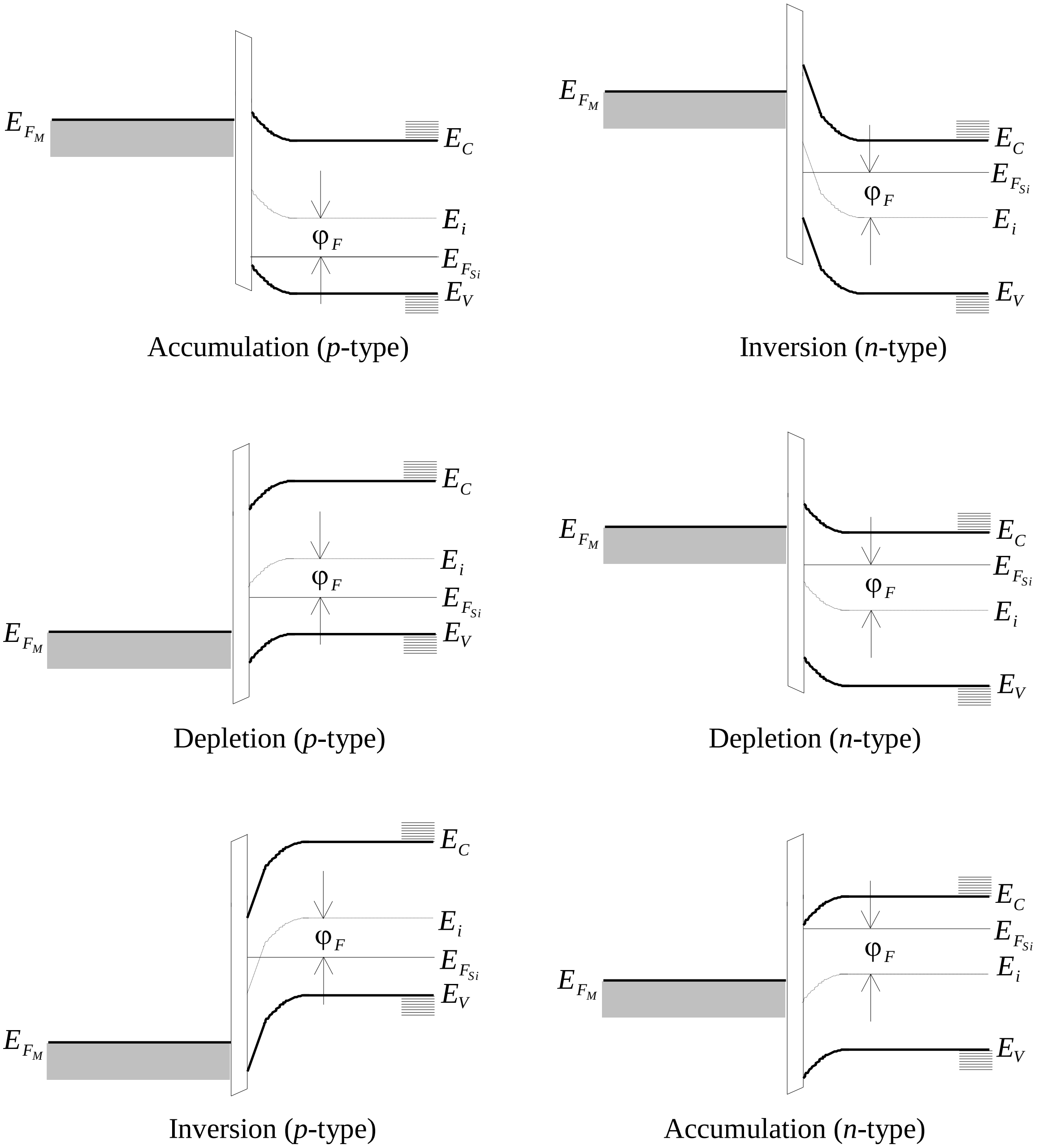


Fig. 39: Behavior of an MOS structure for both *p* and *n*-type substrates under various conditions of bias

Clearly, bias voltage increases from negative to more positive values as one sequentially considers band diagrams from top to bottom and as expected, *p* and *n*-type semiconductors exhibit complementary behavior.

Considering the case of an MOS capacitor fabricated on *p*-type silicon, if one initially applies a negative voltage to the "gate" (*i.e.*, the metal contact), it is clear that majority carriers, *i.e.*, holes, will be attracted to the surface forming an *accumulation layer*, *i.e.*, the bands are strongly bent upward, increasing the effective doping of the surface relative to the bulk. In this case, a sheet of negative charge will appear at the metal-oxide interface and a sheet of positive charge will appear at the silicon-silicon oxide interface. Clearly, this is similar to the simple case of a charged parallel plate capacitor. Thus, the capacitance per unit area, called C_{ox}, just corresponds to the elementary formula:

$$C_{ox} = \frac{\varepsilon_{ox}}{x_o}$$

Here, ε_{ox} is the dielectric constant of silicon dioxide (0.34 pF/cm) and x_o is nominal oxide thickness. If the magnitude of the negative voltage is reduced, *i.e.*, bias voltage is increased toward zero, then curvature of the bands decreases and, likewise, the degree of accumulation decreases. Clearly, at some point, the bands will not be bent and the surface will not be accumulated (or depleted), *i.e.*, the silicon is neutral everywhere. This is, of course, merely the flat band condition, which for this case occurs at a small negative bias since work functions for metals, *e.g.*, aluminum, typically are smaller than for *p*-type silicon. Under this condition, the external potential exactly compensates the intrinsic potential arising from the work function difference. As voltage is further increased through zero to positive values, the bands bend the opposite direction, *i.e.*, downward. (Clearly, it has already been established that the bands are bent downward at zero bias.) In the case of a positive bias, but not too positive, one can easily visualize that majority carriers will be repelled from the surface, thus, creating a region devoid of mobile carriers. This is called the *depletion region*. Consequently, to satisfy charge conservation requirements of the system, the depletion region must have a net negative space charge. Of course, this space charge is provided by uncovered ionized dopant atoms, *i.e.*, in this case most likely negatively charged boron atoms. Since these negative charges are not mobile, the conductivity of the depletion region is much lower than the bulk semiconductor. It follows then that the capacitance per unit area of an MOS structure in depletion, is the series combination of the oxide capacitance per unit area as defined previously and a depletion layer capacitance per unit area, C_d:

$$C_d = \frac{\varepsilon_s}{x_d}$$

Here, ε_s is the semiconductor dielectric constant (1.04 pF/cm for Si) and x_d, is defined as *depletion width*. Clearly, x_d is not a physical thickness of a thin film in the same sense as x_o, but is an electrical equivalent thickness. (This will be treated in more detail subsequently.) Naturally, if positive bias is increased still further, the bands continue to bend downward until the intrinsic level and the Fermi level are just equal at the surface. This defines the onset of inversion since as positive bias is increased further electrons accumulate at the semiconductor surface to form an *inversion layer*. Thus, as asserted previously, the semiconductor surface becomes effectively *n*-type. An additional increase of bias voltage results in greater accumulation of electrons in the inversion layer without substantial depletion of the underlying semiconductor. Therefore, once an inversion layer is fully formed, the semiconductor does not deplete any further and x_d reaches a maximum value, $x_d^{\max}$. Obviously, maximum depletion capacitance per unit area, C_s, is defined as follows:

$$C_s = \frac{\varepsilon_s}{x_d^{\max}}$$

Therefore, to summarize, in accumulation the capacitance per unit area of an MOS structure, C, is just C_{ox}. At the onset of depletion, C is the series combination of C_{ox} and C_d and, hence, must be less than C_{ox}. Therefore, as the surface becomes more depleted C decreases until at the onset of inversion it approaches a minimum value equal (or nearly equal) to the series combination of C_{ox} and C_s. Under equilibrium conditions, when a full inversion layer is formed, C just returns to C_{ox} since the inversion layer once again acts as one electrode of a simple parallel plate capacitor, thus, effectively removing any effect of C_s from observation. Of course, the behavior of an MOS capacitor fabricated on an *n*-type substrate is analogous except the sign sense of the bias voltage must be reversed, *i.e.*, accumulation for an *n*-type substrate occurs at positive bias voltage and inversion occurs at negative bias voltage.

The Surface Potential and Field

Within this context, the electrostatic potential and field within a semiconductor surface layer can be readily analyzed quantitatively. For simplicity, the semiconductor will be considered as a uniformly doped, semi-infinite silicon crystal in thermal equilibrium. Accordingly, the semiconductor surface is defined by a plane located at $x=0$ perpendicular to the x axis. Hence, the bulk of the semiconductor is characterized by positive values of x. Naturally, an electrostatic surface potential, φ, is defined directly from Poisson's equation:

$$\frac{d^2\varphi}{dx^2} = -\frac{\rho(x)}{\varepsilon_s}$$

Here, $\rho(x)$ is net charge density. Clearly, $\rho(x)$ can be written as just the sum of contributions from mobile carriers and ionized dopant atoms:

$$\rho(x) = q(p(x) - n(x) + N_D - N_A)$$

Of course, in contrast to the simple case of bulk semiconductor, carrier concentrations near the surface are explicit functions of x; however N_D and N_A remain independent of x due to the assumption of uniform doping. It follows from fundamental definitions previously given that:

$$\frac{kT}{2}\ln\frac{p(x)}{n(x)} = E_i(x) - E_F$$

Hence, one makes use of carrier equilibrium to construct expressions for carrier concentrations as a function of *x*:

$$kT\ln\frac{p(x)}{n_i} = E_i(x) - E_F \qquad ; \qquad -kT\ln\frac{n(x)}{n_i} = E_i(x) - E_F$$

Moreover, these expressions can be rearranged to obtain expressions for carrier concentrations in terms of the intrinsic level:

$$p(x) = n_i \exp\left(-\frac{E_F - E_i(x)}{kT}\right) \qquad ; \qquad n(x) = n_i \exp\left(-\frac{E_i(x) - E_F}{kT}\right)$$

Clearly, hole concentration appears on the left and electron concentration on the right. Of course, the intrinsic level is also a function of depth, *x*, since it bends with the bands. As usual, for light to moderate doping these expressions can be regarded as explicitly of Maxwellian form. Obviously, it follows that:

$$p(x) - n(x) = 2n_i \sinh\left(\frac{E_i(x) - E_F}{kT}\right)$$

Naturally, deep within the bulk, *i.e.*, in the limit that $x \to \infty$, the condition of charge neutrality must be satisfied, hence:

$$p(\infty) - n(\infty) = -(N_D - N_A) = 2n_i \sinh\left(\frac{E_i - E_F}{kT}\right)$$

Here, E_i denotes the intrinsic level in the bulk, *i.e.*, beyond the region of band bending. Thus, these results can be readily combined to give an explicit expression for the charge density:

$$\rho(x) = 2n_i q\left(\sinh\left(\frac{E_i(x) - E_F}{kT}\right) - \sinh\left(\frac{E_i - E_F}{kT}\right)\right)$$

Of course, it follows from the Maxwellian forms that the electrostatic surface potential, φ, is fundamentally related to the intrinsic level as follows:

$$\varphi = \frac{E_F - E_i(x)}{q}$$

Naturally, the “zero” of the potential may be always chosen arbitrarily. Clearly, the preceding expression implies that φ vanishes if the intrinsic level and the Fermi level are exactly equal; hence, Poisson’s equation takes the form:

$$\frac{d^2\varphi}{dx^2} = \frac{2n_i q}{\varepsilon_s}\left(\sinh\left(\frac{q\varphi}{kT}\right) - \sinh\left(\frac{q\varphi_\infty}{kT}\right)\right)$$

Here, φ_∞ is obviously identified as $(E_F - E_i)/q$. For convenience, dimensionless thermal potentials, ϕ and ϕ_∞ are defined as $q\varphi/kT$ and $q\varphi_\infty/kT$, respectively:

$$\frac{d^2\phi}{dx^2} = \frac{2q^2 n_i}{\varepsilon_s kT}(\sinh\phi - \sinh\phi_\infty)$$

Upon inspection, a characteristic length, λ_i, called the *intrinsic Debye length* can be identified with the following combination of constants:

$$\lambda_i = \sqrt{\frac{\varepsilon_s kT}{2q^2 n_i}}$$

It remains to formally integrate Poisson’s equation.

One begins by defining an integrating factor by means of the following elementary identity:

$$\frac{d}{dx}\left(\frac{d\phi}{dx}\right)^2 = 2\frac{d\phi}{dx}\left(\frac{d^2\phi}{dx^2}\right)$$

At this point, one multiplies Poisson’s equation by $2(d\phi/dx)$ to obtain:

$$\frac{d}{dx}\left(\frac{d\phi}{dx}\right)^2 = \left(\frac{\sinh\phi - \sinh\phi_\infty}{\lambda_i^2}\right)\frac{d\phi}{dx}$$

If one defines ϕ_s and $d\phi_s/dx$, respectively, as ϕ and $d\phi/dx$ characteristic of the semiconductor surface, *i.e.*, $x=0$, then this expression can be cast into definite integral form as follows:

$$\int_{d\phi_s/dx}^{0} d\left(\frac{d\phi}{dx}\right)^2 = \int_{\phi_s}^{\phi_\infty} d\phi\left(\frac{\sinh\phi - \sinh\phi_\infty}{\lambda_i^2}\right)$$

Here, integration is taken from the surface into the bulk of the semiconductor. Obviously, $d\phi/dx$ must be proportional to the electric field, hence:

$$\frac{d\phi}{dx} = \frac{qE}{kT}$$

The integrals can be simplified by substitution of elementary forms, therefore one obtains the result:

$$-\left(\frac{qE_s}{kT}\right)^2 = \frac{2}{\lambda_i^2}(\cosh\phi_\infty - \cosh\phi_s - (\phi_\infty - \phi_s)\sinh\phi_\infty)$$

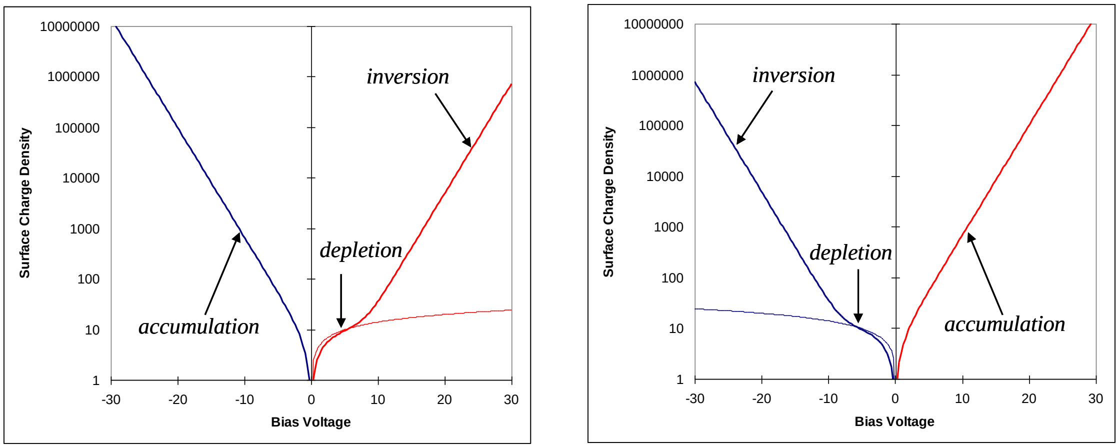
Accordingly, at the surface of the semiconductor the electric field is determined in terms of the dimensionless potential at the surface and in the bulk:

$$E_s = \mathrm{sgn}(\phi_\infty - \phi_s)\frac{kT}{q\lambda_i}\sqrt{2((\phi_\infty - \phi_s)\sinh\phi_\infty - \cosh\phi_\infty + \cosh\phi_s)}$$

Here, "sgn" denotes the signum function, which merely specifies the sign as positive (bands bend downward) or negative (bands bend upward). Accordingly, total charge per unit area, Q_s, is easily obtained using Gauss' Law

$$Q_s = \varepsilon_s E_s = \mathrm{sgn}(\phi_\infty - \phi_s)\frac{\varepsilon_s kT}{q\lambda_i}\sqrt{2((\phi_\infty - \phi_s)\sinh\phi_\infty - \cosh\phi_\infty + \cosh\phi_s)}$$

Surface charge density versus bias voltage for both *p* and *n*-type semiconductor is shown in the following figure. (In practice, bias voltage can be regarded as surface potential, ψ_s, plus some constant offset.)



p-type *n*-type

Fig. 40: Surface potential versus bias voltage (red indicates negative charge, blue positive charge)

If one identifies ψ_s as the the potential difference between the surface and the bulk, $\varphi_s - \varphi_\infty$, one finds that for a *p*-type substrate Q_s declines rapidly as ψ_s, *i.e.*, potential bias, is increased from accumulation to the flat band condition (at which point, by definition Q_s vanishes). As ψ_s is increased beyond the flat band condition, *i.e.*, to positive values, the magnitude of Q_s increases slowly as the semiconductor surface is depleted. Obviously, inversion must begin when ψ_s is equal to φ_F, the Fermi potential in the bulk. (One notes that φ_F is equal to $-\varphi_\infty$ for a *p*-type substrate.) Furthermore, this implies that intrinsic and actual Fermi levels are exactly equal at the surface. However, one finds that the magnitude of Q_s still increases only slowly until ψ_s is approximately twice the bulk Fermi potential. This defines the condition of *weak inversion*. For values of ψ_s more positive than $2\varphi_F$, the magnitude of Q_s increases rapidly, thus, defining *strong inversion*. Clearly, strong inversion in a *p*-type semiconductor occurs if the Fermi level is as far above the intrinsic level at the surface as it is below the intrinsic level in the bulk. (Of course, an *n*-type substrate exhibits analogous behavior with corresponding sign senses reversed.)

The Depletion Approximation

Of particular importance in CV analysis is knowledge of the surface field when the semiconductor is in depletion. In this case, carrier concentrations are small and can be ignored within the depletion region. Furthermore, the depletion region is formally treated as a layer of definite width, x_d. This is called the *depletion approximation*. Therefore, Poisson's equation takes the simplified approximate form:

$$\frac{d^2\varphi}{dx^2} = -\frac{q}{\varepsilon_s}(N_D(x) - N_A(x))$$

Moreover, since carrier concentrations are ignored, the assumption of uniform doping can be suspended and doping densities treated as general functions of *x*. Within this context, it is useful to define the potential at the depletion region edge as, φ_d, and the potential difference across the depletion region, ψ, as $\varphi - \varphi_d$. Thus, Poisson's equation is easily recast in terms of ψ as follows:

$$\frac{d^2\psi}{dx^2} = -\frac{q}{\varepsilon_s}(N_D(x) - N_A(x))$$

One readily integrates this expression from the depletion edge toward the semiconductor surface:

$$\int_0^{-E(x)} d\left(\frac{d\psi}{dx}\right) = \frac{q}{\varepsilon_s}\int_x^{x_d} dx'(N_D(x') - N_A(x'))$$

Of course, the electric field, $E(x)$, vanishes at the edge of the depletion region. Furthermore, $E(x)$ is just the negative of the potential gradient; hence, one can integrate a second time as follows:

$$\psi(x) = -\int_{x_d}^{x} dx' E(x') = -\frac{q}{\varepsilon_s}\int_{x}^{x_d} dx' \int_{x'}^{x_d} dx''(N_D(x'') - N_A(x''))$$

If the order of the integrals over x' and x'' is formally inverted, then the integral over x' is readily simplified to give:

$$\psi(x) = -\frac{q}{\varepsilon_s}\int_{x}^{x_d} dx'' \int_{x}^{x''} dx'(N_D(x'') - N_A(x'')) = \frac{q}{\varepsilon_s}\int_{x}^{x_d} dx''(x - x'')(N_D(x'') - N_A(x''))$$

Obviously, this expression can be simplified no further without explicit knowledge of the net impurity distribution.

If, for simplicity, one again assumes uniform doping, the previous expression can easily be recast as follows:

$$\psi(x) = -\frac{q}{\varepsilon_s}(N_D - N_A)\int_{x}^{x_d} dx''(x'' - x) = \frac{q}{2\varepsilon_s}(N_A - N_D)(x_d - x)^2$$

Thus, within this approximation, band bending has a parabolic shape as a function of distance from the surface of the semiconductor. Furthermore, ψ_s is obtained by formally setting x equal to zero:

$$\psi_s = \frac{q}{2\varepsilon_s}(N_A - N_D)x_d^2$$

One can construct an approximate expression for the corresponding electric field by directly differentiating:

$$E(x) = \frac{q}{\varepsilon_s}(N_A - N_D)(x_d - x)$$

Of course, the field at the surface is, again, obtained if x is set equal to zero:

$$E_s = \frac{q}{\varepsilon_s}(N_A - N_D)x_d$$

Likewise, the total charge per unit area in the depletion region is determined from Gauss' Law:

$$Q_s = \varepsilon_s E_s = q(N_A - N_D)x_d$$

Alternatively, this expression follows just from elementary consideration of net impurity concentration. Furthermore, one finds that capacitance per unit area of the depletion region is just Q_s/ψ_s as might be expected. For completeness, it is useful to observe that within the depletion approximation the magnitude of Q_s as a function of ψ_s is of simple square root form:

$$|Q_s| = \sqrt{2\varepsilon_s \psi_s q(N_A - N_D)}$$

Of course, for *n*-type and *p*-type semiconductors Q_s is negative and positive, respectively. Within this context, this formula is plotted in the previous figure and corresponds to the "parabolic" curve coinciding with depletion. Clearly, as might be expected, the depletion approximation is no longer applicable upon the onset of inversion.

Maximum Depletion Width

The maximum depletion width expected under equilibrium conditions for a *p*-type substrate can be determined by inverting the previous approximate expression for ψ_s and replacing ψ_s by $2\varphi_F$, *i.e.*, corresponding to the onset of strong inversion:

$$x_d^{\max} = 2\sqrt{\frac{\varepsilon_s \varphi_F}{q(N_A - N_D)}}$$

Naturally, this formula can be generalized to both *n* and *p*-type substrates by the simple expedient of replacing $N_A - N_D$ with absolute value:

$$x_d^{\max} = 2\sqrt{\frac{\varepsilon_s \varphi_F}{q|N_A - N_D|}}$$

Of course, it follows from fundamental considerations that:

$$kT \ln\left(\frac{|N_A - N_D|}{n_i}\right) = |E_i - E_F|$$

Hence, the Fermi potential is related to impurity concentrations as follows:

$$\varphi_F = \frac{kT}{q} \ln\left(\frac{|N_A - N_D|}{n_i}\right)$$

One substitutes this formula into the expression for the maximum depletion width, $x_d^{\max}$, to obtain the desired result:

$$x_d^{\max} = \sqrt{\frac{4\varepsilon_s kT}{q^2|N_A - N_D|}\ln\left(\frac{|N_A - N_D|}{n_i}\right)}$$

It follows immediately that maximum capacitance of the depletion layer, C_s, corresponds to the formula:

$$C_s = \frac{q}{2}\sqrt{\frac{\varepsilon_s\,|\,N_A - N_D\,|}{kT\ln(|\,N_A - N_D\,|\,/\,n_i)}}$$

In practice, both C_{ox} and C_s are measured experimentally. These values can then be used to determine oxide thickness and substrate doping. However, these quantities are better measured by different methods and this is not the primary use of CV analysis, which is characterization of the electrical quality of an oxide film and the Si/SiO$_2$ interface.

Capacitance-Voltage Measurement

Capacitance-voltage (CV) measurements are conventionally made using a dedicated test fixture situated within a light excluding enclosure to prevent measurement errors due to extraneous photo-generated currents. Typically, an unpatterned oxide layer is fabricated on a high quality silicon substrate, which is then metallized with a thin film of aluminum. The thin film is patterned to form MOS capacitors, either at the time of deposition by use of a shadow mask or by means of conventional photolithography and chemical etching. Alternatively, MOS capacitors can be fabricated by depositing, heavily doping, and patterning a polysilicon thin film instead of aluminum. The advantage of the use of aluminum is that the deposition and patterning are very convenient; however, heavily doped polysilicon most closely approximates a finished *transistor* structure. Usually, electrical connection to the completed MOS capacitor is made to the topside aluminum or polysilicon contact by means of a thin tungsten probe and to the backside of the wafer by use of a conductive "chuck" which allows a partial vacuum to be drawn under the wafer, *i.e.*, a "vacuum chuck", thus facilitating electrical contact. Although not absolutely necessary for capacitance measurements, it is generally advantageous to remove any pre-existing insulating layers from the back of the wafer. (Indeed, it is usual for a layer of oxide to be grown on the back as well as the front of a wafer during thermal oxidation.) This is easily done by etching in hydrofluoric acid, hydrogen fluoride vapor, or some other method. If backside contact is found to be especially critical, the whole back of the wafer optionally may be metallized. Furthermore, in practice, it is usual for the backside contact, *i.e.*, the silicon substrate, to be held at ground and the frontside contact, *i.e.*, aluminum or polysilicon, to be biased at some potential. In general, it is found that useful information is obtained from CV measurements made for two different conditions, *viz.*, quasistatic and high frequency. (Typically, this data is used to construct a "CV plot" which graphically displays MOS capacitance (or capacitance per unit area) as a function of bias voltage.)

Quasistatic Conditions

As might be expected, quasistatic conditions correspond closely to equilibrium. Conventionally, a quasistatic CV measurement is made by sweeping bias voltage applied to an MOS capacitor so that the surface of the semiconductor changes from inversion to depletion and then to accumulation. During this procedure, displacement current is measured as a function of time. (Obviously, displacement current is just the transient charging current of the capacitor.) Ideally, there is no true conduction current flowing through the oxide layer. Even so, a small amount of "leakage current" invariably flows through any "real" oxide layer. Of course, for quasistatic measurements to be useful the oxide must not be too "leaky", otherwise, conduction current will be added to displacement current and measurements will be inaccurate. (Generally, for a high quality oxide this current is negligible and can be ignored.) Within this context, modern, computer-controlled CV analysis equipment can automatically correct for leakage (if it is not too large or erratic), thus, extending the usefulness of CV measurements to less than perfect oxide layers. Obviously, integration of charging current over time merely results in a measurement of the charge stored in the MOS capacitor; hence, capacitance as a function of voltage is determined by elementary identification of the product of capacitance and voltage as stored charge, *i.e.*, $CV=Q$.

Clearly, if the surface of the semiconductor is either in accumulation or inversion, a layer of charged mobile carriers is present directly beneath the oxide. (For notational convenience, an italicized C denotes capacitance per unit area and a non-italicized C denotes absolute capacitance.) Thus, the measured capacitance, $\mathrm{C}_{\mathrm{max}}$, is just the capacitance of the oxide layer alone (equal to $C_{ox}A$ such that, by definition, A is the area of the gate, *i.e.*, the frontside contact). In contrast, if the semiconductor is depleted, there is no layer of mobile carriers present underneath the oxide, *i.e.*, at the Si/SiO_2 interface. However, mobile carriers are present underneath the depletion region. Therefore, in depletion, the measured capacitance consists of the series combination of the oxide capacitance and the capacitance of the depletion layer. Of course, this combined capacitance must be less than $\mathrm{C}_{\mathrm{max}}$. Accordingly, as the voltage is swept from inversion to accumulation, during depletion the capacitance decreases from $\mathrm{C}_{\mathrm{max}}$ to a minimum, $\mathrm{C}_{\mathrm{min}}$, corresponding to maximum depletion layer width and then rises again as the semiconductor becomes accumulated. (By definition, $\mathrm{C}_{\mathrm{min}}$ is capacitor area, A, times minimum capacitance per unit area, C_{min}.) This essential behavior illustrated in the following figure:

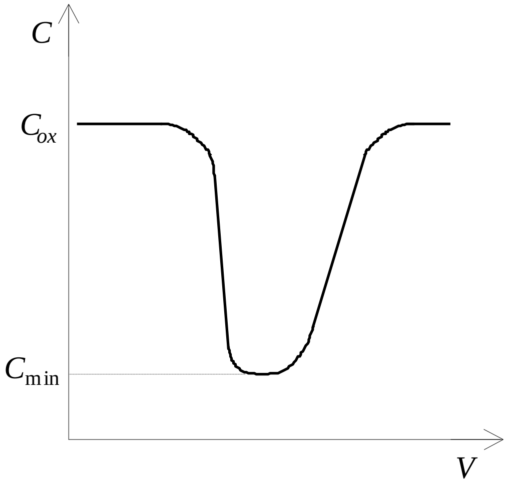


Fig. 41: Idealized quasistatic CV plot (*p*-type substrate; accumulation at left)

Clearly, C_{min} is related to C_{ox} and C_s as follows:

$$\frac{1}{C_{\mathrm{min}}}=\frac{1}{C_{ox}}+\frac{1}{C_s}=\frac{x_o}{\varepsilon_{ox}}+\frac{x_d^{\mathrm{max}}}{\varepsilon_s}$$

As defined previously, x_o and ε_{ox} are, respectively, thickness and dielectric constant of SiO_2. Likewise, x_d^{max} and ε_s are maximum thickness and dielectric constant of the depletion layer. Of course, electrical charges are still present within the depletion region due to ionized impurity atoms; however, these charges are fixed and do not move in response to an applied bias (at low temperatures). Consequently, these fixed charges can make no contribution to displacement current and do not result in a contribution to measured capacitance.

High Frequency Conditions

In quasistatic measurements, capacitance is measured directly by integrating charging current. However, CV measurements can also be made by superimposing a small sinusoidally oscillating (AC) signal on the voltage sweep and measuring the corresponding impedance directly as a function of bias voltage. (This requires the use of a high precision impedance meter.) In this case, capacitance measured under conditions of accumulation and depletion can be expected to be the same as observed in quasistatic measurements, *i.e.*, conditions still remain near equilibrium. However, if the frequency of the AC signal is sufficiently high, capacitance measured under a condition of inversion is not the same as in the quasistatic case. The explanation for this is quite simple and is a direct consequence of non-equilibrium behavior of the inversion layer. Physically, any inversion layer must be formed from minority carriers generated in the depletion region and swept to the surface by the electric field. (Of course, minority carriers may be also generated in the bulk and diffuse into the depletion region.) Equilibrium conditions imply that there is sufficient time (by definition) for the inversion layer carrier concentration to respond to any changes in applied field. However, if the material quality of the silicon is good, carrier generation-recombination processes occur very slowly, with a time constant on the order of milliseconds. Therefore, for an applied AC voltage in the megahertz range, the response of the inversion layer is simply too slow to “follow” the signal and similar to ionized dopant impurity atoms, the inversion layer appears fixed with respect to the AC component of the bias. (Of course, the inversion layer does respond to the primary voltage sweep.) This behavior is shown in the following figure:

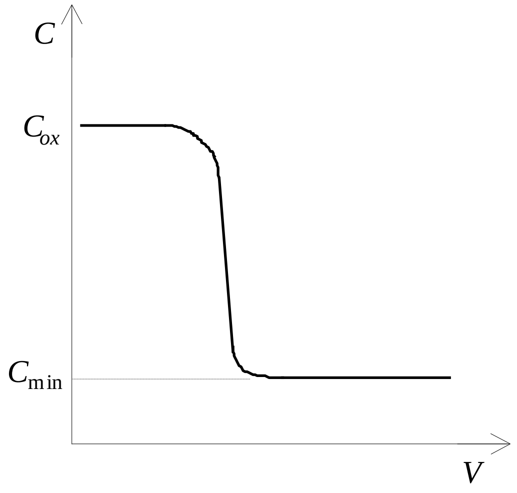


Fig. 42: Idealized high frequency CV plot (*p*-type substrate; accumulation at left)

Therefore, for high frequency conditions, the capacitance per unit area measured in inversion is the series combination of oxide capacitance per unit area and capacitance per unit area of the depletion region. Furthermore, since, the depletion width reaches a maximum value, the combined capacitance per unit area saturates at $C_{\min}$.

Interpretation of Ideal Capacitance-Voltage Measurements

In principle, measurements of capacitance versus voltage can be made either by sweeping the applied voltage from accumulation to inversion (– to + voltage for *p*-type; + to – for *n*-type) or inversion to accumulation (+ to – voltage for *p*-type; – to + for *n*-type). For quasistatic measurements the direction of the sweep makes essentially no difference in the form of the CV plot, again, since the MOS capacitor remains nearly at equilibrium. However, for high frequency measurements, the form of the CV plot can differ in the inversion region depending on the direction and rate of the voltage sweep. As asserted previously, this behavior is due to the kinetics of minority carrier generation. Clearly, if the time constant for generation-recombination processes is long (indicative of a high quality substrate), then the inversion layer may not fully form during a fast voltage sweep from accumulation to inversion. Thus, the depletion region may grow larger than one would otherwise expect for equilibrium conditions. This phenomenon is called *deep depletion* and is illustrated below:

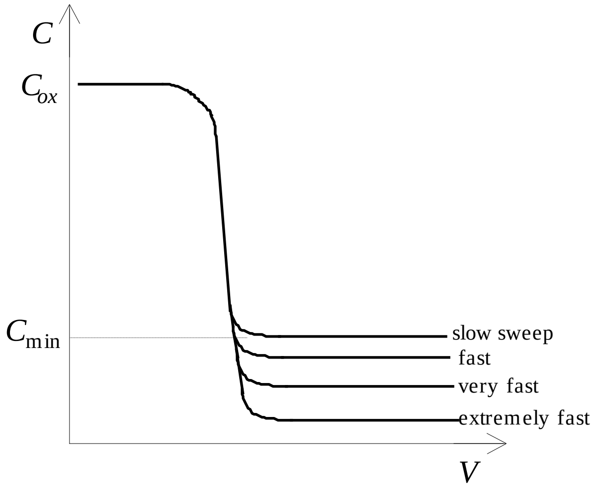


Fig. 43: Deep depletion for sweep from accumulation to inversion (*p*-type substrate; accumulation at left)

Although deep depletion is generally a nuisance in conventional CV analysis and, as such, to be avoided, it is possible to take advantage of this effect to determine the time constant of carrier generation-recombination processes, *i.e.*, minority carrier lifetime, for the substrate. As asserted previously, generation-recombination is slow for a high quality substrate; hence, minority carrier lifetime is long. However, minority carrier lifetime is significantly shortened (on the order of microseconds) in defected or contaminated semiconductor due enhancement generation-recombination processes. Moreover, if minority carrier lifetime is short, not only is deep depletion absent, but, even at high frequency one may observe the onset of equilibrium behavior, *i.e.*, MOS capacitance increases in inversion. Accordingly, minority carrier lifetime can be estimated by determining the dependence of inversion capacitance on sweep rate. The derivative of this function extrapolated back to equilibrium conditions is proportional to minority carrier lifetime.

In practice, to avoid deep depletion, high frequency CV measurements are nearly always made by sweeping the applied bias voltage from inversion to accumulation. Furthermore, prior to application of the voltage sweep, the substrate is fully inverted by appropriate biasing and illumination of the surface. (Illumination enhances the formation of an inversion layer by providing photo-generated minority carriers.) Of course, the voltage sweep itself should be made without illumination. One finds that the form of associated CV plots essentially depends on oxide thickness and the substrate doping (among other factors). Obviously, ideal MOS capacitance per unit area, *C*, is constructed by the usual series combination:

$$\frac{1}{C}=\frac{1}{C_{ox}}+\frac{1}{C_d}=\frac{x_o}{\varepsilon_{ox}}+\frac{x_d}{\varepsilon_s}$$

(In accumulation, x_d vanishes, in inversion, x_d is equal to $x_d^{\max}$, and in depletion x_d varies smoothly between these two values.) Clearly, if substrate doping is constant, an increase in oxide thickness reduces the capacitance of the oxide layer; thus, the total MOS capacitance is also reduced. Furthermore, the position of the depletion region as a function of bias moves to higher values of voltage magnitude since it is really the electric field magnitude at the interface which determines surface conditions (*i.e.*, a higher voltage magnitude must be applied across a thicker oxide to obtain the same electric field magnitude at the Si/SiO$_2$ interface.) Conversely, if oxide thickness is constant, a change in substrate doping causes a corresponding change in depletion layer capacitance. This is easy to understand because increasing (or decreasing) substrate doping causes the maximum width of the depletion layer to be reduced (or increased). Therefore, for a specified oxide thickness, C_{ox} remains constant and $C_{\min}$ changes as a function of substrate doping; however, the position of the onset of depletion as a function of voltage remains unaffected.

The Effect of Fixed Charges

Ideally, there is essentially no uncovered charge within a high quality thermal oxide layer. However, in reality, it is possible for charges to become more or less permanently trapped within the oxide layer or at the Si/SiO$_2$ interface. Furthermore, in many cases these charges behave as if they are fixed. Therefore, in analogy to ionized impurity atoms in the substrate, such fixed oxide charges do not participate in nor are changed by charging the MOS capacitor. However, the existence of extraneous fixed charges does cause an overall shift in the position of the depletion region with respect to applied bias voltage. This is easily understood in elementary terms, since if one solves Poisson’s equation, one finds that a layer of fixed charge inside the oxide layer just results in a constant potential offset. This is most conveniently analyzed by considering the capacitance and voltage for which the semiconductor is in a flat band condition.

Accordingly, one begins by considering surface differential capacitance per unit area in the semiconductor substrate subject to the assumption that the MOS capacitor is essentially at equilibrium. Obviously, this limits consideration to low frequency or quasistatic conditions. Clearly, for an arbitrary surface potential, *i.e.*, arbitrary bias voltage, it follows from the fundamental definition of depletion layer capacitance that:

$$|\delta Q_s| = C_d |\delta\varphi_s|$$

Here, δQ_s and $\delta\varphi_s$ denote differential changes in surface charge density and potential. (Absolute values appear so that this analysis can be applied to either *n* or *p*-type substrates.) Clearly, the partial derivative of surface charge density with respect to dimensionless surface potential is directly obtained from the explicit expression for Q_s constructed previously:

$$\left|\frac{\partial Q_s}{\partial \phi_s}\right| = \left(\frac{\varepsilon_s kT}{q\lambda_i}\right) \frac{|\sinh\phi_s - \sinh\phi_\infty|}{\sqrt{2((\phi_\infty - \phi_s)\sinh\phi_\infty - \cosh\phi_\infty + \cosh\phi_s)}}$$

Clearly, one determines capacitance per unit area directly by multiplying the preceding expression by *q*/*kT*, hence:

$$C_d = \left(\frac{\varepsilon_s}{\lambda_i}\right) \frac{|\sinh\phi_s - \sinh\phi_\infty|}{\sqrt{2((\phi_\infty - \phi_s)\sinh\phi_\infty - \cosh\phi_\infty + \cosh\phi_s)}}$$

It is convenient to define a dimensionless band bending potential, ϑ_s, as $\phi_s - \phi_\infty$, hence:

$$C_d = \left(\frac{\varepsilon_s}{\lambda_i}\right) \frac{|\sinh(\vartheta_s + \phi_\infty) - \sinh\phi_\infty|}{\sqrt{2(\cosh(\vartheta_s + \phi_\infty) - \cosh\phi_\infty - \vartheta_s \sinh\phi_\infty)}}$$

Of course, the flat band condition occurs when ϑ_s exactly vanishes; hence, one must consider the limit:

$$C_d^{FB} = \left(\frac{\varepsilon_s}{\lambda_i}\right) \lim_{\vartheta_s \to 0} \frac{|\sinh(\vartheta_s + \phi_\infty) - \sinh\phi_\infty|}{\sqrt{2(\cosh(\vartheta_s + \phi_\infty) - \cosh\phi_\infty - \vartheta_s \sinh\phi_\infty)}}$$

Here, C_d^{FB} is defined as the value of C_d at flat band conditions. To determine the limit, one formally substitutes exponentials for hyperbolic functions

$$C_d^{FB} = \left(\frac{\varepsilon_s}{\lambda_i}\right) \lim_{\vartheta_s \to 0} \frac{\left|e^{\vartheta_s + \phi_\infty} - e^{-\vartheta_s - \phi_\infty} - 2\sinh\phi_\infty\right|}{2\sqrt{e^{\vartheta_s + \phi_\infty} + e^{-\vartheta_s - \phi_\infty} - 2\cosh\phi_\infty - 2\vartheta_s \sinh\phi_\infty}}$$

Next, one makes use of the Taylor series of the exponential function such that:

$$C_d^{FB} = \left(\frac{\varepsilon_s}{\lambda_i}\right) \lim_{\vartheta_s \to 0} \frac{\left|(1+\vartheta_s+\cdots)e^{\phi_\infty} - (1-\vartheta_s+\cdots)e^{-\phi_\infty} - 2\sinh\phi_\infty\right|}{2\sqrt{(1+\vartheta_s+\tfrac{1}{2}\vartheta_s^2+\cdots)e^{\phi_\infty} + (1-\vartheta_s+\tfrac{1}{2}\vartheta_s^2+\cdots)e^{-\phi_\infty} - 2\cosh\phi_\infty - 2\vartheta_s\sinh\phi_\infty}}$$

Obviously, high order terms are negligible in the series expansions; hence, only low order terms have been retained, *viz.*, linear in the numerator and quadratic in the denominator. Thus, it follows immediately from the elementary relationship of exponential and hyperbolic functions that:

$$C_d^{FB} = \left(\frac{\varepsilon_s}{\lambda_i}\right) \lim_{\vartheta_s \to 0} \frac{|2\vartheta_s \cosh\phi_\infty|}{2\sqrt{\vartheta_s^2 \cosh\phi_\infty}} = \left(\frac{\varepsilon_s}{\lambda_i}\right)\sqrt{\cosh\phi_\infty}$$

Likewise, it follows immediately from the definition of Fermi potentials that:

$$\cosh\phi_\infty = \frac{p(\infty)+n(\infty)}{2n_i}$$

Furthermore, in an extrinsically doped semiconductor, majority carriers predominate, hence, one can approximate the preceding expression explicitly in terms of net doping density:

$$\cosh\phi_\infty = \frac{|N_A - N_D|}{2n_i}$$

Thus, the capacitance of the depletion layer per unit area subject to flat band conditions has the form:

$$C_d^{FB} = \frac{\varepsilon_s}{\lambda_i}\sqrt{\frac{|N_A - N_D|}{2n_i}} = \frac{\varepsilon_s}{\lambda_D}$$

Accordingly, *extrinsic Debye length* is defined as follows:

$$\lambda_D = \sqrt{\frac{\varepsilon_s kT}{q^2|N_A - N_D|}}$$

Physically, Debye length is a characteristic distance that some external electric field can penetrate a neutral semiconductor surface without substantially perturbing the semiconductor away from neutrality. (This external field can arise either from a contact potential or an applied bias.) Thus, Debye length can be interpreted as an "effective shielding distance". In many physical systems multiple Debye lengths can be identified. (Indeed, for a semiconductor, both intrinsic and extrinsic Debye lengths are commonly defined.) Moreover, it is evident that the shortest Debye length can be expected to dominate (provided that it is substantially shorter than the nearest alternative). Indeed, for a typical extrinsic semiconductor at room temperature, $\lambda_D << \lambda_i$. Thus, it is reasonable that C_d^{FB} is simply ε_s/λ_D. Obviously, flat band capacitance per unit area, C_{FB}, is readily constructed, thus:

$$\frac{1}{C_{FB}} = \frac{1}{C_{ox}} + \frac{1}{C_d^{FB}} = \frac{x_o}{\varepsilon_{ox}} + \frac{\lambda_D}{\varepsilon_s}$$

As expected, C_{FB} appears as a series combination of oxide and depletion layer capacitances.

If substrate doping is known, C_{FB} is easily determined from measured values of C_{ox} and $C_{\min}$ obtained from a high frequency CV plot. First of all, one observes that maximum depletion width and extrinsic Debye length are related directly as follows:

$$x_d^{\max} = 2\lambda_D \sqrt{\ln\left(\frac{|N_A - N_D|}{n_i}\right)}$$

Thus, C_{FB} takes the form:

$$\frac{1}{C_{FB}} = \frac{1}{C_{ox}} + \frac{1}{2C_s\sqrt{\ln(|N_A - N_D|/n_i)}}$$

Naturally, C_s is directly related to $C_{\min}$ as follows:

$$\frac{1}{C_s} = \frac{1}{C_{\min}} - \frac{1}{C_{ox}} = \frac{1}{C_{ox}}\left(\frac{C_{ox}}{C_{\min}} - 1\right)$$

Therefore, it immediately follows that:

$$\frac{1}{C_{FB}} = \frac{1}{C_{ox}}\left(1 + \frac{(C_{ox}/C_{\min}) - 1}{2\sqrt{\ln(|N_A - N_D|/n_i)}}\right)$$

Of course, this expression is easily recast in terms of absolute flat band capacitance, C_{FB}, and measured values of absolute capacitances, $\mathrm{C}_{\max}$ and $\mathrm{C}_{\min}$, again, obtained from a high frequency CV plot:

$$\mathrm{C}_{FB} = \frac{\mathrm{C}_{\max}}{1 + \dfrac{(\mathrm{C}_{\max}/\mathrm{C}_{\min}) - 1}{2\sqrt{\ln(|N_A - N_D|/n_i)}}}$$

This formula frequently appears in practical guides to CV measurements. Clearly, C_{FB} is the series combination of capacitances, $C_{ox}A$ and $C_d^{FB}A$. Furthermore, once C_{FB} has been determined, flat band voltage, V_{FB}, is easily specified from experimental data (*i.e.*, one just "reads off" the voltage that corresponds to C_{FB} from the CV plot).

Ideally, the flat band voltage should correspond just to the effective work function difference between the metal contact and the doped silicon substrate. However, if uncovered charges are present within the oxide layer, then the flat band voltage corresponds to the expression:

$$V_{FB} = \phi_M - \phi_{Si} - \frac{1}{\varepsilon_{ox} A} \int_0^{x_o} dx' x' \rho_{ox}(x')$$

Here, $\rho_{ox}(x)$ is charge density in the oxide and is regarded as a function of depth from the oxide surface. (This expression is easily derived from Poisson's equation.) Clearly, the nearer a charge is the Si/SiO$_2$ interface, the larger is its contribution to flat band voltage. In many cases, it is reasonable to assume that all of the oxide charge is located at or very near the Si/SiO$_2$ interface. In this case, one defines fixed charge per unit area, Q_f; hence, the above expression takes the form:

$$V_{FB} = \phi_M - \phi_{Si} - \frac{x_o Q_f}{\varepsilon_s} = \phi_M - \phi_{Si} - \frac{Q_f}{C_{ox}}$$

Almost invariably, Q_f is found to be positive, in which case V_{FB} is more negative than the work function difference. Hence, if positive fixed charges are present near the Si/SiO$_2$ interface, then the CV plot is translated to more negative values of bias voltage, but functional form remains undistorted. This translation from the ideal flat band voltage corresponding to the simple work function difference, to flat band voltage experimentally observed in a CV plot is called *flat band shift*, ΔV_{FB}, and is illustrated below:

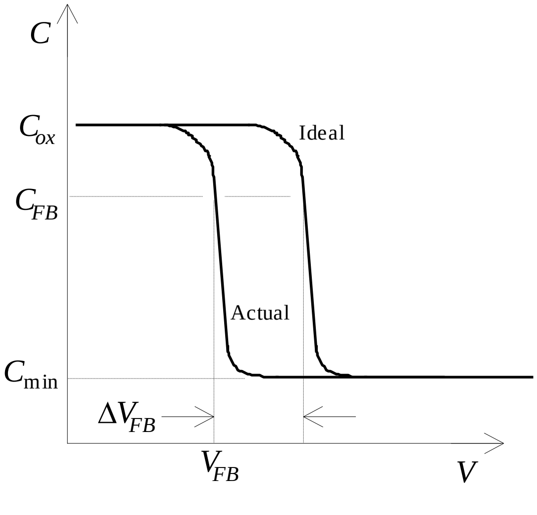


Fig. 44: Flat band shift due to oxide fixed charge (*p*-type substrate; accumulation at left)

Clearly, the actual CV plot is shifted by ΔV_{FB} toward more negative bias voltage in comparison to the ideal CV plot. The flat band shift, ΔV_{FB}, has a magnitude of Q_f/C_{ox} and, therefore, is a direct measure of fixed charge density.

Physically, fixed charges arise from a variety of sources. They may be "grown in" due to the oxidation process itself or they may be the result of contamination or radiation damage. Of particular interest are fixed charges that arise as a result of mobile ion contamination (K^+, Na^+, Li^+, *etc.*). Mobile ion contamination may be differentiated from other types of fixed charge by use of a *bias stress test*. The technique is as follows: First, an initial high frequency CV measurement is made. Second, a large positive bias is applied to the gate and simultaneously the substrate is heated to ~200°C. This treatment should serve to sweep any mobile ions dissolved in the oxide to the Si/SiO_2 interface. Third, the bias is removed, the substrate cooled, and a second CV measurement is made. If the second CV plot is translated to more negative values in comparison to the initial CV plot, but there is no significant distortion in the shape, then the presence of mobile ion contamination within the oxide layer is indicated. This result can be confirmed by applying a large negative bias to the gate and, again, heating the substrate. A CV plot obtained after this treatment should show some evidence of "recovery", *i.e.*, translation of the CV plot back to more positive values.

The Effect of Interface Traps and Fast Surface States

Other important phenomena readily observable in CV measurements are the existence of interface traps and fast surface states at the Si/SiO_2 interface. Theoretically, such interface quantum states must arise naturally due to the broken symmetry caused by termination of the crystal lattice at a surface. Furthermore, these states may have characteristic energies that lie within the band gap of the bulk crystal, thus, reducing minority carrier lifetime. Physically, interface traps and fast surface states can be regarded as arising from unsatisfied, *i.e.*, dangling, bonds which appear upon transition from single crystal silicon to amorphous silicon dioxide. Generally, it is found that the density and energy distribution of interface traps and fast surface states is very dependent on processes and/or materials. Semantically, the difference between interface traps and fast surface states is somewhat vague; however a useful distinction can be made by consideration of time constants. Within this context, a fast surface state has a charging time constant sufficiently short so that its response is observable in a high frequency CV measurement. This causes distortion and "stretch-out" of the corresponding CV plot. In contrast, an interface trap has a longer charging time constant so that distortion of high frequency CV measurements is minimal. Of course, both fast surface states and interface traps are observable in quasistatic CV measurements. Obviously, if significant distortion is visible in a measured high frequency CV plot, then the quality of the Si/SiO_2 interface is very poor and no further analysis is required. Such distortion is illustrated in the following figure

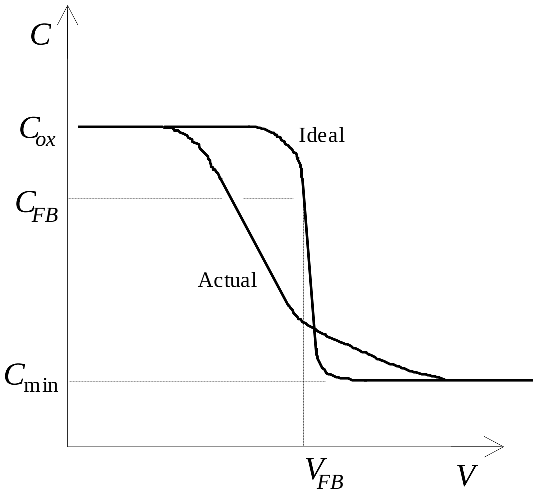


Fig. 45: Distortion due to fast surface states (*p*-type substrate; accumulation at left)

As a practical matter, it has long been known that fast surface states can be "passivated" by means of low temperature heat treatment (400-450°C) in a hydrogen containing ambient, *e.g.*, in "forming gas". (Presumably, the hydrogen diffuses to the Si/SiO$_2$ interface and satisfies dangling bonds.)

However, if high frequency CV measurements exhibit minimal distortion, then measurement of interface trap capacitance per unit area, C_{it}, is particularly useful as a quantitative measure of the quality of a Si/SiO$_2$ interface. In practice, C_{it} can be determined by a direct comparison of quasistatic and high frequency CV data taken for the same capacitor. Obviously, the essence of the method relies on the charging kinetics of interface traps. Again, in a quasistatic measurement, interface traps result in a contribution to capacitor charging current, but since time constants of interface traps are relatively long, the effect of interface traps is absent (or at least greatly reduced) in a high frequency measurement. Therefore, the difference between quasistatic and high frequency CV measurements under conditions of depletion allows direct determination of C_{it}:

$$C_{it} = C_d^{LF} - C_d^{HF} = \left(\frac{1}{C_{LF}} - \frac{1}{C_{ox}}\right)^{-1} - \left(\frac{1}{C_{HF}} - \frac{1}{C_{ox}}\right)^{-1}$$

Here, C_d^{LF} and C_d^{HF} are depletion layer capacitances per unit area obtained respectively by observation of quasistatic and high frequency MOS capacitances per unit area, C_{LF} and C_{HF}. Clearly, in the preceding formulation, observed depletion layer capacitance is regarded as the series combination of interface trap capacitance and "true" depletion layer capacitance. Of course, absolute interface trap capacitance, C_{it}, is just $C_{it}A$ and, thus, corresponds to the formula:

$$\mathrm{C}_{it} = \left(\frac{1}{\mathrm{C}_{LF}} - \frac{1}{\mathrm{C}_{\max}}\right)^{-1} - \left(\frac{1}{\mathrm{C}_{HF}} - \frac{1}{\mathrm{C}_{\max}}\right)^{-1}$$

Obviously, C_{LF} and C_{HF} are observed quasistatic (*i.e.*, low frequency) and high frequency MOS capacitances and, naturally, $C_{\max}$ is just $C_{ox}A$. Alternatively, one can define ΔC simply as the difference between high frequency and low frequency capacitances, *i.e.*, $C_{LF} - C_{HF}$:

$$C_{it} = C_{\max}\left(\left(\frac{C_{\max}}{C_{HF}+\Delta C}-1\right)^{-1}-\left(\frac{C_{\max}}{C_{HF}}-1\right)^{-1}\right) = \Delta C\left(1-\frac{C_{HF}+\Delta C}{C_{\max}}\right)^{-1}\left(1-\frac{C_{HF}}{C_{\max}}\right)^{-1}$$

Of course, it is desirable for C_{it} to be small.

To express these measurements in a more fundamental form, it is possible to relate C_{it} directly to total interface trap density, D_{it}. To begin, one observes that interface traps may exhibit either acceptor-like or donor-like behavior. Physically, this means that acceptor-like traps behave in analogy to shallow level impurity states associated with acceptor dopant atoms and, hence are negatively charged when occupied by electrons. Conversely, donor-like trap states behave similarly to shallow level impurity states arising from donor dopant atoms and are positively charged when occupied by holes, *i.e.*, unoccupied by electrons. Of course, in contrast to shallow level impurity states, which are distributed throughout the bulk of the semiconductor crystal, interface trap states are localized at the Si/SiO_2 interface. Physically, at equilibrium interface trap charge per unit area due to acceptor-like traps, Q_{it}^{a}, corresponds to the integral expression:

$$Q_{it}^{a} = -q\int_{E_V}^{E_C} dEf(E-q\varphi_s)D_{it}^{a}$$

Here, D_{it}^{a} is the density of acceptor-like interface trap states and $f(E)$ is the Fermi-Dirac distribution function for electrons. In this analysis the rigorous Fermi-Dirac distribution function must be used since interface trap states may have energies close to the Fermi level, *i.e.*, inside the band gap. Of course, an explicit factor of $-q$ must also appear since electrons carry a single negative fundamental unit of charge. Naturally, a complementary expression for interface trap charge per unit area due to donor-like traps, Q_{it}^{d}, can be also be constructed as a definite integral as follows:

$$Q_{it}^{d} = q\int_{E_V}^{E_C} dE(1-f(E-q\varphi_s))D_{it}^{d}$$

Of course, D_{it}^{d} is the density of donor-like interface trap states and in this case, the factor q appears since holes are positively charged. Furthermore, $1-f(E)$ appears within the integrand instead of $f(E)$ since the charged state of a donor-like trap corresponds to an unoccupied electronic state. In both of the preceding expressions, the integral is taken over energies within the band gap.

Of course, during a CV measurement, bias voltage is slowly varied, which causes a corresponding variation in the surface potential, $\delta\varphi_s$. Therefore, one can write:

$$\delta Q_{it} = \delta Q_{it}^d + \delta Q_{it}^a = q\delta\varphi_s \frac{\partial}{\partial\varphi_s}\int_{E_V}^{E_C} dE((1 - f(E - q\varphi_s))D_{it}^d - f(E - q\varphi_s)D_{it}^a)$$

Here, δQ_{it}^a and δQ_{it}^d are variations in interface trap charge associated with acceptor-like and donor-like states, respectively, and δQ_{it} is the variation in total interface trap charge. By definition, $q\varphi_s$ is the difference between the actual Fermi level and the intrinsic Fermi level at the Si/SiO$_2$ interface. Therefore, it is clear that if φ_s becomes more negative, acceptor-like states discharge (become neutral) and donor-like states become charged. Conversely, if φ_s becomes more positive, acceptor-like states become negatively charged and donor-like states discharge. Since, capacitance is always defined to be positive, it follows that C_{it} is merely $-\delta Q_{it}/\delta\varphi_s$, hence:

$$C_{it} = q\int_{E_V}^{E_C} dE \frac{\partial f(E - q\varphi_s)}{\partial\varphi_s} D_{it}$$

Here, D_{it} is the total interface trap density and is the sum of D_{it}^a and D_{it}^d. As a matter of mathematics, the partial derivative with respect to φ_s can be formally replaced with a partial derivative with respect to E:

$$C_{it} = -q^2\int_{E_V}^{E_C} dE \frac{\partial f(E - q\varphi_s)}{\partial E} D_{it}$$

From the well-known form of the Fermi-Dirac distribution function, an explicit expression for the partial derivative immediately follows:

$$\frac{\partial}{\partial E} f(E - q\varphi_s) = \frac{-e^{(E - q\varphi_s - E_F)/kT}}{kT(1 + e^{(E - q\varphi_s - E_F)/kT})^2} = -\frac{1}{kT} f(E - q\varphi_s)[1 - f(E - q\varphi_s)]$$

This function is sharply peaked about a value of E equal to $q\varphi_s$ with a peak width of order kT. Therefore, a good approximation to the exact integral is obtained if one assumes that D_{it} is constant over an energy interval of order kT, *i.e.*, one assumes that D_{it} is a smooth, slowly varying function of energy, thus:

$$C_{it} \cong -q^2 D_{it}\int_{E_V}^{E_C} dE \frac{\partial f(E - q\varphi_s)}{\partial E} = q^2 D_{it}(f(E_V - q\varphi_s) - f(E_C - q\varphi_s))$$

Provided that φ_s is has a value that does not locate the Fermi level near the band edges, it is obvious that the Fermi-Dirac distribution function difference is very close to unity, hence:

$$C_{it} \cong q^2 D_{it}$$

It is clear from this expression that the physical interpretation of D_{it} is the number of trap quantum states per unit energy per unit area defined on an energy domain characteristic of the band gap.

A typical plot of D_{it} with respect to surface state energy is shown in the following figure:

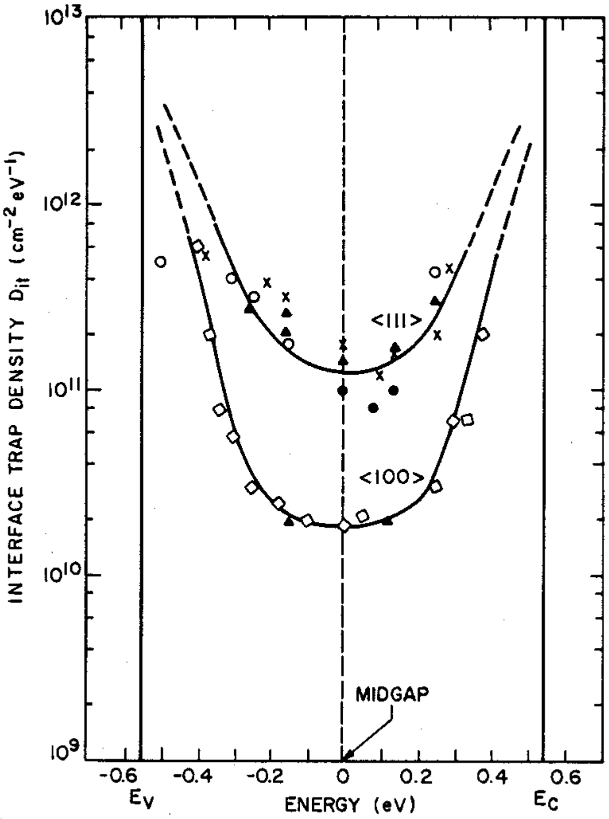


Fig. 46: Interface state density as a function of electronic energy for [111] and [100] silicon surfaces

Clearly, C_{it} and, hence, D_{it} are functions of bias voltage applied during quasistatic and high frequency CV measurements. If one recalls the position of the Fermi level at the semiconductor surface for the various conditions of bias, a physical interpretation of this behavior is readily formulated. Accordingly, irrespective of substrate doping, if the applied bias of an MOS capacitor is swept from negative to positive values, then the surface Fermi level effectively moves upward through the band gap due to changes in band bending. By definition, electronic energy states below the Fermi level tend to be occupied, while those above the Fermi level tend to be empty. Therefore, as the Fermi level moves upward through the band gap, empty electronic states, *i.e.*, interface trap states and fast surface states, of corresponding energy become occupied. Since electrons carry one fundamental unit of charge, these transitions make a directly observable contribution to C_{it}. The bias voltage at which an interface trap becomes charged is directly related to its energy and, hence, its position relative to the band gap. Thus, observation of C_{it} allows D_{it} to be determined as a function of electronic energy measured relative to the band gap as is shown in the preceding figure. Within this context, a common "rule of thumb" is that D_{it} should no more than $1(10^{11})$ cm^{-2} eV^{-1}. Clearly, this is easily realized on a [100] silicon surface, but not on a [111] surface. (Since, a stable switching threshold is essential to reliable operation of modern CMOS transistors, it is critical to obtain a low density of interface traps; hence, this provides a major motivation for the usual preference of [100] silicon substrates for device fabrication.)

Current-Voltage Measurement

Of course, if a dielectric thin film, *e.g.*, thermal silicon dioxide, is to serve as a high quality insulator, then as asserted previously, it is desirable that very little current, *i.e.*, leakage current, should flow if the insulator layer is subjected to normal bias conditions. Therefore, in addition to desirable capacitance-voltage (CV) characteristics, a thermal oxide layer must also have desirable current-voltage (IV) characteristics. Obviously, these can also be measured using the same MOS structures as fabricated for CV measurements. However, physical interpretation is substantially simpler than in the case of CV analysis since the semiconductor substrate, provided that it is sufficiently conductive, is unimportant in IV characterization of a thermal oxide layer.

In principle, IV measurement is quite simple and simply requires biasing the oxide layer at a designated voltage and then measuring resultant current flow. In practice, one observes the measured current as a function of bias voltage. However, this is exactly what was done in a quasi-static CV measurement. So, what is the difference between IV and quasi-static CV measurements? To be specific, there are two major differences. First of all, for a quasi-static CV measurement, transient displacement current, *i.e.*, capacitor charging current, is measured. In contrast, for an IV measurement, one is interested only in steady-state (DC) current, *i.e.*, true conduction current, which continues to flow after the transient has decayed, *i.e.*, after the MOS capacitor has become fully charged at the applied bias voltage. (Indeed, as asserted previously, this current must be formally subtracted from displacement current in order to obtain an accurate quasi-static CV plot.) Second, the bias voltage in an IV measurement is generally carried to values for which the oxide layer *breaks down*, *i.e.*, fails as an insulator. Moreover, once break down has occurred, the oxide layer is permanently damaged, thus IV measurements can be made one time only on any particular MOS structure, *i.e.*, IV testing is essentially destructive. This requires that a substantial number of test structures must be measured in order to generate meaningful statistics. Clearly, break down of the oxide at the outset caused by application of a large initial bias voltage must be avoided. Therefore, IV measurements should be made by sweeping voltage bias slowly from zero toward either positive or negative values, but not as in a CV measurement by sweeping from positive to negative voltages through zero bias (or the reverse).

Conduction Mechanisms

For relatively thick oxide layers, the IV response of “good” oxide is quite simple. At voltage biases well below break down, very little current flows (on the order of a few pA/cm^2). However, once break down occurs, current rises very rapidly. On a semi-logarithmic plot (*i.e.*, current plotted on a logarithmic scale versus voltage or electric field plotted on a linear scale) an ideal thick oxide IV response appears as a flat or slowly rising curve below break down at which point the plot becomes essentially vertical. Obviously, if a large amount of current flows at voltage biases significantly lower than the expected break down voltage, then the oxide quality is poor. Of course, break down voltage must be directly dependent on oxide film thickness; however *break down field strength* is essentially independent of thickness. In general, break down fields for high quality oxides are of magnitude 10-12 MV/cm.

For thin oxides, IV response is complicated by the phenomenon of quantum mechanical *tunneling* (which is a fundamental physical phenomenon that is a direct consequence of the Heisenberg Uncertainty Principle). To be specific, it is impossible to confine particles (such as electrons) completely by a finite potential barrier (such as provided by a layer of thermal oxide in an MOS structure). Therefore, some current can always be expected to “leak”, even through a materially “perfect” insulator completely free of defects. Within this context, in a typical plot of IV response, tunneling current appears as a rising characteristic in the region just below break down. However, particle confinement substantially depends on barrier thickness, and thus, is significant only for very thin oxide layers. This is apparent in the following figure, which illustrates typical IV responses for “thick”, “thin”, and “very thin” oxide layers:

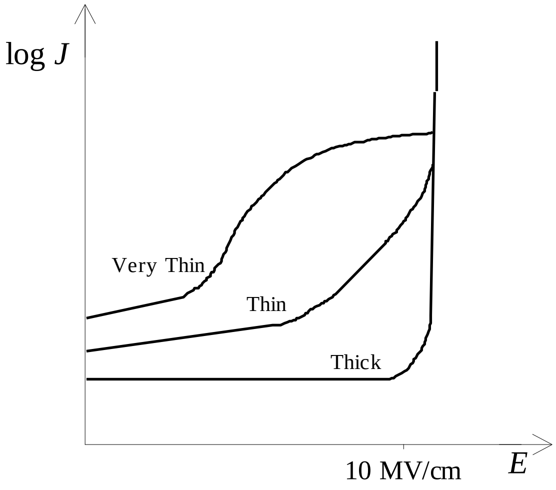


Fig. 47: Idealized current-field characteristics of thick, thin, and very thin thermal oxide layers

Here, for simplicity break down field strength is taken to be the same for each oxide layer; however in actual practice it is likely to be somewhat more variable. In general, tunneling current is essentially independent of oxide quality.

Physically, it is found that at low bias voltages, several electrical conduction mechanisms can exist within thermal oxide. Naturally, the presence of foreign impurities can greatly enhance conduction due to the introduction of physical defects (pinholes, *etc.*). For impurity free oxides, however, *Frenkel-Poole emission* and *Fowler-Nordheim tunneling* are the most common low-field conduction mechanisms. As asserted previously, tunneling is a normal phenomenon that cannot be prevented, and which allows charge carriers, *e.g.*, electrons, to pass through a potential barrier even though available energy for such a process is insufficient. (This accounts for the terminology; since such a particle does not “pass over” the barrier, but “tunnels through” the barrier.) Accordingly, it follows directly from elementary quantum mechanics that the current-field characteristic for Fowler-Nordheim tunneling has an athermal exponential characteristic of the form:

$$J = A_{FN} E^2 \exp\left(-\frac{E_o}{E}\right)$$

Here, A_{FN} and E_o are characteristic constants, J is leakage current density, and E is applied field strength. (A more fundamental expression for E_o can be written in terms of an oxide barrier height, electron effective mass, the fundamental unit of charge, and Planck's constant.) Of course, Fowler-Nordheim tunneling is important in high quality oxides only if they are very thin. However, if there are a large number of "trap states" for electrons distributed within the oxide layer, then tunneling can occur "trap-to-trap". This mechanism causes a dramatic increase in leakage current in comparison to oxides which have a low trap state density, *i.e.*, that are essentially trap free. Furthermore, trap states are generally associated with the fixed oxide charges that are observable in CV measurements. Of course, a large amount of oxide fixed charge and an associated high density of trap states is indicative of a poor quality oxide. Therefore, a large amount of oxide leakage current observed at a low bias voltage can be expected to be correlated with a high trap state density.

Frenkel-Poole emission is also mediated by electronic trap states and occurs if the electric field within the oxide layer becomes large enough so that electrons trapped within the oxide layer are directly injected into the conduction band of the semiconductor. Consequently, it is found that the current-field characteristic for Frenkel-Poole emission has the form:

$$J = A_{FP} E \exp\left[-q\left(\phi_B - \sqrt{\frac{qE}{\pi\varepsilon_{ox}}} \right) \Big/ kT \right]$$

Here, A_{FP} is a "pre-exponential" constant and ϕ_B is a barrier height characteristic of the oxide trap states. Clearly, in contrast to tunneling, Frenkel-Poole emission is a thermally activated process. Another less important conduction mechanism similar to Frenkel-Poole emission is *Schottky emission*, which also has a thermally activated current-field characteristic:

$$J = A^* T^2 \exp\left[-q\left(\phi_B - \sqrt{\frac{qE}{4\pi\varepsilon_{ox}}} \right) \Big/ kT \right]$$

Here, A^* is a coefficient known as "effective Richardson constant" and ϕ_B is, again, trap barrier height. Both Frenkel-Poole and Schottky emission processes should be negligible in high quality thermal oxide, since both mechanisms require a reasonably large density of trap states within the oxide. Typically, a large density of trap states is the result of contamination, damage, and/or generally poor processing.

Two other possible oxide conduction mechanisms are *ohmic* and *ionic* conduction. Both of these are thermally activated:

$$J = A_e E \exp[-q\,\Delta E_{ae}/kT]$$

$$J = \frac{A_i E}{T} \exp[-q\,\Delta E_{ai}/kT]$$

Clearly, these are just Arrhenius forms defined such that ΔE_{ae} and ΔE_{ai} are, respectively, activation energies for ohmic and ionic conduction processes and A_e and A_i are corresponding pre-exponential factors. Neither one of these conduction mechanisms should ever be observed in high quality thermal oxide. By definition, ionic conduction can only occur if the oxide is greatly contaminated with some type of mobile ionic species, *e.g.*, sodium. Ohmic conduction can occur only if the chemical composition of the oxide layer is disturbed.

For completeness, it is worthwhile to consider conduction mechanisms associated with normal oxide break down. If the electric field becomes very high, the current density through the oxide may become *mobility limited*.

$$J = \frac{9\varepsilon_{ox}\mu_e^{ox}V^2}{8x_o^3}$$

Clearly, in this expression the electric field does not itself appear, but rather J depends on bias voltage directly. Furthermore, μ_e^{ox} is identified as electronic mobility within the oxide layer and is analogous to electronic mobility as defined for the semiconductor substrate. Physically, just as in the semiconductor substrate itself, mobility is determined by electron scattering from atomic species, *i.e.*, silicon and oxygen atoms. Of course, since silicon dioxide is an insulator, one expects that μ_e^{ox} should be much smaller than the corresponding electronic mobility of silicon. If the electric field strength due to the applied bias voltage is relatively small, electron scattering processes are essentially elastic and cause no changes in the oxide network structure. However, at or near break down, the electric field strength becomes large. In this case, free electrons injected into the surface of the oxide layer collide with bound atomic electrons causing them also to become free and then to be accelerated by the applied bias. This process is called *electron impact ionization*. The newly freed electrons can then collide with additional bound atomic electrons, thus multiplying the current in a “chain reaction” or *avalanche*. This accounts for the rapidly rising current-field characteristic typical of oxide break down. In these circumstances it is not surprising that the high current density associated with avalanche break down permanently damages the oxide layer.

Oxide Reliability

In addition to determination of interface trap density, leakage current, or break down field, an additional critical criterion for oxide quality is *reliability*, which is quantified as an estimate of expected performance of an oxide layer over some projected usable lifetime. Common methods for determination of oxide reliability are *charge-dependent-breakdown* (QDB) and *time-dependent-breakdown* (TDDB) analysis. In QDB analysis, an MOS capacitor is biased using a constant current source. Obviously, as current is “pumped into” the capacitor, bias voltage must rise until, finally, the capacitor breaks down. Total injected charge is determined simply by multiplying the applied current by the time to reach break down. In general, the larger the total injected charge, the more reliable the oxide layer. In contrast, in TDDB analysis, bias voltage is held constant. In this case, the current flowing through the capacitor is variable. Again, the MOS structure is subject to electrical stress until break down is observed. In principle, TDDB analysis does not require a particular bias voltage, which may be chosen consistent with device characteristic or simply for convenience. Even so, if the chosen bias is too small, then the length of time to observe break down may become extremely long. Conversely, if the bias voltage is too high, TDDB results may not correspond closely to actual operating conditions. In practice, the bias level for TDDB analysis should be set about twice the maximum bias to which an oxide layer will be subjected during normal operation.

For both QDB and TDDB analysis, results will differ depending on whether electrons are injected into the oxide layer from the substrate or from the gate. Therefore, for purposes of comparison, one must adopt a consistent measurement technique. Also, a reasonably large number of MOS capacitors must be measured to obtain an acceptable degree of statistical confidence. In practice, QDB and TDDB data is interpreted by construction of a cumulative probability plot, which has total injected charge (*i.e.*, QDB) or time (*i.e.*, TDDB) required to observe failure (*i.e.*, oxide break down) as the horizontal axis and fraction failed, *i.e.*, failure probability, as the vertical axis. If, for example, a sample of one hundred MOS capacitors is tested until break down is observed, a cumulative probability plot is constructed by ranking measurements from the smallest observed injected charge or shortest time observed for break down and plotting rank against charge or time. In this case, the rank corresponds directly to probability of failure measured in per cent as illustrated in the following figure:

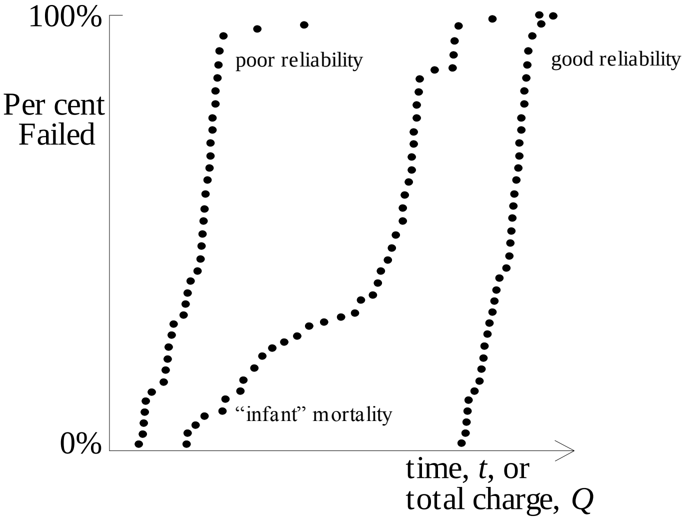


Fig. 48: Cumulative probability plots showing good reliability, poor reliability, and "infant" mortality

Obviously, sample size need not restricted to any particular number, *e.g.*, one hundred, but may be chosen arbitrarily provided statistical confidence is sufficient.

A "good" QDB or TDDB result is characterized by a nearly vertical distribution of data points. This indicates that all of the measured structures are very similar in behavior. Naturally, the larger the average total injected charge or the longer the average time required for break down, the more reliable the oxide. In contrast, the data points may be distributed more horizontally over some range of probability. This indicates that the behavior of the measured structures is inconsistent and that the failure probability distribution is very broad or perhaps even bimodal. Such a result is almost certainly caused by defects or damage in the oxide layer and, in general, represents a “bad” result. Within this context, there is one special case worth consideration, which is characterized by a horizontal distribution of points at low probability above which the points are distributed more vertically and about a reasonably high value of average charge or time. This is characteristic of “infant” mortality. Obviously, it is desirable to eliminate such behavior; however this is not always practical in the fabrication process itself. An alternative (albeit a somewhat costly one) is to perform a “burn-in” in which finished devices are stressed well beyond normal operating conditions. Presumably, this precipitates early or infant failures and the remaining devices should have reliability characteristics of the vertically distributed points.

In general, one observes that absolute QDB and TDDB results will vary with respect to measurement conditions (*i.e.*, injected current density, bias voltage, temperature, *etc.*). However, QDB and TDDB results observed under different conditions can be directly compared by application of a suitable reliability model. In this case, one obtains a *mean time to failure* or MTTF extrapolated from actual measurement conditions to some normal operating condition. There are several reliability models available for this purpose and there is still considerable debate regarding which model is more realistic. However, if TDDB measurements are made at several bias voltages, it is straightforward to extrapolate average break down time at a value of electric field consistent with normal operation. Although not always possible, it is desirable that this result, *i.e.*, MTTF, should be quite long (perhaps, even a few hundred years). In practice, the desired extrapolation may be made by plotting average break down time versus electric field or reciprocal field. Physically, use of the reciprocal field is justifiable since the logarithm of Fowler-Nordheim tunneling current density is proportional to $1/E$. However in practice, more realistic estimates of MTTF seem to be obtained from empirical extrapolations using just the electric field itself. Ideally, for QDB measurements, the actual charge required for oxide failure should be independent of the magnitude of forced current. In this case, MTTF can be estimated just from the measured value of QDB and oxide leakage current characteristic of normal operation. However, as a practical matter QDB may be found to depend on forcing current. In this case, an extrapolation of QDB to operating conditions can be made in close analogy to methods for extrapolation of MTTF from TDDB measurements.

Physical Characterization of Thermal Oxide

In addition to electrical characterization of thermal oxide using CV or IV methods, there are other useful physical techniques for characterization of thermal oxide films. These generally rely on optical measurements and are used to measure physical film thickness, refractive index, *etc.*

Reflectance Spectroscopy and Interferometry

In general, it is well known that a transparent thin film having a thickness commensurate with the wavelength of visible electromagnetic radiation will appear colored when it is illuminated by a broad band white light source, *e.g.*, sunlight or other incandescent source. This is caused by interference between light reflected from the top and bottom interfaces of the thin film. Of course, the intensity of the various reflected spectral components is determined by the relative phase between the two reflections. Naturally, the most intense reflected wavelengths will be those for which reflected components are "in-phase", *i.e.*, interference is constructive. Consequently, it is evident that not all components of illuminating white light are reflected uniformly and, thus, the reflected light appears colored rather than white. (This same phenomenon is readily observed in everyday life in the colors generated by thin oil or soap films.) The apparent color or more precisely the spectral composition of the reflected light is directly related to the thickness of the thin film. This phenomenon is summarized in the following table:

μm	Apparent Color	μm	Apparent Color	μm	Apparent Color
0.00	Metallic or white	0.39	Yellow	0.77	Yellow; washed out
0.05	Tan	0.41	Light orange	0.80	Orange; quite strong
0.07	Brown	0.42	Carnation pink	0.82	Salmon
0.10	Dark violet to red-violet	0.44	Violet-red	0.85	Light red-violet; dull
0.12	Royal blue	0.46	Red-violet	0.86	Violet
0.15	Light blue to metallic blue	0.47	Violet	0.87	Blue-violet
0.17	Metallic to very light yellow-green	0.48	Blue-violet	0.89	Blue
0.20	Light gold or yellow; slight metallic look	0.49	Blue	0.92	Blue-green
0.22	Gold with slight yellow-orange	0.50	Blue-green	0.95	Yellow-green; dull
0.25	Orange to melon	0.52	Green; quite strong	0.97	Yellow; somewhat washed out
0.27	Red-violet	0.54	Yellow-green	0.99	Orange
0.30	Blue to violet-blue	0.56	Green-yellow	1.00	Carnation pink
0.31	Blue	0.57	Yellow; washed out toward gray	1.02	Violet-red
0.32	Blue to blue-green	0.58	Light orange or yellow with a pink cast	1.05	Red-violet
0.34	Light green	0.60	Carnation pink	1.06	Violet
0.35	Green to yellow-green	0.63	Violet-red	1.07	Blue-violet
0.36	Yellow-green	0.68	Blue-gray; washed out with a red cast	1.10	Green
0.37	Green-yellow	0.72	Blue-green to green; strong	1.11	Yellow-green
				1.12	Green
				1.18	Violet
				1.20	Violet-red

Table 3: Apparent colors of thermal oxide of various thickness (in μm; viewed normal to the surface)

Here, oxide thickness is specified in micrometers and apparent colors are specified empirically. Naturally, *reflectance spectroscopy* allows quantification and relies on analysis of the spectral composition of normally reflected light from an oxide thin film. Indeed, in simplest form no equipment other than a "calibrated human eyeball" is required. As is clear from the preceding table, experienced observers can easily estimate oxide thickness to within a few nanometers. However, in recent years, automated instrumentation has been developed which is both more accurate and convenient than simple visual observation. In addition, these instruments collect reflected light through a microscope objective, which allows thickness measurements to be made at very precise locations on the substrate surface. This is particularly useful for characterization of partially fabricated devices. Indeed, if multiple positions are measured using some predefined pattern that essentially samples the whole wafer, then the data can be conveniently rendered into a "map" of thickness.

If instead of a broad band light source, a monochromatic light source (such as a laser) is used, reflectance spectroscopy becomes *reflectance interferometry*. Rather than as an "after-the-fact" characterization method, reflectance interferometry is most useful as an "in-situ" measuring technique, which can be incorporated into processing equipment allowing film thickness to be measured during growth or removal (*e.g.*, by etching or polishing).

Monochromatic and Spectroscopic Ellipsometry

Ellipsometry is a second optical interferometric technique that is frequently used to characterize transparent thin films, *e.g.*, thermal oxide. In its simplest form, a monochromatic beam of light (typically from a laser diode) is resolved into two independent polarized components. These are reflected from the wafer surface at some fixed angle. A second variable polarizer then analyzes the reflected light. It is well known from classical electromagnetic field theory that light components polarized parallel and perpendicular with respect to the surface exhibit different, independent behavior with respect to reflection. As before for reflectance spectroscopy, interference occurs between light reflected from the top and bottom interfaces of the oxide layer. Moreover, the intensity and phase of each polarized component of the reflected light is characteristically dependent on thickness and refractive index of the thin film. Since, the two polarized components are independent, *monochromatic ellipsometry* can make simultaneous measurements of both thickness and refractive index of an oxide layer. In addition, ellipsometry is inherently more precise than a reflectance spectroscopy and is particularly useful for characterization of very thin oxide layers. Ellipsometry can be extended either by measuring at various angles of reflection, (*multiangle ellipsometry*) or by using various wavelengths of light (*multiwavelength ellipsometry*). In both of these cases, each measurement made at different angles or wavelengths provides two independent values of thickness and refractive index. For a single thin film (such as a thermal oxide layer) these can be used to increase the accuracy of the overall measurement. However, a more common application of these techniques is simultaneous, independent measurement of refractive indices and thicknesses of two or more "stacked" transparent thin film layers.

Extending ellipsomety further, if an intense broad band light source is used, ellipsometric measurements can be made over a continuous range of wavelengths. This is *spectroscopic ellipsometry*. Clearly, since two independent measurements can be made at any particular wavelength, the amount of information available in a single spectroscopic ellipsometric measurement is quite large. Again, this is useful for characterization of multilayer thin films, however, for characterization thermal oxide spectroscopic ellipsometry can be used to accurately account for the optical properties of the Si/SiO_2 interface and/or the semitransparent surface layers of the substrate itself. In practice, this requires extensive numerical fitting to some "model" of the oxide or Si/SiO_2 interface. In recent years, several systems have been developed which provide algorithmic support for spectroscopic ellipsometry. This allows very accurate characterization of thin thermal oxide films.

Prism Coupling

A third optical technique that can be used for thin film characterization is *prism coupling*. This method is used less frequently at present than previously since it suffers from the disadvantage that, in contrast to ellipsometry or reflectance spectroscopy which are non-contacting; a small prism must come in contact with the thin film surface. The prism is made of a transparent material chosen so that the interface between the prism and the thin film layer forms a totally reflecting interface. Physically, it turns out that if the prism-thin film couple is rotated with respect to a monochromatic optical source (*i.e.*, a laser) some of the totally reflected light "leaks out" into the thin film due to the phenomenon of *evanescent coupling*. As might be expected, the degree of evanescent coupling depends on refractive index and thickness of the thin film material and appears as a series of interference fringes that are a function of angular position. In practice, if several interference fringes (more than three or four) can be observed, then both thickness and refractive index of the thin film can be determined. Prism coupling is most applicable to relatively thick transparent thin films. This is useful since both reflectance spectroscopy and ellipsometry typically become less accurate for thick films due to cyclic error, *etc.*

Pre-Oxidation Cleaning

It is critical that prior to any thermal oxidation process, the silicon surface should be scrupulously cleaned. Typically, this is done by treating the wafers in two successive chemical solutions conventionally called “SC-1” and “SC-2” (surface cleans 1 and 2). The precise composition of these solutions is somewhat variable; however, SC-1 is generically formulated as a 10:1 mixture of commercial ammonium hydroxide (NH_4OH) and hydrogen peroxide (H_2O_2) solutions heated to about 80°C. Similarly, SC-2 is also approximately a 10:1 mixture of commercial hydrochloric acid (HCl) and hydrogen peroxide (H_2O_2) solutions, again, heated to about 80°C. Naturally, all starting materials must be “electronic grade”. These mixtures are also sometimes diluted with an equivalent volume of de-ionized water; however, this is not a requirement. It has been shown that SC-1 is effective primarily for removal of organic contamination and SC-2 for removal of metallic species. Therefore, any type of contamination is substantially reduced by sequential treatment in SC-1 and SC-2. Originally, of course, these cleans were carried out in simple static tanks. As particle control has become more critical, recirculation, filtration, and automation have been added. Alternatively, it is common practice to use completely automated spray chemical processors in which wafers are cleaned, rinsed, and dried without any external intervention.

In addition to contamination, it is often desirable to remove any pre-existing “native oxide” from the silicon surface. (Native oxide is a naturally occurring thin oxidized layer one to two nm thick on the surface of silicon that arises from routine exposure to oxygen and water vapor in the atmosphere.) This may be done by a short etch in unbuffered 50:1 hydrofluoric acid solution following cleaning in SC-1 and SC-2. (Buffered oxide etch or BOE should not be used because it contains the salt, ammonium fluoride, which can form particles on the surface.) Removal of native oxide is thought to result in a “hydrogen terminated” silicon surface, which is believed to persist perhaps one to four hours depending on conditions, before the native oxide layer is reformed.

In addition to wet chemical treatments, vapor phase processing using anhydrous hydrogen fluoride is also an option for pre-oxidation cleaning. In recent years, various systems have been developed for this purpose. Some of these have even been integrated with the oxidation process. None of these seem to have proved entirely satisfactory both in terms of cost and performance.

Ultra-Thin Insulators

As will become evident in more detail subsequently, in a very real sense an MOS transistor is the solid-state analog of an old-fashioned vacuum tube. (It is an MOS capacitor built on top of a semiconductor resistor hence, the term “transfer resistor” or transistor.) Obviously, any bias voltage applied to the gate modulates current flow in the channel. Of course, the conduction channel is electrically connected to the circuit wiring on each end (source and drain contacts). Therefore, the gate is insulated and ideally should not supply any current to the channel. Within this context, one observes that state-of-the-art MOS transistors require fabrication of ultra-thin gate insulators which, in addition, must satisfy stringent performance specifications. Indeed, insulating properties must not only be excellent, but since the gate electrode is generally made of heavily doped CVD polysilicon, metal, or metal alloy, the gate insulator must also block any migration of metal atoms or shallow level dopant impurity, in particular boron, from the gate electrode into the conduction channel as well. (If there is any significant dopant contamination of the channel from the polysilicon gate electrode, the transistor threshold voltage will have unacceptable variation.)

Reoxidized Nitrided Oxide

Unfortunately, a pure thermal oxide film has poor characteristics with respect blocking dopant migration from the polysilicon (particularly in the case of boron). This problem may be addressed by direct incorporation nitrogen into a thermal gate oxide film. Within this context, one might ask, why not just use pure silicon nitride? Certainly, nitride is a good insulator and, indeed, very thin nitride layers can be produced by CVD or even direct thermal nitridation of a silicon surface. However, a silicon nitride/silicon interface has a very high density of interface traps. Within the context of MOS transistor performance, this causes a severe degradation of effective carrier mobility in the channel. (In practical terms, this appears as a high channel resistance or low drive current.) As a practical matter, it is found that if the nitrogen concentration exceeds one atomic per cent at the gate oxide/silicon interface there is a significant degradation of device performance. Thus, one would ideally like to have a graded nitrogen concentration in the gate oxide with the concentration relatively high at the polysilicon/gate oxide interface and low at the gate oxide/silicon interface. However, this is difficult to achieve in practice.

The most practical process for formation of a nitrided oxide is the addition of either nitrous (N_2O) or nitric (NO) oxide directly to the oxidizing ambient. Unfortunately, since further oxidation tends to convert nitride to oxide, this also results in the occurrence of maximum nitrogen concentration precisely at the gate oxide/silicon interface. Thus, for a one step oxidation process, the total concentration of nitrogen in the oxide must be kept low. This suggests implementation of a two step process as a possible improvement. In the first step a relatively nitrogen rich oxide is grown. This is followed by “reoxidation” in non-nitrogen containing ambient. Since new oxide is formed only at the gate oxide/silicon interface, the interfacial nitrogen concentration falls rapidly. However, again since oxidation converts nitride to oxide, the nitrogen content of the entire film also falls. Thus, the growth rate of oxide at the interface must be balanced with the conversion of nitride to oxide in the bulk of the film. Fortunately, this trade-off can be achieved since it is found in practice that only a small nitrogen concentration is effective at blocking “boron penetration”.

In passing, it is worthwhile to mention other alternative approaches for fabrication of nitrided gate oxide. In particular, very shallow ion implantation of nitrogen either into a preformed gate oxide or into the silicon substrate itself upon which a subsequent gate oxide is grown have both been tried. In either case, results do not appear to be as good as that obtained using some form of N_2O or NO oxidation processes.

Rapid Thermal Oxidation

Alternatively, rapid thermal oxidation (RTO) can also produce a high quality oxide. Typically, RTO is implemented by use of high intensity quartz-halogen lamps rather than ordinary resistive heating elements as a heat source. Of course, the conventional tube configuration is generally not optimal for RTO, which is much more compatible with “single wafer processing”. Accordingly, individual wafers are typically suspended on rings or pins within a small process chamber to reduce thermal mass and allow rapid change of the temperature. Indeed, it is possible to achieve a very high wafer surface temperature (>1000°C) quite quickly, *e.g.*, in less than a minute. Therefore, in principle, thin gate oxides can be controllably grown in a very short period of time. In practice, RTO processes have suffered from problems of repeatability and control and, thus, have not been used widely for conventional oxidation.

Post Oxidation Annealing

As observed previously, for a high quality gate oxide, D_{it} should be no more than $1(10^{11})$ cm^{-2} eV^{-1}. If this is to be achieved in a single oxidation step, this requires oxidation at very high temperature. (Conventionally, this favors dry oxidation over steam since the lower rate allows for better thickness control.) Alternatively, post-oxidation annealing in an inert ambient at high temperature (>1000°C) can reduce an unacceptably high post-oxidation D_{it} to a desirable value. Presumably, this “repairs” the Si/SiO_2 interface. Of course, it goes almost without mentioning that any high quality oxidation process requires scrupulous pre-cleaning of the substrate surface.

Limitation of Conventional Oxidation Technology

The channel length for the current device generation is 20 nm (or less). Such short channel lengths require gate insulator thicknesses equivalent to no more than 1-2 nm of pure silicon dioxide in order to achieve desirable device characteristics. (If the gate insulator thickness is not scaled with the channel length, resulting transistors suffer from severe “short channel effect”.) Fabrication of such thin layers requires careful processing, although, perhaps somewhat surprisingly, process conditions using conventional quartz tube furnaces can be found that result in high yield and very reliable insulating films. In principle, thin oxides can be produced by either diluting the oxidant with inert gas and/or reducing the oxidation temperature. In the latter case, a subsequent high temperature anneal in an inert ambient is needed to reduce interface trap density. In practice, even high quality thin oxide films generally must be modified or replaced to reduce current flowing between the gate electrode and the channel. Indeed, this is not due to formation of defects in thin oxide films and there is no inherent problem in fabrication of such thin films by thermal oxidation; however, current resulting directly from quantum mechanical tunneling of electrons through the oxide layer can be expected to be impracticably large. This presents a very severe limitation on device performance. Therefore, either physical characteristics of thermal oxide must be markedly improved (not likely), a new device architecture must be invented (*e.g.*, the recently introduced “tri-gate” structure), or some other methodology must be found.

Obviously, one solution to the problem of thin gate insulators is partial or complete replacement of thermal silicon dioxide with some other material. Any such material must have a large dielectric constant, *i.e.*, “high-k”, such that physical thickness can be much larger than the physical thickness of an electrically equivalent silicon dioxide layer. These are generally identified with oxides and/or silicates of heavy metals such as zirconium, tantalum, or hafnium. (Other possibilities include a host of perovskite materials.) Of course, as one might imagine such a radical change in device structure is not to be undertaken unless out of absolute necessity since it is precisely the unique compatibility of the silicon-silicon dioxide material system that makes modern solid-state electronics really possible. In particular, for any insulator different from thermal oxide a much poorer silicon/insulator interface, *viz.*, high trap density, is to be expected. Even so, in recent years these kinds of changes have been required for the most advanced device structures. Accordingly, successful integration of such materials requires careful engineering and close attention to processing. In particular, silicates have been attractive since they combine some of the essential properties of silicon dioxide with larger values of the dielectric constant.

Impurity Diffusion in Semiconductors

In the practical fabrication of solid-state electronic devices, it is generally necessary to introduce controlled amounts of various shallow level impurities, *i.e.*, dopants (B, P, or As), into particular regions within the silicon crystal. Indeed, boundaries between regions inside the volume of the wafer for which extrinsic doping changes from *p*-type to *n*-type or vice-versa form electrically active structures called *pn-junctions*. (Along with MOS capacitors, *pn*-junctions are the most important fundamental components of solid-state devices.) In general, although the wafer may have some uniform background doping added to the original melt during manufacture of the substrate itself, it is usual for additional dopants to be introduced through the surface of wafer. These are commonly restricted to specific laterally defined regions of the wafer surface by some type of mask, *i.e.*, one type of dopant might be introduced into some particular area (or areas) and other types of dopants introduced elsewhere. In any case, the vertical and lateral distribution of these dopant atoms may be precisely manipulated by carefully controlled diffusion. Such diffusion processes are thermally activated and, thus, are carried out in quartz tube furnaces very similar to those used for thermal oxidation. (However, the atmosphere inside the furnace generally will be inert or reducing rather than oxidizing.)

Linear Transport Processes

Diffusion of shallow level dopants in semiconductors, *e.g.*, silicon, is a specific example of a broad class of physical processes called *transport processes*. Other examples are conduction of heat and electricity and viscous fluid flow. Physically, transport processes are characteristic of physical systems which are not in thermodynamic equilibrium. Indeed, from a theoretical point of view, transport processes are dissipative in nature, which when occurring within some physical system, proceed to establish the system in equilibrium at which time any net transport comes to a halt. (The general study of transport processes and the approach to equilibrium is called *non-equilibrium thermodynamics*.) Conventionally, transport processes are considered within the context of a *linear phenomenology*, which means that they are described by expressions of the generic form:

$$J_a = \mathrm{L}_{ab} X_b$$

Here, J_a is defined as *flux* (or, more generally, a *flux vector*) identified with transport of some physical property, *a*, *e.g.*, mass, momentum, energy, charge, *etc*. Similarly, X_b is defined as *driving* or *thermodynamic force*, identified with a disequilibrium in some physical property, *b*, *e.g.*, gradients of concentration, fluid velocity, temperature, electrical potential, *etc*. Physically, a thermodynamic force quantifies the magnitude of any disequilibrium driving net transport processes. Thus, fluxes and forces are related by the parameter, L_{ab}, called a *phenomenological transport coefficient*. In the most general formulation (as above), fluxes and forces formally appear as column vector components and transport coefficients as square matrix elements. This allows for the possibility that a force in one physical property, *b*, may drive a flux in some different physical property, *a*. Indeed, such “cross effects” are commonly observed. Representative examples are provided by thermal diffusion or thermoelectric effects in which case a temperature gradient drives material or electrical transport respectively. Clearly, cross effect transport coefficients correspond to off-diagonal matrix elements as defined by the preceding general expression. Of course, “direct effects” for which a force in a physical property drives a flux in the same property correspond to diagonal matrix elements and, therefore, are generally more important than cross effects. Obviously, ordinary diffusion, heat and electrical transport, viscous fluid flow, *etc.* provide elementary examples of just such processes. Therefore, to describe impurity diffusion in semiconductors, it is only necessary to consider direct effects, *i.e.*, only diffusive forces and fluxes. In such a case, the general matrix expression can be simplified to a simple linear proportionality:

$$J_a = \mathrm{L}_a X_a$$

In this expression, the direct effect transport coefficient, L_a, represents an ordinary numerical quantity rather than a matrix element. Clearly, for impurity diffusion, a corresponds to some impurity species, hence, L_a is identified as diffusivity of a (usually symbolized as D_a). In passing, it is, again, useful to observe that this same linear phenomenology can be applied to various specific physical situations that might superficially appear unrelated. Therefore, several specific cases are summarized as follows:

<table>
<tr><td colspan="3">Ohm’s Law of electrical conduction: $j = \sigma E = E/\rho$</td></tr>
<tr><td>J = electric current density, j
(units: A/cm^2)</td><td>X = electric field, $E = -\nabla V$
(units: V/cm)
V = electrical potential</td><td>L = conductivity, $\sigma = 1/\rho$
(units: mho/cm)
ρ = resistivity (Ω cm)</td></tr>
<tr><td colspan="3">Fourier’s Law of heat transport: $\mathrm{q} = -\kappa\nabla T$</td></tr>
<tr><td>J = heat flux, q
(units: W/cm^2)</td><td>X = thermal force, $-\nabla T$
(units: °K/cm)
T = temperature</td><td>L = thermal conductivity, κ
(units: W/°K cm)</td></tr>
<tr><td colspan="3">Fick’s Law of diffusion: $F = -D\nabla C$</td></tr>
<tr><td>J = material flux, F
(units: /sec cm^2)</td><td>X = diffusion force, $-\nabla C$
(units: /cm^4)
C = concentration</td><td>L = diffusivity, D
(units: cm^2/sec)</td></tr>
<tr><td colspan="3">Newton’s Law of viscous fluid flow: $F_u = -\eta\nabla u$</td></tr>
<tr><td>J = fluid velocity flux, F_u
(units: /sec^2 cm)</td><td>X = viscous force, $-\nabla u$
(units: /sec)
u = fluid velocity</td><td>L=viscosity, η
(units: /sec cm)</td></tr>
</table>

Table 4: Summary of common linear transport phenomena

Here, the subscript a has been dropped and J, X, and L correspond to the simple linear transport relation:

$$J = \mathrm{L}X$$

Clearly, the preceding correspondences are useful because once one type of transport process is considered, *e.g.*, impurity diffusion, any results obtained can be immediately applied to other types of transport by the simple expedient of redefinition and/or substitution of the appropriate phenomenological parameters.

Limiting further consideration specifically to the case of impurity diffusion, one considers transport of impurity species through a hypothetical bar of some solid material, *e.g.*, semiconductor, having a uniform cross section. Furthermore, for additional simplicity, one assumes that the concentration of impurity species varies only along the length of the bar and is constant over any given cross section. Thus, assuming that transport fluxes remain constant over some small time interval, Δt, then the net change per unit time in the number of impurity atoms in a small volume element of width, Δx, located at a distance, x, from the end of the bar is given by the simple difference expression:

$$\frac{\Delta N}{\Delta t} = [F(x) - F(x + \Delta x)]A$$

Here, N is the number of impurity atoms found within the volume element, $F(x)$ is impurity flux, and A is the cross sectional area of the bar. This is illustrated by the following figure:

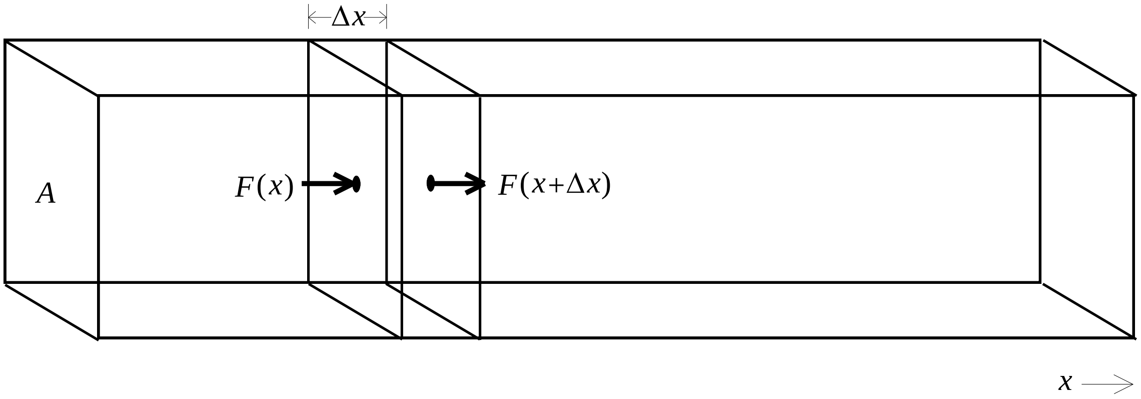


Fig. 49: Diffusion in a rectangular bar of constant cross section

If impurity concentration, C, is defined as usual as the quotient of N divided with volume, then one obtains:

$$\frac{\Delta C}{\Delta t} = -\frac{F(x + \Delta x) - F(x)}{\Delta x}$$

In the limit that the material volume element is allowed to become arbitrarily thin, the right hand side of this expression can just be identified as the negative of the derivative of the material flux with respect to *x*. Similarly, the left hand side can be identified as the derivative of concentration with respect to time, hence:

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x}$$

Here, partial derivatives are written since *C* and *F* are functions of both position and time. Of course, material flux and concentration are related by Fick's Law, which in a single dimension has the form:

$$F = -D\frac{\partial C}{\partial x}$$

Here, *D* is the diffusivity of the impurity species. If the two preceding equations are formally combined, one obtains a single second order linear partial differential equation as follows:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2}$$

This equation is conventionally called Fick's equation, Fick's Second Law, or just the diffusion equation. Clearly, it is a closed form expression for concentration as a function of position and time. (By convention, Fick's First Law is just the linear transport relation defined previously.) Furthermore, within the present context, the diffusion equation has been derived in a one dimensional form. For an elementary description of impurity diffusion in semiconductors this is adequate. However, in more complicated situations diffusion in more than one dimension must be considered. Obviously, Fick's Second Law can be generalized to all three dimensions just by replacing the second order partial derivative with the Laplacian:

$$\frac{\partial C}{\partial t} = D\nabla^2 C$$

In summary, Fick's Laws are useful for the description of diffusion of relatively dilute solutes. If the concentration becomes sufficiently high, due to interactions between solute atoms *D* may become dependent on the concentration, *C*. In this case, diffusion becomes non-linear and is much more difficult to treat mathematically.

Solution of Fick's Equation

Construction of a general solution of the diffusion equation in one dimension is quite straightforward. First of all, one must separate the space and time variables. This can be accomplished by assuming that the concentration, *C*, is a formal product of a function of position, $g(x)$, and a function of time, $f(t)$:

$$C(x,t) = g(x)\,f(t)$$

Upon substitution of this form, it follows that:

$$\frac{1}{Df}\frac{\partial f}{\partial t} = \frac{1}{g}\frac{\partial^2 g}{\partial x^2}$$

Clearly, all of the *x* dependence appears on the left hand side and all of the *t* dependence appears on the right hand side. Since the variables have now been separated, one can set each side equal to an unknown "separation constant". Therefore, the arbitrary constant, λ, is defined such that:

$$\frac{1}{Df}\frac{\partial f}{\partial t} = -\lambda^2 \qquad ; \qquad \frac{1}{g}\frac{\partial^2 g}{\partial x^2} = -\lambda^2$$

The form, $-\lambda^2$, is used purely for mathematical convenience. These two ordinary differential equations are easily integrated by elementary methods. In the case of the time equation one has:

$$\ln f = -D\lambda^2 t + \ln \gamma_t$$

$$f = \gamma_t e^{-D\lambda^2 t}$$

Here, γ_t is an unknown constant. In the case of the space equation, one immediately recognizes that the solution can be expressed as either a sine or a cosine:

$$g = \gamma_{x1} \sin \lambda x \qquad ; \qquad g = \gamma_{x2} \cos \lambda x$$

However, it is more convenient to express this in equivalent form as a complex exponential:

$$g = \gamma_x e^{i\lambda x}$$

Here, γ_x is a second unknown constant. Thus, it follows that a particular solution of the diffusion equation, C_λ, can be written as follows:

$$C_\lambda = \alpha e^{-D\lambda^2 t} e^{i\lambda x}$$

In this expression, α is the product of the arbitrary constants γ_t and γ_x and, thus, is itself an unknown constant. Obviously, the separation constant, λ, labels each particular solution.

It is well known that particular solutions of linear differential equations satisfy the Principle of Superposition. Simply stated, this means that if any two functions are independent solutions of some differential equation, then the sum (or difference) of the two is also a solution. Thus, if one considers λ to be a continuous variable, it follows that a general solution of the diffusion equation can be written as an integral, *i.e.*, a limiting sum, over all particular solutions:

$$C(x,t) = \int_{-\infty}^{\infty} d\lambda \alpha(\lambda) e^{-D\lambda^2 t} e^{i\lambda x}$$

Here, α is now treated as an unknown function of λ. It is desirable to express α(λ) in terms of some initial condition, $C_0(x)$, defined such that:

$$C_0(x) = C(x,0) = \int_{-\infty}^{\infty} d\lambda \alpha(\lambda) e^{i\lambda x}$$

Upon inspection, one observes that $C_0(x)$ is just the ordinary Fourier transform of α(λ). Furthermore, Fourier transformation is easily inverted, therefore, α(λ) can be written explicitly as follows:

$$\alpha(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dx C_0(x) e^{-i\lambda x}$$

Clearly, the "inverse Fourier transform" is identical to the "forward Fourier transform" except that a factor of 1/2π appears. (These forms can be made exactly identical if one removes a factor of $\sqrt{2\pi}$ from the denominator of the above expression in which case one finds that $C_0(x)$ and $\alpha(\lambda)\sqrt{2\pi}$ define a formal "Fourier transform pair".) Thus, substitution of the above result into the expression for $C(x,t)$ yields:

$$C(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\lambda \int_{-\infty}^{\infty} dx' C_0(x') e^{i\lambda(x-x')} e^{-D\lambda^2 t}$$

This expression can be further simplified by completion of the square in the exponent:

$$C(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dx' C_0(x') e^{-(x-x')^2 / 4Dt} \int_{-\infty}^{\infty} d\lambda e^{-Dt\left(\lambda - \frac{i(x-x')}{2Dt}\right)^2}$$

The integral over λ can be determined using standard methods (*e.g.*, complex contour integration), however it is essentially an integral over a Gaussian function and is found to be equal to $\sqrt{\pi / Dt}$:

$$C(x,t)=\frac{1}{2\sqrt{\pi Dt}}\int_{-\infty}^{\infty}dx'C_0(x')e^{-(x-x')^2/4Dt}$$

Clearly, impurity concentration at any time, *t*, is completely determined by the initial concentration. Indeed, this is a completely general result and is applicable for any initial concentration, $C_0(x)$.

Instantaneous Source

A specific form for $C_0(x)$ that is of particular interest can be represented mathematically as follows:

$$C_0(x)=2N\delta(x-x_0)$$

Here, $\delta(x-x_0)$ is a Dirac delta function, which is defined to be zero everywhere except in the case that $x=x_0$ where it becomes infinite. Furthermore, the integral of $\delta(x-x_0)$ over *x* is finite and equal to one. This is called an *instantaneous source*. An instructive way to view a delta function is as a normalized Gaussian function, which has a standard deviation of zero. Therefore, this initial concentration corresponds to an infinitely thin sheet of impurity located at a position, x_0. Within this context, the coefficient, *N*, is the number of dopant atoms per unit area of the sheet and is called *dose*. (The factor of 2 included in the definition of $C_0(x)$ is a geometrical factor which accounts for the fact that a wafer is, perhaps, better regarded as a semi-infinite diffusion domain rather than an infinite domain.) Thus, $C(x,t)$ can be trivially determined if one substitutes the preceding form for $C_0(x)$:

$$C(x,t)=\frac{1}{2\sqrt{\pi Dt}}\int_{-\infty}^{\infty}dx'(2N\delta(x'-x_0))e^{-(x-x')^2/4Dt}=\frac{N}{\sqrt{\pi Dt}}\int_{-\infty}^{\infty}dx'\delta(x'-x_0)e^{-(x-x')^2/4Dt}$$

Integration over x' is trivial due to the delta function, hence:

$$C(x,t)=\frac{N}{\sqrt{\pi Dt}}e^{-(x-x_0)^2/4Dt}$$

Clearly, an instantaneous source results in a *Gaussian concentration profile*. (The terminology "concentration profile" is generally used to describe dependence of impurity concentration in a one dimensional sense.)

In practice, a Gaussian concentration profile describes an impurity diffusion process for which an "infinitely thin" initial layer of dopant, *i.e.*, shallow level impurity, is deposited on the wafer surface. This surface deposition is followed by diffusion at elevated temperature for some time, *t*. Obviously, since x_0 is zero by definition, the concentration profile takes the simplified form:

$$C(x,t) = \frac{N}{\sqrt{\pi Dt}} e^{-x^2 / 4Dt} = C_s e^{-x^2 / 4Dt}$$

Here, C_s is *surface concentration* and is equal to $N / \sqrt{\pi Dt}$. Obviously, for t equal to zero, C_s is infinite just as one expects from the original delta function concentration profile.

Constant Source

A second initial concentration profile, which is generally useful for the description of impurity diffusion has the explicit form:

$$C_0(x) = 2C_s(1 - H(x - x_0))$$

Here, $H(x - x_0)$ is a Heaviside or unit step function and is formally defined to be equal to zero if $x<x_0$ and equal to one if $x>x_0$. Thus, $C_0(x)$ is equal to $2C_s$ if $x<x_0$ and equal to zero if $x>x_0$. This is called a *constant source*. Clearly, upon substitution the step function "cuts off" the integral above a value of x_0 and one obtains the result:

$$C(x,t) = \frac{C_s}{\sqrt{\pi Dt}} \int_{-\infty}^{x_0} dx' e^{-(x-x')^2 / 4Dt}$$

This integral is modified by defining a new integration variable x'' equal to $x-x'$, hence:

$$C(x,t) = \frac{C_s}{\sqrt{\pi Dt}} \int_{x-x_0}^{\infty} dx'' e^{-x''^2 / 4Dt}$$

The integral cannot be constructed in closed form but has a standard definition in terms of the error function, erf(x) or complementary error function, erfc(x):

$$C(x,t) = C_s\left(1 - \text{erf}\left(\frac{x - x_0}{2\sqrt{Dt}}\right)\right) = C_s \text{erfc}\left(\frac{x - x_0}{2\sqrt{Dt}}\right)$$

Hence, a constant source results in a *complementary error function concentration profile*.

In contrast to the previous case, this type of concentration profile describes impurity diffusion processes for which the surface of the wafer remains in equilibrium with some dopant source (solid, liquid, or gaseous) during exposure to elevated temperature, *i.e.*, during the diffusion process. Therefore, the surface concentration can generally be identified as the maximum solid solubility of the dopant in silicon. Hence, it is usual to set x_0 to zero and, thus, the concentration profile takes the form:

$$C(x,t) = C_s \mathrm{erfc}\left(\frac{x}{2\sqrt{Dt}}\right)$$

Clearly, the surface impurity concentration remains constant in the case of a complementary error function profile. In contrast, for a Gaussian profile, C_s decreases as $1/\sqrt{t}$. Furthermore, the total integrated amount of impurity present within the wafer is constant in the case of an instantaneous source; however, it continues to increase in the case of a constant source. Physically, this is easily understood if one observes that all impurity species diffused into the wafer are initially present in an instantaneous source. However, a constant source continues to introduce impurity atoms into the wafer surface during the diffusion process.

Characterization of Dopant Diffusion

The most elementary method for characterization of dopant diffusions are measurements of surface contact current-voltage, *i.e.*, resistance. These are usually made using some type of four terminal configuration such as a direct surface probe *i.e.*, a four point probe, or patterned structures fabricated during processing as a test key (or test element group, TEG).

It is useful to digress here and point out that confusion sometimes occurs between the terms resistivity and *sheet resistance*. Resistivity has been defined previously for extrinsically doped silicon; however, it is a property of any conductor. Moreover, it should be re-emphasized that resistivity is an *intensive* physical property and, therefore, is independent of the amount or shape of the material, *e.g.*, silicon. Furthermore, it is the reciprocal of the electrical conductivity, which is a standard linear transport coefficient as defined previously. Therefore, in terms of a linear transport relation, *i.e.*, Ohm's Law, one can write:

$$E = \rho j$$

Here, j is current density, which has units of A/cm^2 and E is electric field, which has units of V/cm. Of course, one recalls that an ohm, Ω, is defined as a volt per ampere, V/A; hence, as expected, ρ, has units of Ω cm. This expression is quite general and is applicable if linear phenomenology provides an adequate description of electrical transport processes (which is almost invariably the case).

In contrast, sheet resistance, R_s, is a characteristic property of a thin conducting film and is commonly expressed in terms of "ohms per square" or Ω/sq. (Shallow diffusions or epitaxial layers can be generally treated as thin conductive films.) However, this terminology can lead to confusion, since "per square" does not mean "per unit area". Indeed, it turns out that one should regard sheet resistance as a "hybrid" material property which is intensive in two dimensions, specifically, the two horizontal dimensions parallel to the surface of the thin film, and *extensive* in the third vertical dimension, *i.e.*, perpendicular to the thin film surface. Thus, R_s depends on film thickness, but not area. Therefore, "per square" denotes any square unit of film regardless of size. Resistance is measured, assuming a uniform current distribution, "across" the square. For materials with uniform physical properties, resistivity and sheet resistance are related by a simple formula:

$$R_s = \frac{\rho}{x_f}$$

Here, x_f is just thin film thickness. However, sheet resistance remains well-defined even if material resistivity varies vertically (as in a diffusion process). In this case, the measured sheet resistance is to be considered as a parallel combination of thin sheets of uniform resistivity stacked one on top of another. Accordingly, if one identifies the thickness of each sheet as Δx_j, then one can write:

$$\frac{1}{R_s} = \sum_j \frac{\Delta x_j}{\rho(x_j)}$$

Clearly, in the limit that Δx_j tends to zero, a corresponding integral formula is obtained:

$$\frac{1}{R_s} = \int_0^{x_J} \frac{dx}{\rho(x)}$$

Hence, $\rho(x)$ can be directly determined from dopant concentration profile. The upper limit, x_J, is called *junction depth*.

Surface Probing

Consider a single, arbitrarily sharp conductive probe contacting a semi-infinite volume, *i.e.*, substrate, of conductive material. Suppose that this probe is injecting a total current, *I*, into the material:

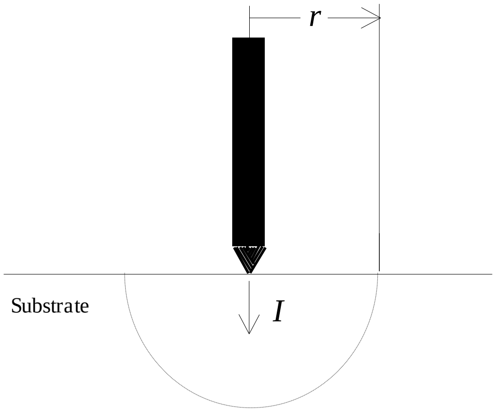


Fig. 50: Current injection into a semi-infinite substrate from a single probe

Ignoring any effects of the surface, the current diverges from the probe tip into the medium through a series of concentric hemispherical shells of equipotential. Clearly, the current density through each shell is $I/2\pi r^2$. (Of course, $2\pi r^2$ is just the surface area of a hemisphere of radius, *r*.) The electric field across each hemispherical shell is just $-\Delta V/\Delta r$, such that ΔV is the "voltage drop" and Δr is the shell thickness. The negative sign appears because voltage decreases as radius increases. If one applies Ohm's Law and allows the shell thickness to tend toward zero, then one obtains the simple formula:

$$\frac{dV}{dr} = -\rho\left(\frac{I}{2\pi r^2}\right)$$

This formula can be trivially integrated as follows:

$$V - V' = \frac{\rho I}{2\pi}\left(\frac{1}{r} - \frac{1}{r'}\right)$$

If the boundary condition is adopted that voltage falls to zero at large distances, *i.e.*, as r' tends to infinity, then this equation takes the simple form:

$$V = \frac{\rho I}{2\pi r}$$

Physically, V is voltage measured at a radial distance, r, from the probe tip subject to this boundary condition. Therefore, one observes that if r is zero, the measured voltage becomes infinite. This, of course, is not a realistic situation, but would be true for an "infinitely sharp" probe tip with a radius of zero. However, any real probe tip evidently must have some finite radius. Consequently, the preceding formula can only be applied when r is large in comparison to the probe tip radius. In practice, this condition is easily satisfied and presents no practical problem.

At this point, suppose that, instead of a single probe, one contacts the surface of the semi-infinite volume with four probes. Furthermore, suppose that these probes are all arranged in a line, equally spaced at some fixed distance, *s*, and, for convenience are labeled 1 through 4. As a condition of measurement, suppose current is being forced through the end probes, *i.e.*, 1 and 4, such that a current, *I*, enters the semi-infinite volume through probe 1 and leaves the volume through probe 4:

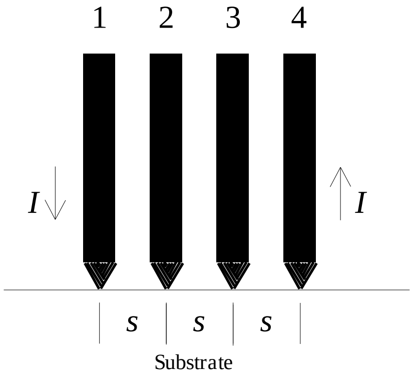


Fig. 51: Probe arrangement for a conventional 4-point probe

Of course, assuming equivalent boundary conditions for each probe, the voltage at any point within the volume is just the superposition, *i.e.*, the sum, of the voltages due to each probe separately. Thus, one can apply the formula for a single probe to obtain:

$$V = \frac{\rho I}{2\pi}\left(\frac{1}{r_1} - \frac{1}{r_4}\right)$$

Here, r_1 is identified as radial distance from probe 1 and r_4 as radial distance from probe 4 to some arbitrary point within the volume. In practice, potential is measured at probes

2 and 3. Furthermore, these measurements are made using a high impedance volt meter such that very little current (ideally no current) flows through probes 2 and 3. Therefore, any "IR drops" due to the probes themselves are insignificant and the voltage measured at probes 2 and 3 just corresponds to the preceding formula, hence:

$$V_2 = \frac{\rho I}{2\pi}\left(\frac{1}{s} - \frac{1}{2s}\right) \quad ; \quad V_3 = \frac{\rho I}{2\pi}\left(\frac{1}{2s} - \frac{1}{s}\right)$$

Clearly, for probe 2, r_1 is equal to s and r_4 is equal to $2s$. Conversely, for probe 3, these quantities are just reversed. Hence, the *voltage difference*, ΔV, between probes 2 and 3 is just the simple difference:

$$\Delta V = V_2 - V_3 = \frac{\rho I}{2\pi}\left(\frac{2}{s} - \frac{1}{s}\right) = \frac{\rho I}{2\pi s}$$

In practice, this formula is rearranged to solve for resistivity as follows:

$$\rho = 2\pi s\left(\frac{\Delta V}{I}\right)$$

Moreover, it is usual for I and s to be preset parameters and ΔV to be measured.

Of course, the preceding formula is applicable to determination of resistivity of a bulk substrate; however, if the conductive material cannot be considered as a semi-infinite volume, then this expression must be modified. Naturally, the case of a thin film represents the limit that thickness of conductive material, x_f. Even so, analysis of a probe on a thin film still remains similar to the case of a semi-infinite volume. Again, suppose that a single probe is injecting a current, I, but in this case into a thin film:

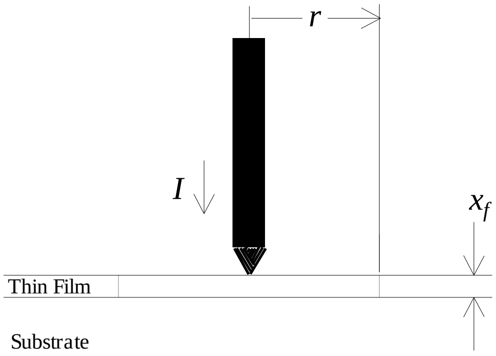


Fig. 52: Current injection into a thin film substrate from a single probe

Obviously, injected current diverges through a series of circular (or more correctly very short cylindrical) equipotential shells such that current density through each shell is $I/2\pi r x_f$. (Clearly, $2\pi r$ is just the circumference of a circle of radius, r, hence $2\pi r x_f$ is the surface area of a cylindrical element.) As before, electric field across each shell is just

$-\Delta V/\Delta r$, such that ΔV is voltage drop and Δr is shell thickness. (Again, a negative sign appears because voltage decreases as radius increases.) Clearly, if one applies Ohm's Law and allows shell thickness to tend toward zero, one obtains a simple formula:

$$\frac{dV}{dr} = -\rho\left(\frac{I}{2\pi r x_f}\right) = -R_s\left(\frac{I}{2\pi r}\right)$$

Here, the fundamental relationship between sheet resistance and resistivity has been formally substituted. Again, one trivially integrates to obtain:

$$V - V' = \frac{IR_s}{2\pi}\left(\ln\frac{1}{r} - \ln\frac{1}{r'}\right) = \frac{IR_s}{2\pi}(\ln r' - \ln r)$$

However, in contrast to the case of a semi-infinite volume, one cannot assume that voltage falls to zero at large distances since $\ln\infty$ does not vanish. Nevertheless, this is of no consequence, since it is only voltage differences that will actually be measured.

Naturally, one supposes that four probes are arranged on the surface just as before; with the additional requirement that film thickness, x_f, is negligible in comparison to probe spacing, s. Again, voltages due to each probe are superimposed and one finds that divergent terms corresponding to primed variables subtract out. Thus, for probes 1 and 4 "forcing" a current, I, it follows that:

$$V = \frac{IR_s}{2\pi}(\ln r_4 - \ln r_1)$$

As before, voltages are measured at probe 2 and probe 3. Of course, the values of r_1 and r_4 are just the same as in the semi-infinite case:

$$V_2 = \frac{IR_s}{2\pi}(\ln 2s - \ln s) \quad ; \quad V_3 = \frac{IR_s}{2\pi}(\ln s - \ln 2s)$$

It follows immediately that:

$$\Delta V = V_2 - V_3 = \frac{IR_s}{2\pi}(2\ln 2s - 2\ln s) = \frac{IR_s}{\pi}\ln 2$$

Therefore, upon rearrangement sheet resistance, R_s, is readily obtained:

$$R_s = \frac{\pi}{\ln 2}\left(\frac{\Delta V}{I}\right)$$

This formula appears frequently as an expression for sheet resistance. Obviously, the factor, $\pi/\ln 2$, is an irrational number which to ten decimal places has a value 4.532360142. One further observes the rather remarkable result that, in contrast to the case of a semi-infinite volume, probe spacing does not explicitly appear. Therefore, as long as the four probes are equally spaced and spacing is large in comparison to thin film thickness, for any forcing current, *I*, the voltage drop, Δ*V*, remains the same irrespective of *s*. Therefore, it is not surprising that sheet resistance measurements are inherently very reproducible.

Of course, it can happen that film thickness is not negligible in comparison to the probe spacing. In this "thick film" case, the simple one dimensional analysis appearing previously is not satisfactory. Consequently, to obtain correct results, one must apply a two dimensional analysis. This is complicated (requiring application of two dimensional Green's functions) and will not be considered further. However, without going into details, the result of such an analysis can be cast in terms of the semi-infinite formula and a correction factor, *G*:

$$\rho = 2\pi s\left(\frac{\Delta V}{I}\right)G$$

It is found that *G* has the explicit form:

$$G = \frac{u}{\ln\sinh 2u - \ln\sinh u} \qquad ; \qquad u = \frac{x_f}{2s}$$

Naturally, one must recover the thin film result in the limit that *u* tends to zero. Of course, it follows from the fundamental definition of sinh*u* that in this limit, sinh*u* is asymptotic to *u* itself. Hence, one just replaces sinh*u* in the formula for *G* with *u* to obtain:

$$\lim_{u\to 0} G = \frac{u}{\ln 2u - \ln u} = \frac{u}{\ln 2} = \frac{x_f}{2s\ln 2}$$

This immediately gives the required result:

$$\rho = 2\pi s\left(\frac{\Delta V}{I}\right)\lim_{x_f\to 0} G = \frac{\pi x_f}{\ln 2}\left(\frac{\Delta V}{I}\right)$$

Obviously, in the limit that film thickness tends toward infinity, *G* must tend to unity and the semi-infinite result recovered. Of course, for large values of *u*, hyperbolic sines can be replaced with exponentials, thus:

$$\lim_{u\to\infty} G = \frac{u}{\ln\frac{e^{2u}}{2} - \ln\frac{e^{u}}{2}} = \frac{u}{\ln e^{2u} - \ln e^{u}} = \frac{u}{2u-u} = 1$$

Indeed, *G* becomes unity as film thickness becomes indefinitely large.

In all of the preceding analysis, it has been assumed that the conductive material is infinite in extent in any direction parallel to the surface. In practical circumstances, this may not be the case. To treat these situations, *geometrical correction factors* have been determined. Such a "corrected" formula for sheet resistance has generic form:

$$R_s = \frac{\pi}{\ln 2} F_g \left(\frac{\Delta V}{I} \right)$$

In this case, F_g is identified as a geometrical correction factor. Again, without going into explicit details, it is found that for a circular substrate of finite diameter, d, measured exactly at the center, F_g has the form:

$$F_g = \frac{\ln 2}{\ln 2 + \ln\left(\frac{d^2}{s^2} + 3 \right) - \ln\left(\frac{d^2}{s^2} - 3 \right)}$$

Clearly, if the ratio of d to s is reasonably large, then F_g is close to unity and the uncorrected formula is recovered. (One can determine the required ratio of d/s for a particular prescribed measurement precision by careful analysis of the above formula.) This is easily satisfied for silicon wafers, but perhaps not for small irregular samples. Another significant geometrical configuration occurs if it is necessary to measure resistivity or sheet resistance near the edge of a sample. Again, an appropriate correction factor has been determined. In this case, if the probe array is placed perpendicular to the sample edge (assumed to be straight), then F_g is found to have the form:

$$F_g = \frac{1}{1 + \frac{1}{1 + 2d/s} - \frac{1}{2 + 2d/s} - \frac{1}{4 + 2d/s} + \frac{1}{5 + 2d/s}}$$

Similarly, if the probe array is placed parallel to the sample edge, then F_g is found to be:

$$F_g = \frac{1}{1 + \frac{2}{\sqrt{1 + (2d/s)^2}} - \frac{1}{\sqrt{1 + (d/s)^2}}}$$

Here, d is redefined as the distance from the probe array to the sample edge. Generally, if d is greater than $4s$, then these corrections are unnecessary. (Obviously, $4s$ is roughly the width of the probe head.) As might be expected, there are a number of additional corrections for orientation of the probe array and other geometrical shapes.

Another more useful method for obtaining accurate sheet resistance measurements on finite samples requires simultaneous application of two measurement configurations. In the "normal" configuration, as defined previously, current is forced through probes 1 and 4 and voltage is measured between probes 2 and 3. There is nothing physically unique about this probe configuration. In principle, one can force current between any two probes and measure voltage between the other two and extract a value for sheet resistance. Accordingly, one can obtain an independent measurement of sheet resistance by using an alternate probe configuration. The most common practice is inversion of two of the probes. In such a configuration, ΔV is measured between probes 2 and 4 and current is forced between probes 1 and 3. Applying a similar analysis as before, it is a simple matter to construct an expression for sheet resistance for this configuration as follows:

$$R_s = \frac{2\pi}{\ln 3}\left(\frac{\Delta V}{I}\right)$$

Assuming that actual sheet resistance is the same for both normal and alternate probe configurations, one can determine the correction factor electrically by measuring R_s in both alternate and normal configurations. Obviously, if the two measurements are the same, then the correction factor is unity. However, if they do not agree, F_g can be determined as a function of R_s^{norm}/R_s^{alt} such that R_s^{norm} and R_s^{alt} are sheet resistances measured in normal and alternate configurations respectively. Typically, F_g is represented as a truncated power series:

$$F_g = A_0 + A_1\left(\frac{R_s^{norm}}{R_s^{alt}}\right) - A_2\left(\frac{R_s^{norm}}{R_s^{alt}}\right)^2$$

Expansion coefficients, A_0, A_1, and A_2 are determined empirically. In practice, the equipment manufacturer incorporates determination of the correction factor directly into the hardware and software of modern probes and wafer sheet resistance mappers. Thus, an end user does not usually need to determine F_g explicitly.

Junction Depth Measurement

Typically, diffusion profiles are characterized by diffusing a desired dopant species into the surface of a "test wafer" that has some uniform background dopant concentration, C_B, of the opposite type. Clearly, a *pn*-junction is formed at some junction depth, x_J, below the wafer surface at which the diffused concentration profile, $C(x,t)$, exactly equals C_B. As a consequence of the electrical properties of a *pn*-junction, the diffused surface layer is essentially isolated electrically from the remainder of the underlying substrate. Hence, four point probe measurements can be applied directly to determine the sheet resistance of the doped layer. However, the diffusion profile cannot be characterized precisely by sheet resistance measurements alone, but if combined with a junction depth measurement, a reasonable estimate of diffusion profile, subject to specific processing conditions, *e.g.*, instantaneous or constant sources, can be made.

A classical method for measuring junction depth, requires milling a small area of the surface of a test wafer having the desired diffusion (and, thus, a known sheet resistance) using either a spherical (ball bevel) or cylindrical (groover) grinder. One then treats the exposed silicon with a chemical solution that "stains" regions of one extrinsic doping type, *i.e.*, either *p* or *n*-type silicon. (These stains typically consist of a solution of acids or bases and metal salts, which either preferentially etch or "plate out" on one doping type.) The position of the stained boundary can be determined using a calibrated optical microscope and then related to the actual junction depth by analysis of the original milling geometry. Alternatively, the wafer can be cleaved and stained and the junction depth measured directly using a scanning electron microscope (SEM). If one assumes a standard doping profile and provided that the product of diffusivity and processing time, *Dt*, is known, then junction depth, x_J, is related to surface doping concentration, C_s, as follows for a Gaussian profile:

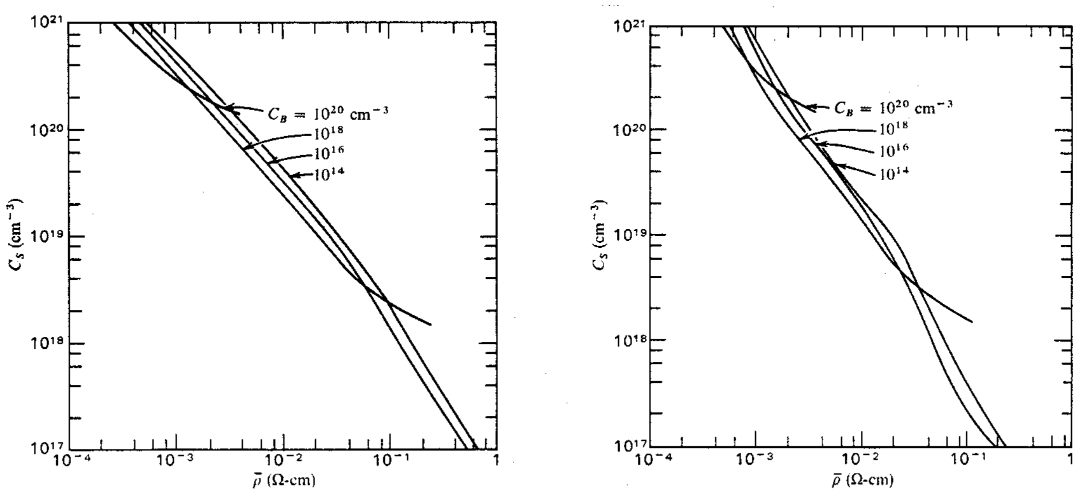
$$C_s = C_B e^{x_J^2 / 4Dt}$$

A corresponding expression can be constructed for a complementary error function profile:

$$C_s = \frac{C_B}{\operatorname{erfc}\left(x_J / 2\sqrt{Dt} \right)}$$

Of course, C_B is the background dopant concentration of opposite type into which the desired species has been diffused. Therefore, assuming a given mathematical form for the diffusion profile, junction depth alone determines surface doping concentration.

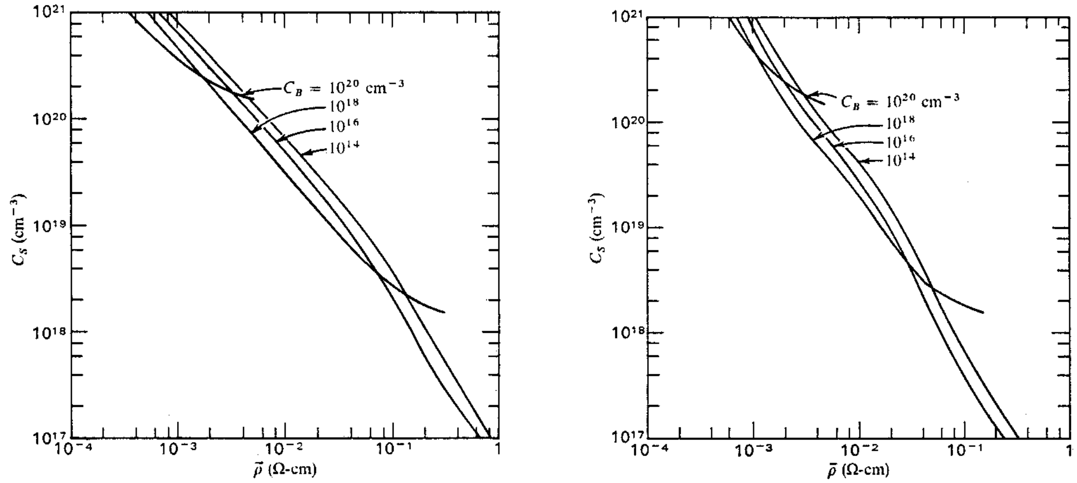
Alternatively, average resistivity, $\overline{\rho}$, can be identified as the product of sheet resistance and junction depth, $R_s x_J$. Indeed, for standard diffusion profiles, *viz.*, Gaussian and complementary error function profiles, $\overline{\rho}$ and C_s can be related directly for a specified C_B. These relationships are summarized graphically for both *p*-type and *n*-type diffusions as follows:



p-type diffusion

n-type diffusion

Fig. 53: Surface concentration versus average resistivity for a Gaussian profile



p-type diffusion *n*-type diffusion

Fig. 54: Surface concentration versus average resistivity for a complementary error function profile

Here, the relationship between $\overline{\rho}$, x_J, and C_s corresponds to the integral formula:

$$\overline{\rho} = \frac{x_J}{\int_0^{x_J} \frac{dx}{\rho(x)}}$$

Of course, $\rho(x)$ is directly related to the concentration profile and background doping. Therefore, in practice surface concentration can be determined separately by measuring R_s and x_J, constructing $\overline{\rho}$, and "reading off" C_s from the associated graph. However, it is clear that C_s can also be determined directly from a junction depth measurement without any prior knowledge of R_s or $\overline{\rho}$. Of course, these two results can agree only if the assumed concentration profile is substantially correct. (In this way, either Gaussian or complementary error function profiles can be confirmed.)

Shallow and Non-Ideal Concentration Profiles

Unfortunately, these classical techniques are really only applicable if junctions are reasonably deep. However, for state-of-the-art processes, junctions typically have become quite shallow making direct measurement of junction depth very difficult. Furthermore, even if the junction depth is not too shallow, if the concentration profile is non-ideal, *i.e.*, it is not of either Gaussian or complementary error function form, then incorrect or inconclusive results will be obtained. To remedy this, sophisticated analysis techniques have been developed that allow direct measurement of the concentration profile as a function of vertical depth from the wafer surface. The most accurate of these is *secondary ion mass spectrometry* (SIMS). In this method, the surface of the wafer is controllably sputtered away in a vacuum chamber using energetic argon ions. The sputtered material is collected and mass analyzed. The amount of each atomic species present within the sample can be quantitatively related to observed peak heights in the mass spectrum. Furthermore, detection limits in SIMS analysis are well below what is necessary for analysis of typical dopant concentrations. In practice, "mass signal" is observed as a function of sputtering time, thus the concentration profile of a diffused species (*i.e.,* a shallow level dopant) is easily obtained by relating sputtering time to depth in the wafer.

For routine analysis, SIMS may be too expensive and time consuming. However, *spreading resistance* or *incremental sheet resistance* measurements can also be used to determine concentration profile. In contrast to SIMS, which can detect any atomic species, these techniques can only detect the net concentration of electrically active extrinsic dopants. To make spreading resistance measurements, the surface of the diffused wafer is milled at a very shallow angle so that at the deepest point the milled surface extends well below the junction. Following this procedure, two point probes are moved down the incline and the resistance is measured as a function of position. This resistance varies in a way that is directly related to the concentration profile. Incremental sheet resistance is a very similar technique in which incremental amounts of the wafer surface are successively removed by etching or milling. Sheet resistance is measured between each etching or milling step using a standard four point probe. Again, the measured sheet resistance varies as a function of the total amount of surface material removed in a way that can be directly related to the concentration profile. Both of these techniques are only semi-quantitative due to difficulties involved controlling the milling angle and/or milling or etch rates. Typically, spreading resistance or incremental sheet resistance measurements are calibrated using SIMS.

Electrical Behavior of *pn*-Junctions

To consider what happens electrically in a *pn*-junction, one can, again, consider a thought experiment. Suppose that two blocks of semiconductor of opposite extrinsic doping, *i.e.*, one block is *n*-type and one is *p*-type, are initially separated widely. If these blocks are then joined, a net transfer of carriers from one block to the other must occur in order to establish equilibrium. Obviously, this situation is similar to the "simpler" cases of a metal-semiconductor contact and an MOS capacitor. Furthermore, just as in these previous cases, if the whole system is to be in equilibrium, then the Fermi level must be constant throughout the entire combined volume of the two blocks. It comes as no surprise that this requires that valence and conduction bands must become bent in the region of the junction. This situation is illustrated below:

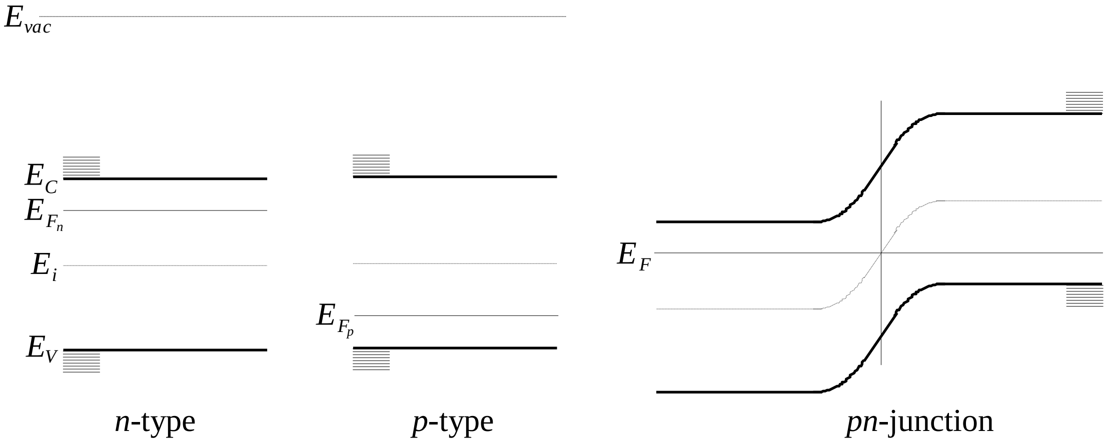


Fig. 55: Band diagrams for an unbiased *pn*-junction

The position of the junction is determined as the point that the intrinsic Fermi level and the actual Fermi level intersect. Of course, this defines the exact position at which the semiconductor changes from one type to the other, *i.e.*, it is intrinsic at the junction. Moreover, even though this structure is hypothetically constructed by joining separate blocks of material, the physical mechanism used to form a *pn*-junction is immaterial to its electrical behavior. Obviously, in practice dopant diffusion processes can be used to form *pn*-junctions intentionally.

An Unbiased *pn*-Junction

A clear consequence of the existence of a junction is the existence of an internal or intrinsic electric field in the junction region. Returning to the previous thought experiment, one can easily see the cause of this behavior. Suppose that the two opposite type semiconductor blocks are brought together suddenly. At the instant of initial contact, mobile carriers are not in equilibrium since the Fermi levels do not coincide, *i.e.*, carrier concentrations are constant right up to the interface between the two blocks. Furthermore, ignoring the effects of non-equilibrium generation-recombination processes, electron and hole fluxes, F_e and F_h, everywhere within the joined blocks of semiconductor (including the region near the junction) are governed by linear transport equations of the form:

$$F_e = -D_e \frac{\partial n}{\partial x} - n\mu_e E$$

$$F_h = -D_h \frac{\partial p}{\partial x} + p\mu_h E$$

Here, D_x and μ_x are carrier diffusivity and mobility (x is either e or h, thus, denoting either electrons or holes), and E is electric field. Clearly, the first term on the right hand side of both of these expressions comes from Fick's Law and describes *diffusion* of mobile carriers due to a concentration gradient. Similarly, the second term comes from Ohm's Law and describes *drift* of mobile carriers under the influence of an electric field, E. Obviously, the change of sign of the drift terms is due to opposite charges of electrons and holes. In contrast, there is no sign change for diffusion terms since diffusion is independent of carrier electrical charge. (These equations illustrate coupling of different thermodynamic forces with the same transport flux.) Of course, corresponding current densities, j_e and j_h, are trivially obtained from fluxes simply by multiplying by carrier charge:

$$j_e = qD_e \frac{\partial n}{\partial x} + qn\mu_e E$$

$$j_h = -qD_h \frac{\partial p}{\partial x} + qp\mu_h E$$

Clearly, the ohmic terms have the expected form of electric field divided by resistivity.

Within this context, it is clear that immediately when the two blocks touch, majority carriers from each side must diffuse across the junction since a large concentration gradient exists. (Of course, when majority carriers diffuse to the other side of the junction, they then become minority carriers since, by definition, the semiconductor type changes at the junction.) Initially, there is no drift since there is no intrinsic electric field. However, carrier mass action equilibrium requires that a large non-equilibrium excess concentration of minority carriers cannot be built up. Therefore, excess minority carriers rapidly recombine with majority carriers resulting in the depletion of majority carriers in both types of semiconductor in the vicinity of the junction. Obviously, depletion is greatest at the junction and decreases as distance from the junction increases. Furthermore, just as in the case of an MOS capacitor or metal-semiconductor contact, depletion also causes ionized impurity atoms to become uncovered. This implies the existence of a space charge within the depletion region around the junction. Naturally, an electric field must exist in this region of space charge. (One should note that the terms "space charge region" and "depletion region" are synonymous and describe different aspects of the same physical phenomenon.) Obviously, in the thought experiment the depletion region begins to form just as soon as the two blocks of semiconductor touch, however, this process cannot continue indefinitely. Returning to the transport equations, one observes that as the space charge increases, the strength of the associated internal electric field must also increase. Therefore, the magnitude of the drift component of net carrier current increases in response to the build up of an internal electric field. Furthermore, the drift current naturally opposes the diffusion current. Thus, the internal electric field can increase only until drift and diffusion currents become exactly equal and opposite, *i.e.*, net carrier current vanishes. At this point, carrier equilibrium is established in the region of the junction as well as elsewhere throughout the volume of the semiconductor and no further net carrier transport occurs.

The electrostatic characteristics of an unbiased *pn*-junction are illustrated in the following figure:

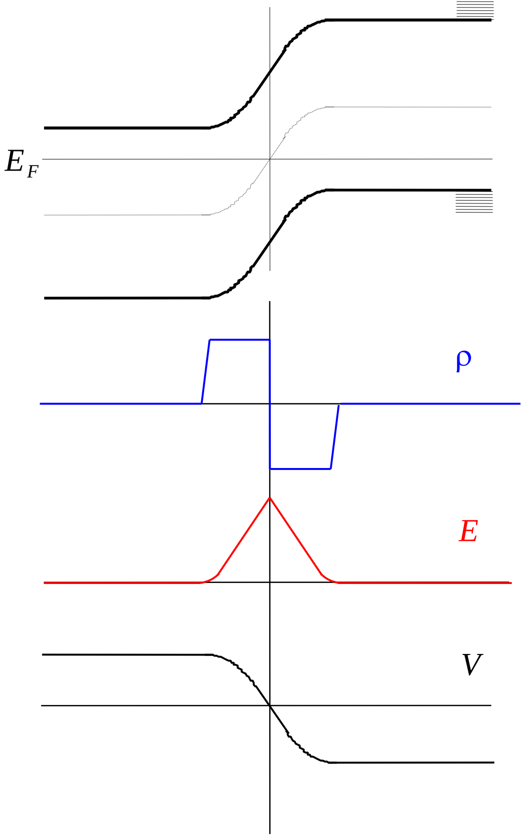


Fig. 56: Electrostatic characteristics of an unbiased *pn*-junction

It is clear from the band diagram for an ideal abrupt *pn*-junction, that the space charge density, ρ, in the junction is dipolar and abruptly changes sign from positive on the *n*-type side to negative on the *p*-type side. Of course, Maxwell's equation relating electric field, *E*, to charge density is:

$$\frac{\partial E}{\partial x} = \frac{\rho}{\varepsilon}$$

Thus, the electric field is in the same direction (defined as positive to the right) on both sides of the junction, *i.e.*, E "points" from n-type to p-type. Obviously, E reaches a maximum value exactly at the junction. Of course, the existence of the internal field implies a "built-in" or *diffusion potential*, V:

$$\frac{\partial V}{\partial x} = -E$$

$$\frac{\partial^2 V}{\partial x^2} = -\frac{\rho}{\varepsilon}$$

Clearly, V is more positive on the n-type side of the junction.

It should be noted in passing, that in the preceding figure, net doping levels on each side of the junction are represented as approximately equal in magnitude. In actual practice, this is not usually the case. Typically, one doping concentration will be much larger that the other, but the total amount of uncovered positive and negative charges in the junction depletion region must be the same irrespective of doping. Hence, in such a case the junction must be asymmetric with the width of the depletion layer on the lightly doped side being much larger than on the heavily doped side. However, this does not substantially change the situation. Similarly, diffused junctions are not abrupt but, generally are "graded" (*i.e.*, the net doping concentration decreases smoothly from some higher concentration far from the junction, to a lower concentration in the junction region). Again, this does not greatly change the behavior of a *pn*-junction (although graded junctions do exhibit some behavioral characteristics that are different than those of abrupt junctions).

Naturally, Fermi potentials, φ_{Fn} and φ_{Fp} can be defined on each side of the junction, and far away from the junction region (*i.e.*, far outside the depletion layer), one can write:

$$\varphi_{Fn} = \frac{E_F - E_{in}}{q} = \frac{kT}{q} \ln \frac{N_n + \sqrt{N_n^2 + 4n_i^2}}{2n_i}$$

$$\varphi_{Fp} = \frac{E_{ip} - E_F}{q} = \frac{kT}{q} \ln \frac{N_p + \sqrt{N_p^2 + 4n_i^2}}{2n_i}$$

Here, N_n and N_p are net donor and net acceptor concentrations on each side of the junction and E_{in} and E_{ip} are corresponding intrinsic Fermi energies. (Recall that the Fermi potential, φ_F, is defined for an extrinsically doped semiconductor as $|E_i - E_F|/q$.) Clearly, the maximum value of the diffusion potential, V_{pn}, is just the sum of n-type and p-type Fermi potentials, hence:

$$V_{pn} = \varphi_{Fn} + \varphi_{Fp} = \frac{kT}{q} \ln \frac{N_n + \sqrt{N_n^2 + 4n_i^2}}{2n_i} + \frac{kT}{q} \ln \frac{N_p + \sqrt{N_p^2 + 4n_i^2}}{2n_i}$$

Naturally, this can be simplified if extrinsic doping levels far exceed the intrinsic carrier concentration:

$$V_{pn} = \frac{kT}{q} \ln \frac{N_n N_p}{n_i^2}$$

Obviously, V_{pn} is analogous to the contact potential defined in the case of a simple metal-metal junction (*i.e.*, contact). Of course, it arises from the effective work function difference between extrinsic *n*-type and *p*-type semiconductor. However, since electric fields can penetrate semiconductors, in contrast to a metal-metal junction, a depletion layer of measurable thickness appears at a *pn*-junction and the potential difference is "stretched out" over a much larger volume of material. Depending on actual levels of extrinsic doping, for silicon V_{pn} is typically in the neighborhood of 0.6-0.7 volts.

The Effect of Forward and Reverse Bias on a *pn*-Junction

So far, only an unbiased *pn*-junction has been considered, however, if a voltage is applied across the junction, then one expects some current to flow. Within this context, the internal potential acts as a barrier to current flow from the *p*-type side to the *n*-type side of the junction. Hence, the existence of the diffusion potential indicates that a *pn*-junction can be expected to conduct electrical current through the junction more easily in one direction than the other. Indeed this is found to be the case and such a device is called a *semiconductor diode*. (For this reason, V_{pn} is sometimes called "diode potential" or "diode drop" or V_{BE}.)

In *forward bias*, the *p*-type side of the junction is held positive with respect to the *n*-type side. Thus, the applied field opposes the internal field. This reduces the drift component of the carrier flux and results in a higher net flux of majority carriers toward the junction from each side due to diffusion. Thus, the depletion region shrinks in forward bias. (However, in principle it cannot ever be completely removed even by a very large forward bias.) Of course, when majority carriers reach the junction, they recombine which (since they are of opposite charge) corresponds to a net current flow through the junction. Clearly, only a limited amount of current can flow before the applied potential substantially offsets the diffusion potential. At this point, the junction "turns on" and current flows very easily. In contrast, in the case of *reverse bias*, the *n*-type side of the junction is held positive with respect to the *p*-type side. Hence, the applied potential enhances the diffusion potential. This further reduces the net flux of majority carriers into the vicinity of the junction and, therefore, increases the size of the depletion region. Thus, current flowing through the junction is greatly reduced. However, it does not fall all the way to zero. This is a result of minority carrier generation. Of course, there is always generation and recombination of electrons and holes due random thermal fluctuations. (This is just the source of the intrinsic carrier concentration.) This process occurs even in the depletion region. Therefore, in reverse bias, minority carriers spontaneously generated by thermal excitation in or very near the depletion region are swept across the junction by the field and, as such, are said to be "injected" by the field. This leads to the appearance of a small reverse current through the junction. Since there is no potential barrier to the flow of minority carriers, the reverse current should have a simple constant characteristic unaffected by any applied potential since it is fixed solely by the generation-recombination rate in the vicinity of the junction.

The preceding description of current flow through a biased *pn*-junction can be cast in a mathematical form if one considers diffusion (or recombination) and drift (or generation) fluxes and/or current densities through the junction. Suppose one applies an external potential, *V*, to the junction, such that *V* is defined as positive in forward bias and negative in reverse bias. One recalls that the energy distribution for mobile carriers is governed by Fermi-Dirac statistics, however, as is usual one assumes that it is allowable to use an approximate Maxwell-Boltzmann form. Thus, diffusion fluxes, F_{ed} and F_{hd}, for electrons and holes, respectively, can be written as follows:

$$F_{ed} = F_{e0} e^{q(V-V_{pn})/kT} \quad ; \quad F_{hd} = F_{h0} e^{q(V-V_{pn})/kT}$$

Clearly, the exponential factor accounts for the fraction of electrons and holes having sufficient energy to surmount the potential barrier associated with the junction. Of course, the net diffusion flux of electrons is toward the *p*-type side of the junction (*i.e.*, to the right with respect to the preceding figure) and the net diffusion flux of holes is toward the *n*-type side of the junction (*i.e.*, to the left). Consequently, if the junction is unbiased, the diffusion fluxes just correspond to the expressions:

$$F_{ed}(0) = F_{e0} e^{-qV_{pn}/kT} \quad ; \quad F_{hd}(0) = F_{h0} e^{-qV_{pn}/kT}$$

However, since there is no net flow of carriers across an unbiased junction, drift flux due to thermal generation of carriers within the depletion region and diffusion flux from the neutral volume are exactly equal and opposite; hence, one finds that:

$$F_{eg} = -F_{e0} e^{-qV_{pn}/kT} \quad ; \quad F_{hg} = -F_{h0} e^{-qV_{pn}/kT}$$

Here, F_{eg} and F_{hg} are identified as electron and hole generation fluxes. As observed previously, generation fluxes are not affected by the applied potential, thus net fluxes of electrons and holes are obtained by combining drift and diffusion fluxes as follows:

$$F_e = -F_{eg}\left(e^{qV/kT} - 1\right) \quad ; \quad F_h = -F_{hg}\left(e^{qV/kT} - 1\right)$$

As usual, carrier fluxes are converted to current densities simply by multiplying by carrier charge. In addition, signs will be rationalized so that current is considered positive when flowing from the *p*-type side of the junction to the *n*-type side. One finds that:

$$j_e = j_{e0}\left(e^{qV/kT} - 1\right) \quad ; \quad j_h = j_{h0}\left(e^{qV/kT} - 1\right)$$

Here, j_{e0} and j_{h0} are the magnitudes of electron and hole generation current densities. Of course, total current density through the junction just corresponds to the simple sum:

$$j = j_0\left(e^{qV/kT} - 1\right)$$

This is the *Shockley* or *ideal diode equation* and is a fundamental characteristic of *pn*-junctions. Obviously, j_0 is the total generation current density due to both electrons and holes. For practical diodes, the diode equation is trivially rendered into absolute current rather than current density by consideration of appropriate geometrical parameters, hence:

$$I = I_0\left(e^{qV/kT} - 1\right)$$

Thus, one finds that a *pn*-junction has an exponential current-voltage characteristic:

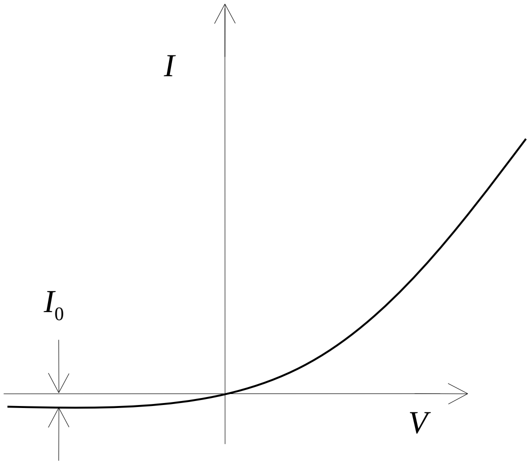


Fig. 57: Current-voltage characteristic of a biased *pn*-junction

Here, I_0 is called *saturated reverse current*. In common terminology, if the value of I_0 becomes too large, then the *pn*-junction is said to be "leaky". The primary cause of this problem is the enhancement of minority carrier generation due to contamination and/or defects.

In practice, the diode equation is commonly modified by inclusion of an empirical *ideality factor, n,* as follows:

$$j = j_0\left(e^{qV/nkT} - 1\right)$$

Obviously for some specific device this takes the form:

$$I = I_0\left(e^{qV/nkT} - 1\right)$$

In general, *n* is a fitting parameter that allows for measured departure of diode IV characteristics from the ideal Shockley equation. Physically, the ideality factor is related to geometry of the junction and distribution of electron-hole recombination with respect to the associated depletion region. Moreover, for any junction that approximates an infinite planar junction, recombination is generally negligible in the depletion region and, consequently, *n* can be expected to be very close to 1. Conversely, if geometry is non-ideal and recombination in the depletion region dominates then *n* can be expected to be near 2.

Within this context, irrespective of ideality the diode equation does not account for any "series" resistance due to the bulk semiconductor itself. Accordingly, in very high forward bias in which case current may become very large, the exponential characteristic can be expected to become combined with a linear characteristic due to IR drop. Furthermore, if either forward or reverse bias becomes too large, then the *pn*-junction "breaks down" due to a high electric field in the junction region. This behavior is similar to the case of oxide break down. In simplistic terms, carriers are accelerated so rapidly by the applied field that they collide with bound electrons of lattice atoms and cause *impact ionization*. This results in a chain reaction of carrier generation or an *avalanche*. As might be expected, this can be destructive, permanently degrading the junction. (Even so, there are devices such as *Zener diodes* that do operate in or near junction breakdown.)

Capacitance-Voltage Response of a *pn*-Junction

It is clear that there is net charge storage associated with a *pn*-junction. Physically, this takes the form of two oppositely charged space charge layers. Hence, to a good approximation a *pn*-junction behaves similar to a classical parallel plate capacitor. Therefore, just as in the case of an MOS capacitor, one can apply the depletion approximation to determine the potential in the space charge region. Accordingly, one easily adapts the general expression obtained for MOS capacitors, thus:

$$\psi_n(x_n) = \frac{q}{\varepsilon_s}\int_{x_n}^{x_{dn}} dx'(x_n - x')N_n(x') \quad ; \quad \psi_p(x_p) = \frac{q}{\varepsilon_s}\int_{x_p}^{x_{dp}} dx'(x_p - x')N_p(x')$$

Here, $\psi_p(x_p)$ and $\psi_n(x_n)$ are internal potentials defined respectively on the *p*-type and *n*-type sides of the junction. The coordinates, x_p and x_n, are distances measured from the junction boundary. (Obviously, the "sign sense" of x_p and x_n must be opposite.) By definition, x_{dp} and x_{dn} are widths of *p*-type and *n*-type space charge layers. (Again, one should note that the space charge layer on the *p*-type side is negative and the space charge layer on the *n*-type side is positive.) Clearly, the total depletion layer width is just the sum of x_{dp} and x_{dn}. For generality, net doping densities are assumed to be position dependent in the above equations. Thus, these expressions can be applied to the more general case of a graded junction; however, for an abrupt junction the net doping densities are constant on each side of the junction and, therefore, can be simplified as follows:

$$\psi_n(x_n) = \frac{q}{2\varepsilon_s} N_n (x_{dn} - x_n)^2 \quad ; \quad \psi_p(x_p) = \frac{q}{2\varepsilon_s} N_p (x_{dp} - x_p)^2$$

By definition the potentials vanish at the edges of the depletion region.

In the simple case of a symmetrically doped *pn*-junction, *i.e.*, N_n and N_p equal, and in the absence of any external bias, $\psi_p(0)$ and $\psi_n(0)$ must exactly equal the Fermi potentials, φ_{Fp} and φ_{Fn}, which are, in fact, precisely equal, hence:

$$\varphi_F = \frac{q}{2\varepsilon_s} N_d x_{0d}^2$$

Here, φ_F, N_d, and x_{0d} are, respectively, Fermi potential, net dopant concentration (either acceptors or donors), and space charge width on one side of an unbiased symmetric junction. Clearly, for this simple case one can use explicit expressions for the Fermi potentials to determine unbiased depletion widths as follows:

$$x_{0d} = \sqrt{\frac{2\varepsilon_s kT}{q^2 N_d} \ln \frac{N_d + \sqrt{N_d^2 + 4n_i^2}}{2n_i}}$$

Of course, the total space charge width, x_{0tot}, is the sum of space charge widths on each side of the junction, which in this simple case is just $2x_{0d}$. Moreover, if extrinsic doping dominates as is usual, then the preceding formula can be simplified, thus:

$$x_{0d} = \sqrt{\frac{2\varepsilon_s kT}{q^2 N_d} \ln \frac{N_d}{n_i}}$$

It is evident that space charge width becomes smaller if net doping increases and conversely, becomes larger if net doping is decreased. Furthermore, one finds that these expressions are very similar to the corresponding expression for the maximum depletion width of an MOS capacitor.

The situation becomes more complicated for an asymmetrically doped junction. In this case, φ_{Fp} and φ_{Fn} are not precisely equal to $\psi_p(0)$ and $\psi_n(0)$, but satisfy the weaker condition that the internal electrical potential must be continuous across the junction. Nevertheless, the diffusion potential must just be the sum of $\psi_p(0)$ and $\psi_n(0)$ and, thus, one can write:

$$V_{pn} = \frac{q}{2\varepsilon_s}(N_n x_{0dn}^2 + N_p x_{0dp}^2)$$

Of course, V_{pn} is also the sum of the Fermi potentials, but this relation does not imply that $\psi_p(0)$ and $\psi_n(0)$ are equal to φ_{Fp} and φ_{Fn}, respectively. Clearly, they may differ by some compensating potential offset. (Indeed, this offset vanishes only if junction doping is symmetric.) In addition, the total uncovered charge per unit area on each side of the junction must be of opposite sign, but equal magnitude, Q_0, hence:

$$Q_0 = qN_p x_{0dp} = qN_n x_{0dn}$$

These two relations can be combined to construct expressions relating space charge widths to diffusion potential for both sides of an asymmetrically doped junction, thus:

$$V_{pn} = \frac{q}{2\varepsilon_s}\left(N_n + \frac{N_n^2}{N_p}\right)x_{0dn}^2 \quad ; \quad V_{pn} = \frac{q}{2\varepsilon_s}\left(\frac{N_p^2}{N_n} + N_p\right)x_{0dp}^2$$

It is a simple matter to rearrange these expressions to obtain the conventional identities:

$$x_{0dn} = \sqrt{\frac{2\varepsilon_s V_{pn}}{qN_n}\left(\frac{N_p}{N_p + N_n}\right)} \quad ; \quad x_{0dp} = \sqrt{\frac{2\varepsilon_s V_{pn}}{qN_p}\left(\frac{N_n}{N_p + N_n}\right)}$$

Naturally, the total space charge width, x_{0tot}, is, again, merely the sum of space charge widths on each side of the junction:

$$x_{0tot} = x_{0dn} + x_{0dp} = \sqrt{\frac{2\varepsilon_s V_{pn}}{q(N_p + N_n)}}\left(\sqrt{\frac{N_n}{N_p}} + \sqrt{\frac{N_p}{N_n}}\right) = \sqrt{\frac{2\varepsilon_s V_{pn}(N_p + N_n)}{qN_p N_n}}$$

Of course, the diffusion potential is also a function of dopant concentrations. Therefore, the explicit expression for V_{pn} obtained previously can be substituted explicitly such that:

$$x_{0tot} = \sqrt{\frac{2\varepsilon_s kT(N_p + N_n)}{q^2 N_p N_n}\ln\frac{N_n N_p}{n_i^2}}$$

Upon inspection of the preceding expressions, it is clear that if doping asymmetry is large, essentially all of the space charge region will be located on the lightly doped side of the junction and, consequently, x_{0tot} will be approximately equal to just the depletion width of the lightly doped side alone.

Similar considerations apply to a biased *pn*-junction. Again, the internal potential is identified as the sum of $\psi_p(0)$ and $\psi_n(0)$:

$$V_{pn} - V = \frac{q}{2\varepsilon_s}(N_n x_{dn}^2 + N_p x_{dp}^2)$$

However, the total potential is a combination of diffusion potential and external bias, *V*. Of course, the combination is written as a formal difference because by convention a forward bias opposes the diffusion potential and reverse bias enhances it. Thus, one observes that as expected, for any reverse bias, the total depletion layer width increases in comparison to the unbiased case. Conversely, for forward bias, total depletion layer width can be expected to decrease. This is clearly the case if the external bias is less than the diffusion potential. However, forward bias is increased beyond V_{pn}, negative values are obtained from the preceding formula. Clearly, this situation is unphysical since it implies that at least one of the space charge widths must have an imaginary value. However, this is of no real consequence and just reflects the inapplicability of the depletion approximation in the case of a large forward bias and associated very thin space charge layers.

Naturally, one can define the positive quantity, *Q*, as the magnitude of total charge per unit area of the junction stored in each space charge layer (*i.e.*, the total charge per unit area stored in the positive space charge region, *i.e.*, on the *n*-type side of the junction, is *Q*, and the total charge per unit area stored in the negative space charge region, *i.e.*, on the *p*-type side of the junction, is −*Q*). Clearly, just as for an unbiased junction, *Q* can be related to net concentrations and space charge widths as follows:

$$Q = qN_p x_{dp} = qN_n x_{dn}$$

One can recast the previous expression for $V_{pn}-V$ in terms of *Q* thus:

$$V_{pn} - V = \frac{Q^2}{2q\varepsilon_s}\left(\frac{1}{N_n} + \frac{1}{N_p}\right) = \frac{Q^2}{2q\varepsilon_s}\left(\frac{N_p + N_n}{N_p N_n}\right)$$

This equation can now be rearranged to give an explicit charge-voltage relation:

$$Q = \sqrt{2q\varepsilon_s (V_{pn} - V)\left(\frac{N_p N_n}{N_p + N_n}\right)}$$

Capacitance per unit area of a *pn*-junction is evidently obtained as the as the formal derivative of *Q* with respect to −*V*. (The derivative must be taken with respect to −*V* because the capacitor, *i.e.*, junction, is charged by a reverse bias and is discharged by a forward bias.)

$$C = \frac{dQ}{d(-V)} = \frac{1}{2}\sqrt{\frac{2q\varepsilon_s}{V_{pn} - V}\left(\frac{N_p N_n}{N_p + N_n}\right)}$$

Accordingly, one obtains the standard form of the capacitance-voltage relation for a *pn*-junction:

$$V_{pn} - V = \frac{q\varepsilon_s}{2C^2}\left(\frac{N_p N_n}{N_p + N_n}\right)$$

Obviously, *pn*-junction capacitance can be measured accurately only in reverse bias since in forward bias very large currents flow. Therefore, for convenience the reciprocal of the square of capacitance is plotted versus the negative of the bias voltage:

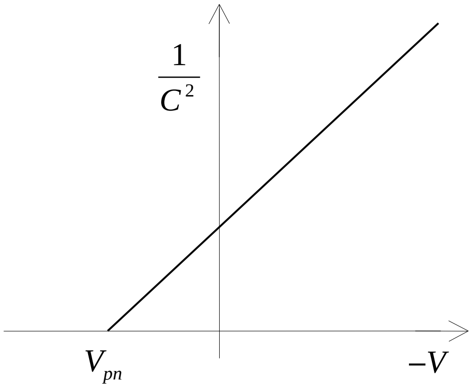


Fig. 58: Capacitance-voltage characteristic of a biased *pn*-junction

In this case, a linear plot is obtained. Moreover, if one extrapolates back to the abscissa intercept, it is clear that this should just correspond to the diffusion potential. Furthermore, the slope of the plot is related to net dopant concentrations on each side of the junction. (In particular, for an asymmetric junction the slope is approximately proportional to the net dopant concentration on the more lightly doped side.)

Relationship of Low Field Mobility to Carrier Diffusivity

The behavior of a *pn*-junction has been determined in terms of diffusion and drift of carriers in the junction region. Of course, diffusion is characterized by carrier diffusivity, D_x, such that x denotes either electrons or holes. Physically, within a very general phenomenological context, carrier diffusivity can be related to an internal "dynamic friction" of the medium, f_x, thus:

$$D_x = \frac{kT}{f_x}$$

This expression is known as the *Nernst-Einstein relation*. Of course, in a "resistive" medium by definition dynamic friction relates the velocity of a particle, *v*, to some applied force, F:

$$v = \frac{\mathrm{F}}{f_x}$$

Clearly, if friction is increased, then as one expects, particle velocity is decreased. Furthermore, in the absence of friction, *i.e.*, f_x vanishes, v diverges to infinity. This is just a consequence of Newton's Second Law. Classically, if no friction opposes the applied force, the particle is accelerated indefinitely and velocity increases without limit.

In the case of carrier drift, the applied force just arises from the electrostatic field, E, and hence, can be identified as $\pm qE$, (such that the upper sign corresponds to holes and the lower sign to electrons).

$$v = \frac{\pm qE}{f_x}$$

Obviously, mobility, μ_x, is trivially identified as q/f_x. This just follows from the elementary definition of mobility. Therefore, it immediately follows from the Nernst-Einstein relation that:

$$\mu_x = \frac{qD_x}{kT}$$

Here, a fundamental relationship is found between carrier mobility and diffusivity. However, this equation is applicable only in the limit of a low field for which the drift velocity of carriers is smaller than average thermal velocity. If the field becomes too large, this condition is no longer satisfied and drift velocity saturates as a function of applied field.

The Photovoltaic Effect

In addition to conventional solid-state electronics, single crystal silicon is widely used for harvest of solar energy. This is a result of the *photovoltaic effect*, which can be understood by considering light absorption in a semiconductor material. As noted previously semiconductors may be classified as having either a direct or an indirect band gap depending on whether an electron can be promoted directly from the valence band to the conduction band without interaction with the lattice, *i.e.*, with phonons. Although this distinction is important for construction of light emitting devices such as light emitting diodes or semiconductor lasers, both direct and indirect band gap semiconductors absorb photons having energy larger than the band gap. Physically, absorption of light results in formation of hole-electron pairs in excess of equilibrium thermal generation. Consequently, in analogy to the application of an electrical bias to a *pn*-junction, illumination of a junction results in non-equilibrium conditions. This can be represented mathematically by addition of photocurrent density, j_L, to the usual diode current:

$$j = j_L - j_0\left(e^{qV/kT} - 1\right)$$

Of course, for a practical photovoltaic device, this expression takes the form:

$$I = I_L - I_0\left(e^{qV/kT} - 1\right)$$

These are "ideal" photodiode equations. Moreover, for conceptual convenience in both of these expressions, the sense of current direction is formally inverted with respect to the original Shockley diode equation; hence, I_L is photocurrent characteristic of some definite illumination condition. Clearly, the effect of photocurrent is simply to shift the current-voltage characteristic of a *pn*-junction diode as illustrated in the following figure:

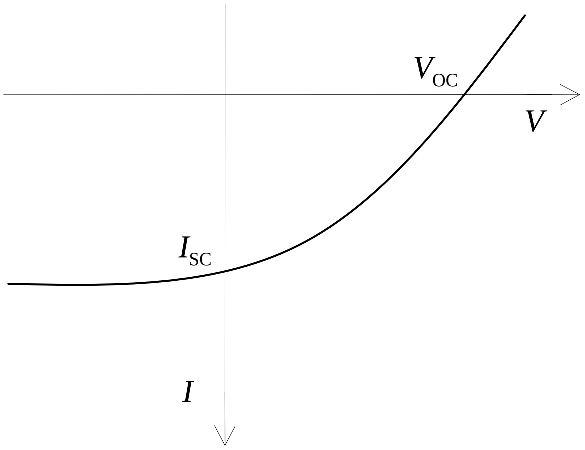


Fig. 59: The photovoltaic effect

By definition, I_{SC} is "short circuit" current, which flows through an illuminated junction when both sides are held at the same potential and, as such, is ideally the same as the photocurrent, I_L. Similarly, V_{OC} is "open circuit" voltage and is the potential measured across the junction when no current is allowed to flow. Moreover, it is clear that photocurrent generated by absorption of light normally flows out of the *p*-type side of the junction and into the *n*-type side.

Solar Cells

In general, a simple *pn*-junction does not provide the most favorable structure for harvest of solar energy. This is because photocurrent is collected almost exclusively from the depletion region, which is, typically, only a few microns thick at most. Therefore, it is advantageous to place a region of intrinsic silicon between *p*-type and *n*-type sides of the junction. Such a structure is called a *pin-junction*. As a practical matter, this region is typically 150 to 250 microns thick. Accordingly, light is efficiently collected throughout the entire intrinsic volume, which, naturally, is effectively depleted. It should be evident that it is highly desirable for the carrier recombination rate to be low, which, represents a fundamental limitation of the *efficiency* of a semiconductor photovoltaic device. Indeed, for silicon solar cells the maximum theoretical efficiency corresponding to the *Shockley-Queisser limit* is about 30%. The best practical devices (backside contact single crystal silicon with sophisticated anti-reflection technology) have efficiencies approaching 24-25%.

As is often convenient, complex electronic devices can be satisfactorily modeled as combinations of simple devices arranged in an "equivalent circuit". Accordingly, practical silicon solar cell may be represented schematically as the following equivalent circuit:

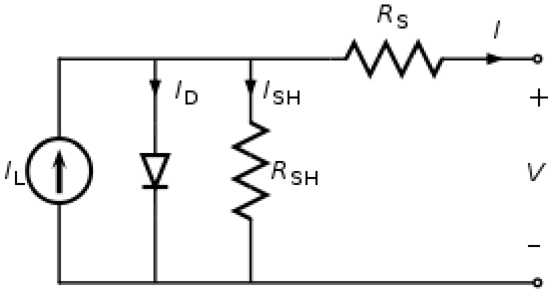


Fig. 60: Equivalent circuit of a silicon solar cell

Here, the current source (circle with arrow inside) just represents the photocurrent, I_L, as defined above. Likewise, I_D is the diode current as defined by the Shockley diode equation. However, in addition parasitic resistances, R_S and R_{SH}, are also included. These are known, respectively, as *series resistance* and *shunt resistance*. Both are undesirable, and it would seem obvious that series resistance should be as low as possible since it impedes the flow of current to the outside world. In contrast, shunt resistance should be as high as possible since it "shorts out" the device. Physically, series resistance generally arises in non-ideal connections between the silicon substrate and external wiring. This can be a result of either poor design or poor fabrication processes. Concomitantly, shunt resistance is caused by internal losses within the cell and is commonly the result of poor diode characteristics. Typically, low shunt resistance is the result of defective manufacturing rather than poor design. In any case, IV characteristics of a practical solar cell are readily represented by modifying the ideal expression as follows:

$$I = I_L - I_0\left(e^{q(V+IR_S)/kT} - 1\right) - \frac{V + IR_S}{R_{SH}}$$

Here, *V* is the potential difference generated across the output terminals of the device due to illumination. Unfortunately, because *I* appears explicitly in the exponent, this equation cannot be solved in closed form. Of course, by definition, V_{OC} implies that *I* is zero, hence:

$$\frac{V_{OC}}{R_{SH}} = I_L - I_0\left(e^{qV_{OC}/kT} - 1\right)$$

If, as is desirable, shunt resistance is large, then the left hand side of this expression can be replaced by zero and one obtains the ideal formula:

$$V_{OC} \cong \frac{kT}{q}\ln\left(\frac{I_L}{I_0} + 1\right)$$

At reasonable illumination, *viz.*, "one sun", I_L is generally much larger than I_0 and "1" can be ignored in the logarithm. Typically, for a crystalline silicon solar cell V_{OC} has a value similar to junction diffusion potential, *i.e.*, 500 to 700 millivolts. Conversely, I_{SC} implies that *V* vanishes, which gives:

$$I_{SC} = \frac{R_{SH}}{R_{SH} - R_S}\left(I_L - I_0\left(e^{qI_{SC}R_S/kT} - 1\right)\right)$$

Again, this formula can be simplified subject to some reasonable approximation, which in this case is that shunt resistance is very much larger than series resistance. If, in addition, the exponential can be approximated by a truncated power series, then it follows that:

$$I_{SC} \cong \frac{I_L}{1 + R_S\left(\frac{1}{R_{SH}} + \frac{qI_0}{kT}\right)}$$

Clearly, if series resistance can be ignored altogether, then I_{SC} trivially reduces to the photocurrent.

Optimal Operation of a Photovoltaic Device

Obviously, V_{OC} and I_{SC} cannot correspond to the optimal operating point of a solar cell. This is determined when output power is maximized. In general, for any electrical generating device, output power corresponds to the product of current supplied to the outside world and the supply voltage, which in the present case is just the product, *IV*. Clearly, the device can supply no power to the external world under either short or open circuit conditions. Concomitantly, for some external load attached to a solar cell supply voltage must fall below V_{OC} and operating current below I_{SC}. Nevertheless, the product, *IV*, does not vanish indicating that light energy absorbed is supplied as power to an external load. The relationship between characteristic IV and PV curves and optimal supply voltage and operating current, $V_{\max}$ and $I_{\max}$, for a solar cell are illustrated in the following figure:

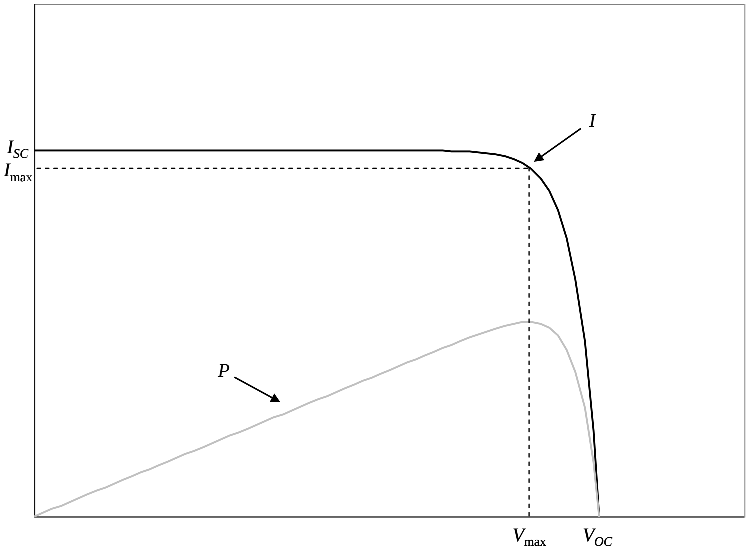


Fig. 61: Solar cell IV and PV curves

Accordingly, the derivative of output power, *P*, with respect to supply voltage, *V*, vanishes at the optimal operating point. Within this context, the derivative of *P* with respect to *V* is constructed implicitly from the IV characteristic as follows:

$$\frac{dP}{dV} = I + V\frac{dI}{dV}$$

Thus, it immediately follows that:

$$\frac{dI}{dV} = -\left(1 + R_S\frac{dI}{dV}\right)\left(\frac{q}{kT}I_0 e^{q(V+IR_S)/kT} + \frac{1}{R_{SH}}\right)$$

This expression can be rearranged to formulate an explicit expression for the current derivative, thus:

$$\frac{dI}{dV} = -\frac{\frac{q}{kT} I_0 e^{q(V+IR_S)/kT} + \frac{1}{R_{SH}}}{1 + R_S\left(\frac{q}{kT} I_0 e^{q(V+IR_S)/kT} + \frac{1}{R_{SH}}\right)}$$

As is evident from the preceding figure the derivative must be uniformly negative. Combining expressions one obtains:

$$\frac{dP}{dV} = I_L - I_0\left(e^{q(V+IR_S)/kT} - 1\right) - \frac{V + IR_S}{R_{SH}} - \frac{\frac{qV}{kT} I_0 e^{q(V+IR_S)/kT} + \frac{1}{R_{SH}}}{1 + R_S\left(\frac{q}{kT} I_0 e^{q(V+IR_S)/kT} + \frac{1}{R_{SH}}\right)}$$

Of course, the left hand side vanishes at the optimal operating point, which, in principle, allows $V_{\max}$ and $I_{\max}$ to be uniquely determined. In practice, the preceding equation can only be solved numerically. Within this context, *fill factor*, *F*, is defined by the ratio:

$$F = \frac{I_{\max} V_{\max}}{I_{SC} V_{OC}}$$

Typically, fill factor is quoted in per cent and is larger the more "square" the IV curve appears. Indeed, although physically impossible, it is evident that if $V_{\max}$ and $I_{\max}$ corresponded exactly with V_{OC} and I_{SC}, then the IV curve would be precisely a rectangle and F would be exactly 100%. Conversely, if IV curves become "squashed" or "rounded" due to unfavorable values of parasitic resistances or poor diode characteristics, then fill factor falls to low values, *e.g.*, below 40%. Accordingly, fill factor provides a useful figure of merit for quality of a solar cell.

Fill factor is directly related to efficiency, η, which can be defined precisely as follows:

$$\eta = \frac{I_{\max} V_{\max}}{P_{\text{in}}} = F\,\frac{I_{SC} V_{OC}}{P_{\text{in}}}$$

As indicated previously efficiency is usually quoted in percent and; hence, is defined as the quotient of maximal power, *viz.*, $I_{\max}V_{\max}$, with incident power due to illumination, P_{in}. Typically, efficiency is defined under conditions of "one sun", which assumes a standard solar spectrum and a power density of nominally one kilowatt per square meter under conditions of perpendicular solar illumination. Therefore, a solar cell (or array) having 20% efficiency and one square meter in area could, in principle supply two hundred watts of power. In practice, it is difficult to achieve high efficiency due to various factors such as sun angle, shading, *etc.* Accordingly, different strategies are adopted to mitigate these factors. For example, a solar panel might be combined with a mechanism that changes orientation and tilt during the day to track the sun. Even so, intermittency remains a serious problem for all forms of renewable energy. For this reason integration of solar energy with the existing electrical grid is difficult since it adds complexity to large scale power management. One solution for this is large-scale battery storage; however, cost and inefficiency still remain serious obstacles.

Light Emitting Diodes

It comes as no surprise that the photovoltaic effect can be inverted. This means that flow of a current through a diode results in emission rather than absorption of electromagnetic radiation. Such a device is called a *light emitting diode* or *LED*. An important distinction to be made between an LED and a photovoltaic is that although a photovoltaic can be made from both direct and indirect band gap semiconductors; because an LED requires a direct photonic transition, a direct band gap semiconductor, such as gallium arsenide, is necessary. Consequently, silicon cannot be used for an LED. Physically, light is emitted from an LED if the junction is fully turned on in forward bias, *i.e*, if the applied bias voltage substantially exceeds the diffusion potential. Typically, the "on" voltage is 2 to 3 volts. The emitted wavelength is determined by the band gap of the semiconductor, which in the case of GaAs is 1.424 eV and, therefore, infrared.

The earliest LEDs appeared as practical electronic components in 1962 and emitted low-intensity infrared light. More specifically, in the fall of 1961, James R. Biard and Gary Pittman, employees of Texas Instruments, Inc., observed that in strong forward bias a gallium arsenide diode emitted infrared light. Subsequently, Biard and Pittman filed a patent entitled "Semiconductor Radiant Diode", which described a zinc diffused GaAs *pn*-junction LED. Texas Instruments then began to manufacture infrared diodes and in October of 1962 announced the first LED commercial product which emitted at 900 nm. (Indeed, such infrared LEDs are still frequently used as transmitting devices in remote controls for consumer electronics.) A visible-spectrum (red) LED was invented in 1962 by Nick Holonyak, Jr., an employee of the General Electric Company. As might be expected, the first visible-light LEDs were also of low intensity, and limited to red (commensurate with band gaps of GaAs and alloys such as AlGaAs). Even so, early visible-light LEDs were frequently used as indicator lamps for electronic devices, replacing small and unreliable incandescent bulbs. Concomitantly, they were packaged in seven-segment displays and, thus, used as digital readouts as commonly seen in digital clocks. Ten years later in 1972, M. George Craford, a former graduate student of Holonyak, invented the first yellow LED and improved the brightness of red and red-orange LEDs by a factor of ten. A few years later in 1976, T. P. Pearsall developed the first high-brightness, high-efficiency LEDs for optical fiber telecommunications by synthesizing semiconductor materials specifically adapted to light wavelengths optimized for optical fiber transmission. More recently, other III-V semiconductors having larger direct band gap, such as gallium nitride, have allowed development of LEDs of almost any color. Consequently, very high brightness LEDs are currently available for visible, ultraviolet, and infrared wavelengths. Accordingly, blue LEDs have been combined with a broad spectrum phosphor to produce very bright white LEDs useful for environmental and task lighting. In such applications, LEDs have substantial advantage over incandescent lamps such as lower energy consumption, longer lifetime, improved mechanical and electrical reliability, reduced size, and faster switching. Consequently, light emitting diodes are currently used in applications as diverse as aviation lighting, automotive headlamps, advertising, general lighting, traffic signals, and camera flashes. Even so, LEDs bright enough for ambient room lighting are still relatively expensive and require more precise current and heat management than incandescent or compact fluorescent lamps of comparable luminosity. Nevertheless, it is generally expected that LED lighting will eventually replace conventional lamps and become ubiquitous.

Atomic Processes of Diffusion

Just as for other processes such as oxidation, diffusion is thermally activated and can be considered within a thermodynamic context. Therefore, the free energy change associated with a diffusion process can be represented by the fundamental expression:

$$\Delta A_D = \Delta E_D - T\Delta S_D$$

For diffusion in a silicon crystal, the internal energy change, ΔE_D, is determined by the net number of bonds broken during diffusion. Obviously, an overall diffusion process may require bond breakage for atoms to migrate, however, these bonds subsequently reform; hence, net bond breakage, discounting the effects of lattice defects, *etc.*, can be expected to be nearly zero. Therefore, the energetic contribution to free energy change is negligible. In contrast, entropy is the thermodynamic measure of disorder; hence, addition of an impurity has a similar effect on the entropy of a perfect crystal as found previously for the addition of point defects. (Indeed, the effect is very similar since an impurity can be treated as just another kind of point defect in addition to vacancies and interstitials.) Therefore, mixing of an impurity into a pure material increases disorder, and thus, must increase entropy. The classical expression for the entropy change associated with mixing two pure materials, *A* and *B*, is:

$$\Delta S_{mix} = k\ln\left(\frac{(N_A + N_B)!}{N_A!N_B!}\right)$$

Here, N_A and N_B are the number of atoms of species *A* and species *B* respectively. (Of course, the argument of the logarithm is the "number of distinguishable arrangements" of N_A atoms of species *A* and N_B atoms of species *B*, *i.e.*, a binomial coefficient.) As usual, since N_A and N_B are very large numbers, one can express the logarithm as a sum of logarithms and further simplify the result by means of Stirling's approximation:

$$\Delta S_{mix} = k\ln(N_A + N_B)! - k\ln N_A! - k\ln N_B! =$$

$$k((N_A + N_B)\ln(N_A + N_B) - N_A\ln N_A - N_B\ln N_B) =$$

$$-kN_A\ln\left(\frac{N_A}{N_A + N_B}\right) - kN_B\ln\left(\frac{N_B}{N_A + N_B}\right)$$

Naturally, it is desirable to recast this expression in terms of concentrations, C_A and C_B, rather than absolute numbers of impurity species. Therefore, if *V* is identified as the volume of material, then it follows that "entropy of mixing" has the form:

$$\Delta S_{mix} = -kV\left[C_A\ln\left(\frac{C_A}{C_A + C_B}\right) + C_B\ln\left(\frac{C_B}{C_A + C_B}\right)\right]$$

In passing, it is worthwhile to note that material (or mole) fractions, X_A and X_B, are also defined respectively as $N_A/(N_A+N_B)$ and $N_B/(N_A+N_B)$ or $C_A/(C_A+C_B)$ and $C_B/(C_A+C_B)$. Hence, entropy of mixing can be expressed in terms of X_A and X_B, thus:

$$\Delta S_{mix} = -kN(X_A \ln X_A + X_B \ln X_B)$$

Here, N is just the total number of atoms, N_A+N_B. (This is the conventional expression for entropy of mixing often appearing in textbooks.)

Naturally, if one identifies species A as silicon and species B as impurity, then it follows that:

$$\Delta S_D = -kV\left[C_{\text{Si}} \ln\left(\frac{C_{\text{Si}}}{C_{\text{Si}}+C_I}\right) + C_I \ln\left(\frac{C_I}{C_{\text{Si}}+C_I}\right)\right]$$

Of course, the concentration of silicon atoms is much larger than the concentration of impurity atoms; hence, the first logarithmic term can be ignored as negligible, *i.e.*, ln 1 vanishes, therefore:

$$\Delta S_D = -kVC_I \ln\left(\frac{C_I}{C_{\text{Si}}+C_I}\right) = -kN_I \ln\left(\frac{C_I}{C_{\text{Si}}+C_I}\right)$$

Here, N_I is the total number of impurity atoms, *i.e.*, $C_I V$. Thus, the free energy change associated with diffusion is:

$$\Delta A_D = kTN_I \ln\left(\frac{C_I}{C_{\text{Si}}+C_I}\right) \cong -kTN_I \ln\left(\frac{C_{\text{Si}}}{C_I}\right)$$

Obviously, the concentration of silicon atoms can be determined by observing that there are eight atoms per unit cell in a diamond cubic structure, hence, C_{Si}, is merely $8/a^3$ and a is just the lattice parameter as usual. One finds that for a dopant concentration of 10^{16} cm^{-3} at 1000°C, the free energy change per unit volume, *i.e.*, $\Delta A_D/V$, is approximately -200 J cm^{-3}. This is a negative free energy change and clearly illustrates that diffusion occurs spontaneously.

Diffusion Mechanisms

To characterize impurity diffusion in crystalline materials atomistically, two distinct diffusion mechanisms must be considered. These are *interstitial* and *substitutional diffusion*. Both of these mechanisms require participation of point crystalline defects, *viz.*, interstitials and vacancies. The interstitial mechanism is perhaps the easiest to visualize. As one might expect, interstitial diffusion occurs when migrating atoms "jump" between interstitial sites within the crystal lattice. If the "jumping" atom is silicon, this is trivially equivalent to the migration of an interstitial defect (specifically, a silicon self-interstitial defect, as considered previously). However, if the migrating atom is an impurity, the result is diffusion of impurity species. One expects this process to be thermally activated; hence, temperature dependence of interstitial diffusion is characterized by an Arrhenius form:

$$D_i = D_{oi} e^{-Q_i/kT}$$

Here, D_i can be the diffusivity of either impurity or silicon self-interstitials, Q_i is activation energy for this same process, and D_{oi} is an associated pre-exponential factor, *i.e.*, infinite temperature diffusivity. Of course, Q_i can be regarded physically as formation energy of some transition state. In the case of interstitial diffusion, this is related to a transient increase in strain energy of the crystal lattice as a diffusing atom migrates from one interstitial site to another. Since, the diamond cubic structure is relatively open; activation energy for interstitial diffusion can be expected to be relatively small. Of course, specific values of D_{oi} and Q_i will depend on the nature of the diffusing species and can be expected to differ for impurity atoms or silicon self-interstitials.

Similarly, substitutional diffusion is also a thermally activated process requiring participation of vacancies instead of interstitials. In this case, the activation energy consists of two components. The first is related to the formation energy of a vacancy since diffusion by this mechanism cannot occur unless there is a vacancy in close proximity to an impurity atom. This contrasts with interstitial diffusion since no additional energy is required for the formation of an interstitial site (these are permanently present in the lattice). The second component is an activation energy associated with the migration of a vacancy from one lattice site to an adjacent one. (This two component formulation is similar to the form appearing previously for vacancy equilibrium.) Thus, simple vacancy diffusivity, D_ν, can also expressed as an Arrhenius form:

$$D_\nu = D_{o\nu} e^{-(Q_{f\nu} + Q_{m\nu})/kT}$$

In analogy to the previous case of interstitial diffusion $D_{o\nu}$ is infinite temperature diffusivity. However, $Q_{f\nu}$ is the contribution to activation energy due to vacancy formation and $Q_{m\nu}$ is the contribution due to vacancy migration. Physically, activation energies, $Q_{m\nu}$ and Q_i are of similar size and relatively small (typically less than 1 eV). In contrast, $Q_{f\nu}$ is much larger (typically 3 to 5 eV's); hence, the overall activation energy for vacancy diffusion is significantly larger than for interstitial diffusion. Accordingly, for any definite temperature, one expects D_ν to be much smaller than D_i. Therefore, in general, one expects interstitial impurities, *e.g.*, heavy metals, *etc.*, to diffuse much more rapidly than substitutional impurities, *e.g.*, shallow level dopants, which diffuse by a vacancy mechanism. This is indeed found to be the case.

A third diffusion mechanism called the *interstitialcy mechanism* occurs when a silicon self-interstitial exchanges with a substituted impurity atom. Consequently, this mechanism requires interaction of interstitial defects with substitutional impurities and can be regarded within the context of a dynamical equilibrium:

$$I_l + \mathrm{Si}_i \overset{\rightarrow}{\leftarrow} I_i + \mathrm{Si}_l$$

Here, the subscript, *l*, denotes an atom substituted in the lattice and the subscript *i* denotes an interstitial atom. One can formally define an equilibrium constant, *K*, as follows:

$$K = \frac{[I_i][\mathrm{Si}_l]}{[I_l][\mathrm{Si}_i]}$$

As a matter of convention, the square brackets denote atomic concentrations of each species in the crystal lattice. Since, the concentration of silicon atoms in lattice sites is large and practically constant, an effective equilibrium constant, *K′*, can be defined as the quotient, $K/[\mathrm{Si}_l]$. Therefore, one expects that the contribution of the interstitialcy mechanism to the total diffusivity, *D*, for a substitutional impurity is proportional to the ratio of the concentration of impurity atoms in interstitial sites to the concentration in lattice sites. Thus, in terms of the effective equilibrium constant, *D* is given by:

$$D = D_o + K'[\mathrm{Si}_i]D_i$$

Of course, D_o is diffusivity in the absence of silicon self-interstitial defects. Thus, one expects that in the presence of silicon self-interstitial defects, the diffusivity of substitutional impurities should be enhanced.

Defect Charge and Dopant Diffusivity

Naturally, intrinsic point defects are associated with localized electronic states, which may have energies lying somewhere (*i.e.*, either shallow or deep) within the band gap. Furthermore, these defect states can interact with the normal band states of the crystal and, in the process change charge state. Therefore, point defects may become charged in analogy to shallow level impurity atoms and/or interface traps. Thus, if one considers total diffusivity as a sum of contributions from point defects of various types and charge states, one expects that activation energies and pre-exponential factors will not necessarily be the same for each contribution. Therefore, impurity diffusion requires detailed consideration of these different contributions. Moreover, since, shallow level dopants are substitutional impurities and diffuse primarily by a vacancy mechanism, the effect of neutral and charged vacancies must be included in any formulation for dopant diffusivity.

To begin an analysis of dopant diffusion, one first observes that silicon vacancies exist in equilibrium with mobile carrier concentrations. These defect-carrier equilibria can be formally represented as follows:

$$V^x \underset{\leftarrow}{\rightarrow} V^- + h^+ \qquad K_V^- = p\frac{[V^-]}{[V^x]}$$

$$V^- \underset{\leftarrow}{\rightarrow} V^= + h^+ \qquad K_V^= = p\frac{[V^=]}{[V^-]}$$

$$V^x \underset{\leftarrow}{\rightarrow} V^+ + e^- \qquad K_V^+ = n\frac{[V^+]}{[V^x]}$$

$$V^+ \underset{\leftarrow}{\rightarrow} V^{++} + e^- \qquad K_V^{++} = n\frac{[V^{++}]}{[V^+]}$$

Here, V^x, V^-, $V^=$, V^+, and V^{++}, denote respectively, neutral, singly and doubly negative, and singly and doubly positive charged vacancies. Of course, h^+ and e^- denote mobile holes and electrons. Naturally, thermodynamic equilibrium constants, K_V^-, $K_V^=$, K_V^+, and K_V^{++} are defined as usual. Individual equilibria can also be combined to give overall expressions. Clearly, the total vacancy concentration, $[V]$, is just the sum of all charge state contributions:

$$[V]=[V^x]+[V^-]+[V^=]+[V^+]+[V^{++}]$$

As has been observed previously, vacancies can diffuse in a semiconductor crystal just as dopants can. Therefore, the total diffusion coefficient for vacancies can be expressed as a sum of contributions from each distinct charge state:

$$D_V = \frac{[V^x]}{[V]}D_V^x + \frac{[V^-]}{[V]}D_V^- + \frac{[V^=]}{[V]}D_V^= + \frac{[V^+]}{[V]}D_V^+ + \frac{[V^{++}]}{[V]}D_V^{++}$$

Naturally, each contribution is weighted by the fraction of each vacancy charge state relative to the total vacancy concentration.

At this point, rather than considering each type of impurity species separately, substitutional silicon self-diffusion can be considered as a typical vacancy mediated diffusion process. Thus, one formally relates the silicon self-diffusion coefficient, D_{Si}, to simple vacancy diffusivity as follows:

$$D_{\mathrm{Si}} = \frac{f}{C_{\mathrm{Si}}}[V]D_V$$

Here, f is identified as an as yet undetermined "correlation factor". Upon substitution, one finds:

$$D_{\mathrm{Si}} = \frac{f}{C_{\mathrm{Si}}}([V^x]D_V^x + [V^-]D_V^- + [V^=]D_V^= + [V^+]D_V^+ + [V^{++}]D_V^{++})$$

Therefore, the contribution to the silicon self-diffusion coefficient due to a particular vacancy charge state, *r*, is defined formally, thus:

$$D_{\mathrm{Si}}(r) = \frac{f}{C_{\mathrm{Si}}}[V^r]D_V^r$$

In principle, if diffusivity contributions from each particular type of charged vacancy are known, one can construct the silicon self-diffusion coefficient.

Clearly, since mobile carrier concentrations participate directly in the defect equilibria, they must affect the concentrations of charged vacancies. Thus, $D_{\mathrm{Si}}(r)$ is evidently a function of extrinsic doping. Therefore, it is useful to consider silicon self-diffusion in intrinsic silicon as a reference or standard state. Indeed, values have been experimentally determined for $D_{\mathrm{Si}}(r)$ in intrinsic silicon and are conventionally denoted as D_{Si}^r:

$$D_{\mathrm{Si}}^r = \frac{f_i}{C_{\mathrm{Si}}}[V^r]_i D_V^r$$

Here, f_i is, again, a correlation factor, but specifically for intrinsic silicon, and $[V^r]_i$ is the concentration of V^r in intrinsic silicon. One immediately can express D_{Si} in terms of the D_{Si}^r's:

$$D_{\mathrm{Si}} = \frac{f}{f_i}\left(\frac{[V^x]}{[V^x]_i}D_{\mathrm{Si}}^x + \frac{[V^-]}{[V^-]_i}D_{\mathrm{Si}}^- + \frac{[V^=]}{[V^=]_i}D_{\mathrm{Si}}^= + \frac{[V^+]}{[V^+]_i}D_{\mathrm{Si}}^+ + \frac{[V^{++}]}{[V^{++}]_i}D_{\mathrm{Si}}^{++}\right)$$

Clearly, since neutral vacancies do not strongly interact electrically with mobile carrier concentrations, it is plausible to assume that that the diffusivity of neutral vacancies is relatively unaffected by carrier concentrations, hence, correlation factors and neutral vacancy concentrations satisfy the relation:

$$\frac{f}{f_i}\left(\frac{[V^x]}{[V^x]_i}\right) = 1$$

This assumption yields an expression for D_{Si} in terms of intrinsic and extrinsic defect concentrations:

$$D_{\mathrm{Si}} = D_{\mathrm{Si}}^x + \frac{[V^x]_i[V^-]}{[V^-]_i[V^x]}D_{\mathrm{Si}}^- + \frac{[V^x]_i[V^=]}{[V^=]_i[V^x]}D_{\mathrm{Si}}^= + \frac{[V^x]_i[V^+]}{[V^+]_i[V^x]}D_{\mathrm{Si}}^+ + \frac{[V^x]_i[V^{++}]}{[V^{++}]_i[V^x]}D_{\mathrm{Si}}^{++}$$

Of course, the defect concentrations are related to carrier concentrations by defect-carrier equilibria. Thus, equilibrium defect concentrations can be replaced by equilibrium carrier concentrations:

$$D_{\mathrm{Si}} = D_{\mathrm{Si}}^{x} + \frac{n_i}{p} D_{\mathrm{Si}}^{-} + \frac{n_i^2}{p^2} D_{\mathrm{Si}}^{=} + \frac{n_i}{n} D_{\mathrm{Si}}^{+} + \frac{n_i^2}{n^2} D_{\mathrm{Si}}^{++}$$

Typically, it is found that, in addition to neutral vacancies, charged vacancies having the same polarity as majority carriers, *i.e.*, opposite of ionized impurity atoms, make significant contributions to diffusivity, *i.e.*, V^{-} and $V^{=}$ in *n*-type silicon and V^{+} and V^{++} in *p*-type. Thus, it is useful to apply carrier equilibrium and rearrange the preceding expression as follows:

$$D_{\mathrm{Si}} = D_{\mathrm{Si}}^{x} + \frac{n}{n_i} D_{\mathrm{Si}}^{-} + \frac{n^2}{n_i^2} D_{\mathrm{Si}}^{=} + \frac{p}{n_i} D_{\mathrm{Si}}^{+} + \frac{p^2}{n_i^2} D_{\mathrm{Si}}^{++}$$

In practice, each of the D_{Si}^{r}'s has been found to have an Arrhenius form:

$$D_{\mathrm{Si}}^{r} = D_{o\mathrm{Si}}^{r} e^{-Q_{\mathrm{Si}}^{r}/kT}$$

Since values of $D_{o\mathrm{Si}}^{r}$ and Q_{Si}^{r} have been experimentally observed, D_{Si} is easily calculated for any given temperature.

Of course, diffusion of any substitutional impurity is similar to silicon self-diffusion. (Obviously, characteristic activation energies and pre-exponential factors will differ, but the general form of the diffusivity can be expected to be similar.) Therefore, upon application of this same analysis, an expression for impurity diffusivity is easily obtained which is entirely analogous to the preceding expression for silicon self-diffusivity. Thus, one can write:

$$D_I = D_I^{x} + \frac{n}{n_i} D_I^{-} + \frac{n^2}{n_i^2} D_I^{=} + \frac{p}{n_i} D_I^{+} + \frac{p^2}{n_i^2} D_I^{++}$$

Obviously, this expression has been obtained just by replacing the subscript "Si" with "*I*" to denote an arbitrary substitutional impurity species. Naturally, the D_I^{r}'s can also be expected to be of Arrhenius form:

$$D_I^{r} = D_{oI}^{r} e^{-Q_I^{r}/kT}$$

Furthermore, just as for silicon self-diffusion, values of D_{oI}^{r} and Q_I^{r} have been measured for various dopant species. Therefore, contributions to impurity diffusivity, D_I^{r}, can easily be obtained as a function of temperature. Characteristic values for $D_{o\mathrm{Si}}^{r}$ and Q_{Si}^{r}

and D_{oI}^{r} and Q_{I}^{r} for the most important diffusion mechanisms are collected in the following table:

Atomic Species I	Diffusion Mechanism V^{r}	D_{oI}^{r} (cm²/sec)	Q_{I}^{r} (eV)
Si	V^{x}	0.015	3.89
	V^{-}	16	4.54
	$V^{=}$	10	5.1
	V^{+}	1180	5.09
As	V^{x}	0.066	3.44
	V^{-}	12.0	4.05
B	V^{x}	0.037	3.46
	V^{+}	0.76	3.46
Ga	V^{x}	0.374	3.39
	V^{+}	28.5	3.92
P	V^{x}	3.85	3.66
	V^{-}	4.44	4.00
	$V^{=}$	44.2	4.37
Sb	V^{x}	0.214	3.65
	V^{-}	15.0	4.08
N	V^{x}	0.05	3.65

Table 5: Arrhenius forms for defect mediated contributions to substitutional diffusivities

Of course, it is important to note that values of the intrinsic carrier concentration must be determined by the carrier equilibrium at the process temperature. Typically, the process temperature is relatively high, (>900°C); hence, it is usually the case that n_i will dominate over any extrinsic doping. Therefore, carrier concentration ratios in the preceding expression will often approach unity and, thus D_I reduces to the simple expression:

$$D_I \cong D_I^{x} + D_I^{-} + D_I^{=} + D_I^{+} + D_I^{++}$$

Furthermore, in most cases D_I is dominated by only one or two terms.

Electric Field Effect

Because dopant atoms are ionized within the crystal lattice, any internal electric field can affect diffusivity. Of course, in the region of the junction just such an electric field exists due to the depletion region. This internal electric field is determined by the gradient of the Fermi potential. One recalls that the Fermi potential is given by the expression:

$$\varphi_F = \frac{kT}{2q}\ln\left(\frac{|N_A - N_D| + \sqrt{(N_A - N_D)^2 + 4n_i^2}}{2n_i}\right)^2 = \frac{kT}{q}\ln\frac{|N_A - N_D| + \sqrt{(N_A - N_D)^2 + 4n_i^2}}{2n_i}$$

Thus, to within a sign the internal electric field, *E*, is just the spatial derivative of the Fermi potential:

$$E = \mp\frac{\partial}{\partial x}\varphi_F$$

Of course, the upper sign denotes an *n*-type diffusion and the lower sign a *p*-type diffusion.

Naturally, diffusion of ionized dopant impurities is a transport process entirely analogous to transport of mobile carriers. Thus, in the presence of an electric field, the total diffusion flux will have both a diffusive contribution (from Fick's Law) and a drift contribution (from Ohm's Law):

$$J = -D_I\frac{\partial C_I}{\partial x} \pm \frac{qD_I}{kT}C_I E = -D_I\frac{\partial C_I}{\partial x} - \frac{qD_I}{kT}C_I\frac{\partial\varphi_F}{\partial x}$$

Clearly, sign options cancel out and the second term is explicitly negative for both *n*-type and *p*-type semiconductor. Moreover, if only one type of impurity is dominant (as is usual), the Fermi potential has the form:

$$\varphi_F = \frac{kT}{q}\ln\frac{C_I + \sqrt{C_I{}^2 + 4n_i^2}}{2n_i}$$

Here, C_I can be either acceptor or donor concentration. Thus, one substitutes as follows:

$$J = -D_I\frac{\partial C_I}{\partial x} - \frac{qD_I}{kT}C_I\frac{\partial}{\partial x}\left(\frac{kT}{q}\ln\frac{C_I + \sqrt{C_I^2 + 4n_i^2}}{2n_i}\right)$$

Considering the dopant concentration, C_I, as an explicit function of *x*, one can construct the derivative to obtain:

$$J = -D_I \frac{\partial C_I}{\partial x} - D_I C_I \frac{\partial}{\partial x}\left(\ln \frac{C_I + \sqrt{C_I^2 + 4n_i^2}}{2n_i} \right) = -D_I \frac{\partial C_I}{\partial x} - D_I \frac{\partial C_I}{\partial x}\left(\frac{C_I}{\sqrt{C_I^2 + 4n_i^2}} \right)$$

Thus, one can define the *electric field coefficient*, *h*:

$$h = 1 + \frac{C_I}{\sqrt{C_I^2 + 4n_i^2}} = 1 + \frac{C_I}{2n_i} \frac{1}{\sqrt{1 + \left({C_I}/{2n_i} \right)^2}}$$

Therefore, the linear transport relation takes the form:

$$J = -D_I h \frac{\partial C_I}{\partial x}$$

Here, $D_I h$ can be considered as an effective diffusivity, which takes into account the electric field effect. Clearly, any internal electric field serves to enhance diffusion. If C_I greatly exceeds the intrinsic concentration, *i.e.*, $C_I >> n_i$, then h evidently approaches a value of 2. In contrast, if intrinsic carriers dominate, *i.e.*, $C_I << n_i$, then h is essentially unity. Furthermore, the electric field effect introduces concentration dependence into the effective diffusivity which causes the transport equations to become non-linear.

Non-Linear Diffusion

At very high impurity concentrations (*viz.*, near the solubility limit) dopant diffusivity generally becomes concentration dependent and corresponding diffusion processes necessarily become non-linear. Moreover, in addition to the effect of an internal electric field, at high concentrations dopant clusters or complexes can also form, which normally tend to reduce effective diffusivity. Consequently, enhancement of diffusivity due to interaction of ionized impurity atoms and mobile carriers can be offset by complex or cluster formation at very high concentrations. In any case, irrespective of cause or precise behavior, *i.e.*, enhancement or suppression of diffusive transport, treatment of concentration dependent diffusion requires appropriate modification of the fundamental transport relation (Fick's First Law), which then takes the more general form:

$$F = -D(C) \frac{\partial C}{\partial x}$$

Clearly, this expression no longer is applicable to a strict linear phenomenology; however, one can still construct a non-linear form of the diffusion equation by the usual method, thus:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D(C) \frac{\partial C}{\partial x} \right)$$

Unfortunately, this equation cannot be solved using the Principle of Superposition. (Indeed, non-linear differential equations are at best difficult and usually impossible to solve.) Even so, concentration profiles associated with non-linear diffusion processes are readily constructed by numerical methods.

Of course, for any accurate numerical description of non-linear diffusion knowledge of the functional dependence of diffusivity on concentration is required, which in practice must generally be determined by experiment. Concomitantly, a convenient technique for determination of $D(C)$ requires formal integration of the non-linear diffusion equation subject to the specific restriction that concentration can be expressed in terms of a single variable, η, defined formally as $x/2\sqrt{t}$. Subject to this restriction, the non-linear diffusion equation can be transformed as follows:

$$\frac{\partial C}{\partial t}=\frac{\partial \eta}{\partial t}\frac{\partial C}{\partial \eta}=\frac{-x}{4t^{3/2}}\frac{\partial C}{\partial \eta} \quad ; \quad \frac{\partial C}{\partial x}=\frac{\partial \eta}{\partial x}\frac{\partial C}{\partial \eta}=\frac{1}{2t^{1/2}}\frac{\partial C}{\partial \eta}$$

These expressions are obtained by elementary application of the chain rule; hence, one finds that:

$$\frac{-x}{4t^{3/2}}\frac{\partial C}{\partial \eta}=\frac{\partial}{\partial x}\left(D(C)\frac{1}{2t^{1/2}}\frac{\partial C}{\partial \eta}\right)=\frac{1}{2t^{1/2}}\frac{\partial}{\partial x}\left(D(C)\frac{\partial C}{\partial \eta}\right)$$

Naturally, the remaining coordinate derivative can be rewritten as a derivative with respect to η, thus:

$$\frac{-x}{4t^{3/2}}\frac{\partial C}{\partial \eta}=\frac{1}{4t}\frac{\partial}{\partial \eta}\left(D(C)\frac{\partial C}{\partial \eta}\right)$$

Within this context, the original non-linear partial differential equation has been reduced to a non-linear ordinary differential equation:

$$\frac{dC}{d\eta}=\frac{-1}{2\eta}\frac{d}{d\eta}\left(D(C)\frac{dC}{d\eta}\right)$$

Clearly, a complementary error function concentration profile can be considered as a function of $x/2\sqrt{t}$ only and, as such, satisfies the required restriction. (Of course, this corresponds to the special case that $D(C)$ is strictly constant.) In contrast, the Gaussian concentration profile cannot be considered as a function only of $x/2\sqrt{t}$. Consequently, constant source boundary conditions are naturally adapted to non-linear diffusion by defining the concentration, C, as a constant, C_o, for $x<0$ and $t=0$ and as vanishing for $x>0$ and $t=0$. In terms of η, this implies that C equals C_o if η tends toward $-\infty$ and that C equals 0 if η tends toward ∞. Accordingly, one can formally integrate the preceding expression to obtain:

$$-2\int_0^C \eta(C')dC' = D(C)\frac{dC}{d\eta} - D(C)\left.\frac{dC}{d\eta}\right|_{\eta\to\infty}$$

Here, η(*C*) is to be regarded as the formal inverse function of the concentration profile, *C*(η). In this form, this equation can be used to characterize any dependence of diffusivity on concentration. Clearly, $\partial C/\partial\eta$ must vanish as η tends toward ∞. Thus, it follows that:

$$D(C) = \frac{-2\int_0^C \eta(C')dC'}{dC/d\eta} = -2\left(\frac{d\eta}{dC}\right)\int_0^C \eta(C')dC' = -\frac{1}{2t}\left(\left.\frac{\partial C}{\partial x}\right|_t\right)^{-1}\int_0^C x(C',t)dC'$$

This is the *Boltzmann-Matano formula*. Within this context, *x*(*C*,*t*) should be interpreted as the formal inverse of *C*(*x*,*t*) with *t* regarded as a fixed parameter.

The classical procedure for experimental determination of diffusivity is construction of a couple (*i.e.*, heterojunction) consisting of two different materials. Furthermore, this method can be readily adapted to silicon by deposition of a uniform, heavily-doped layer (glass, polysilicon, *etc.*) on the silicon surface as a diffusion source or by carrying out a constant source diffusion in which the surface is always maintained at the solubility limit of the diffusing species. In any case, the sample is heat treated at a fixed temperature for a prescribed time interval and the resulting concentration profile is measured (by, perhaps, SIMS or some other applicable method) after which, the experimental concentration profile is numerically converted to an "inverse profile" for which diffused distance, *x*, measured relative to the original interface is determined as a function of concentration. Graphical integration and differentiation of this data allows determination of the concentration dependence of diffusivity using the Boltzmann-Matano formula. Alternatively, to analyze experimental results one might fit a measured concentration profile to a model profile constructed as an explicit functional form incorporating adjustable parameters. Accordingly, an exponential-power function provides a convenient model profile, thus:

$$C(x,t) = C_o e^{\beta^\alpha} \exp\left(-\left(\frac{x}{\sqrt{2\alpha\overline{D}t}} + \beta\right)^\alpha\right)$$

Obviously, α, β, C_o, and $\overline{D}$ are adjustable parameters, which can be interpreted as dimensionless exponent and offset coefficients, surface concentration, and effective diffusivity. Moreover, this formulation is consistent with required boundary conditions from which it follows that:

$$C(\eta) = C_o e^{\beta^\alpha} \exp\left(-\left(\eta\sqrt{\frac{2}{\alpha\overline{D}}} + \beta\right)^\alpha\right)$$

In passing, it is instructive to observe that the parameter groupings, $C_o e^{\beta^\alpha}$ and $\alpha\overline{D}$, are logically equivalent to C_o and $\overline{D}$ and, as such, can be considered as more "natural" adjustable parameters. (Even so, physical interpretation of C_o and $\overline{D}$ is more straightforward and less complicated.) In any case, formal inversion of $C(\eta)$ is a simple matter of elementary algebra, hence:

$$\eta(C) = \sqrt{\frac{\alpha\overline{D}}{2}}\left(\left(\ln\frac{C_o e^{\beta^\alpha}}{C}\right)^{\frac{1}{\alpha}} - \beta\right)$$

Likewise, the derivative is easily constructed from elementary differential identities as follows:

$$\frac{d\eta}{dC} = -\frac{1}{C}\sqrt{\frac{\overline{D}}{2\alpha}}\left(\ln\frac{C_o e^{\beta^\alpha}}{C}\right)^{\frac{1-\alpha}{\alpha}}$$

These expressions are combined in the Boltzmann-Matano formula to give an explicit expression for $D(C)$:

$$D(C) = \frac{\overline{D}}{C}\left(\ln\frac{C_o e^{\beta^\alpha}}{C}\right)^{\frac{1-\alpha}{\alpha}}\int_0^C\left(\left(\ln\frac{C_o e^{\beta^\alpha}}{C'}\right)^{\frac{1}{\alpha}} - \beta\right)dC' = \overline{D}\left(\ln\frac{C_o e^{\beta^\alpha}}{C}\right)^{\frac{1-\alpha}{\alpha}}\left(\frac{1}{C}\int_0^C\left(\ln\frac{C_o e^{\beta^\alpha}}{C'}\right)^{\frac{1}{\alpha}}dC' - \beta\right)$$

Obviously, the second term within the integrand can be trivially identified as $-\beta C$. Unfortunately, the remaining integral term cannot be constructed in closed form; even so, it is desirable to transform the integration variable, thus:

$$u = \ln\left(\frac{C_o e^{\beta^\alpha}}{C'}\right) \qquad du = -\frac{dC'}{C'} = -\frac{dC'}{C_o e^{-u+\beta^\alpha}}$$

Upon substitution, one immediately obtains the expression:

$$D(C) = \overline{D}\left(\ln\frac{C_o e^{\beta^\alpha}}{C}\right)^{\frac{1-\alpha}{\alpha}}\left(\frac{C_o e^{\beta^\alpha}}{C}\int_{\ln\frac{C_o e^{\beta^\alpha}}{C}}^{\infty} du\, e^{-u} u^{\frac{1}{\alpha}} - \beta\right)$$

As a matter of pure mathematics, the integral term can be formally identified with an incomplete gamma function, $\Gamma(a,x)$, which for real-valued parameters, a and x, has the standard definition:

$$\Gamma(a,x) = \int_x^\infty du e^{-u} u^{a-1}$$

Clearly, in the preceding expression for $D(C)$, ${}^{\alpha+1}\!/_{\alpha}$ corresponds to a and $\ln {}^{C_o e^{\beta^\alpha}}\!/_{C}$ to x. (Obviously, the complementary error function itself corresponds to a special case of the incomplete gamma function for which a has a value of exactly ½.) Within this context, it is evident that if $C(x,t)$ corresponds precisely with a complementary error function, then diffusivity must be rigorously independent of concentration. Indeed, for high impurity concentrations near the couple interface (*i.e.*, substrate surface), a complementary error function profile can be fit reasonably well to an exponential-power model profile using typical values for α of between 0.4 and 0.5 and for β of between 1.8 and 2. In contrast, agreement is generally not so good for the profile "tail"; however, in this region of low impurity concentration far from the interface linear diffusion should predominate for any concentration profile and, thus, determination of concentration dependence of diffusivity is of little importance. (Consequently, an exponential-power model profile can be expected to provide a realistic approximation to actual diffused profiles in the "near surface" region.)

Fast Diffusers and Other Contaminants

Of course, shallow level dopant species are generally substitutional impurities in semiconductors since they appear in adjacent columns in the periodic chart and, thus have similar atomic sizes and valencies as the elemental semiconductor atoms, *viz.*, silicon, germanium, *etc*. In contrast, many other species (such as metal atoms) do not "fit" well into the semiconductor crystal structure and, hence, often occupy interstitial sites. Naturally, in analogy to vacancies and substitutional impurities, interstitial species (including silicon self-interstitials) can also interact with mobile carriers and, thus become charged. Therefore, the diffusion mechanism for an interstitial species can also be analyzed in terms of contributions from neutral and charged defects. However, as observed previously, activation energies associated with vacancy diffusion mechanisms are generally larger than activation energies associated with interstitial diffusion mechanisms. Therefore, overall diffusivities of interstitial species can be expected to be much larger than for substitutional impurities. Accordingly, many metallic species are very fast diffusers. This would be of no consequence if metal atoms did not also interact with mobile carriers. Unfortunately, such interactions generally degrade the performance of solid-state devices.

It is worthwhile to digress briefly to consider just how this degradation comes about. Of course, as is the case with any other imperfection in the crystal structure, (*e.g.*, defects, interfaces, *etc.*), the presence of interstitial metal atoms causes electronic states to appear within the band gap. Typically, these states are localized at the site of the metal atom, and as usual, can be rationalized physically as dangling bonds and/or extra atomic orbitals derived from the metal atom electronic structure. The detailed nature of these states is immaterial within the present context. What is important is that these states interact with mobile carriers and substantially increase the overall rate of carrier generation-recombination. Of course, thermally activated carrier generation and recombination is occurring continuously within the crystal. However, in a perfect intrinsic semiconductor crystal, only the process of thermal promotion of an electron from the valence band to the conduction band generates all mobile carriers. Furthermore, the rate of electron-hole recombination in a perfect crystal can be expected to be slow since band states are strongly delocalized, *i.e.*, in order to recombine an electron and a hole must come into close proximity and "collide". The situation is not substantially changed by the addition of dopant impurities to the crystal since the localized states associated with these species are shallow, *i.e.*, near a band edge. (Of course, the major effect of shallow level states is just to change mobile carrier concentrations, *i.e.*, render intrinsic semiconductor either *p*-type or *n*-type, but not substantially to increase the rate of electron-hole recombination.) In contrast, metal atoms are generally *deep level impurities*, which by definition are associated with electronic states lying near the center of the band gap. Such states do not greatly affect overall carrier concentrations, but they do act as *recombination centers*. This may be understood by considering the following figure.

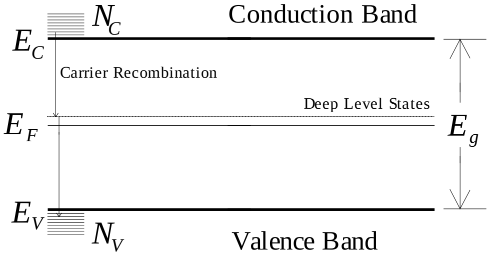


Fig. 62: Mobile carrier recombination due to deep level states

Here, intrinsic silicon has been contaminated by some species that causes localized electronic states to appear just above the center of the band gap. Of course, since these states lie above the Fermi level, the probability that a given state will be empty is greater than one half. Suppose that a mobile electron "wandering" through the conduction band encounters an empty deep level state. Since this state is lower in energy than the conduction band state, it is likely that this electron will "fall" into the deep level state and become captured. Of course, if this happens, the deep level state acquires a relative negative charge due to the occupying electron. Likewise, if a mobile hole encounters the same deep level state, it may easily recombine with the captured electron. (This process may be equally well viewed either, as the captured electron falling into the empty valence band state represented by the hole or the hole falling into the deep level state to annihilate the electron.) One might ask why this process should enhance the overall rate of carrier recombination? After all, the aggregate result is just the same; a conduction band electron has recombined with a valence band hole. In simple terms, a justification can be made, again, by considering mobile carriers as a kind of gas. In an ordinary gas, the collision rate of molecules is directly related to size or more precisely "collision cross section". This is true for mobile carriers inside a semiconductor crystal as well. Naturally, collision cross sections of mobile carriers can be expected to be much smaller than the cross sectional area of an atomic species. However, an interaction of a mobile carrier (either an electron or a hole) and deep level recombination center can be considered as a sort of collision between the mobile carrier and an atom. Clearly, the collision cross section (or more precisely, the "capture cross section") for this process can be expected to be much larger than the cross section for a direct electron-hole recombination and, therefore the rate of the overall recombination becomes much larger. (In equivalent terms, one observes that minority carrier lifetime is reduced.) Of course, the rate of the inverse generation process is similarly enhanced. Accordingly, if one returns to consideration of a *pn*-junction, and recalls that the saturated reverse current is directly related to the rate of carrier generation within the depletion region of the junction, it is clear that deep level states must cause a significant increase junction leakage current.

Some of the most destructive contaminants are commonly occurring metals such as iron, nickel, chromium, copper, *etc.* As expected and as shown by the following table these can have very high diffusivities:

Atomic Species	Mechanism, Temperature, *etc.*	D_{oI} (cm^2/sec)	Q_I (eV)
Ge	substitutional	6.25(10^5)	5.28
Cu	(300°-700°C) (800°-1100°C)	4.7(10^{-3}) 0.04	0.43 1.0
Ag		2(10^{-3})	1.6

Au	substitutional
	interstitial
	(800°-1200°C)
		2.8(10^{-3})
		2.4(10^{-4})
		1.1(10^{-3})
			2.04
			0.39
			1.12

Pt		150-170	2.22-2.15
Fe		6.2(10^{-3})	0.87
Co		9.2(10^4)	2.8
C		1.9	3.1
S		0.92	2.2
O_2		0.19	2.54
H_2		9.4(10^{-3})	0.48
He		0.11	1.26

Table 6: Arrhenius forms for various contaminant species

In general, scrupulous care must be taken to exclude dangerous metallic species from device fabrication processes. Of course, gettering methods and rigorous pre-diffusion surface cleaning can be used to reduce effective metallic contamination. However, the best method is prevention, that is to say avoidance of metal contamination altogether by disciplined handling and process control.

Non-Implanted Dopant Sources and Practical Diffusion Processes

Solid, liquid, and gaseous materials may all be used as non-implanted sources for shallow level dopants. Typically, all of these sources will produce surface doping concentrations near the solubility limit. Physically, solid solubility corresponds to a thermodynamic equilibrium constant and; hence, can be characterized as a function of temperature for various impurity species is shown below:

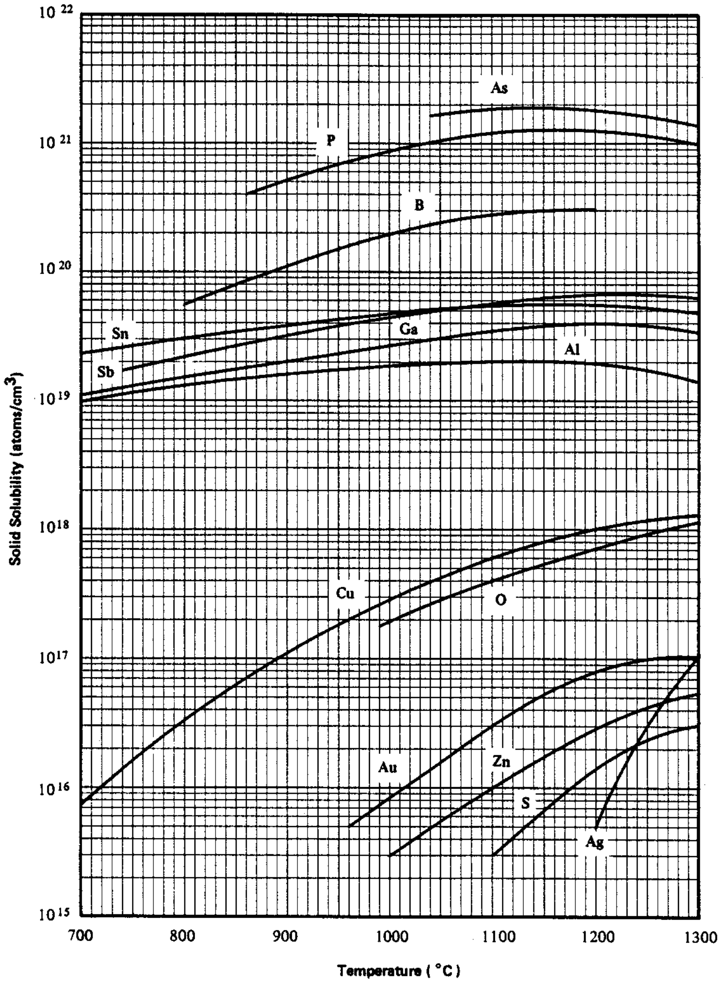


Fig. 63: Solid solubilities of various impurity species in silicon

In practice, solid sources fall into two categories: deposited thin films and preformed substrates. Preformed substrates, such as boron nitride or phosphorous and arsenic impregnated ceramic disks, are used by alternating them with silicon substrates in a quartz tube furnace. The dopants diffuse through the ambient at the high temperature of the diffusion process and deposit on the substrate surface. Typically, solid sources must be activated by a high temperature anneal in oxygen to form a volatile surface oxide and are then used for dopant "pre-deposition". Following pre-deposition, the preformed substrates are removed from the furnace and the substrates (with dopant deposited on the surface) are subjected to a second higher temperature "drive". Obviously, this creates a Gaussian diffusion profile. One problem encountered in this kind of doping process is the formation of a heavily doped surface oxide on the wafer surface. This oxide layer must be removed before further processing to prevent undesirable, accidental doping of other unrelated wafers. In the case of phosphorus and arsenic, this is not difficult, however, borosilicate glass formed by boron doping is difficult to remove and typically must be "cracked" using a steam anneal.

Out diffusion of impurity species from deposited thin films such as in-situ doped chemical vapor deposited polycrystalline silicon or doped "spin-on-glass" (SOG) can also be used to dope semiconductor substrates. Clearly, these diffusion processes closely approximate constant source diffusions with surface concentrations very near the solubility limit. Once the diffusion process is completed the doped SOG must be removed by etching; however, polysilicon is often not removed and is retained as a component of finished devices.

The methodology for use of gases or volatile liquids as dopant sources is very similar. Typical liquid or gaseous source materials are listed in the following table:

Dopant Species *I*	Liquid Source	Gaseous Source
Boron, B	Boron Trichloride, BCl_3 Boron Tribromide, BBr_3	Diborane, B_2H_6
Phosphorus, P	Phosphorus Oxychloride, $POCl_3$ Phosphorus Tribromide, PBr_3	Phosphine, PH_3
Arsenic, As		Arsine, AsH_3

Table 7: Dopant containing volatile liquids and gases used as diffusion sources

In practice, dopant containing gases or volatile liquids are introduced into a gas stream containing a small, controlled amount oxygen mixed with nitrogen or an inert gas. This gas stream is then introduced into a quartz tube furnace containing semiconductor substrates at high temperature. The oxygen is necessary to produce volatile dopant oxides that then deposit on substrate surfaces. (This is similar to the previous case of solid sources.) Obviously, dopant containing gases are easily added directly to the gas stream entering the furnace using high precision valves and flow meters. Typically, liquids are introduced into a diffusion furnace using a "bubbler". A bubbler is a glass or metal vessel held at a fixed temperature through which a stream of "carrier gas" is passed. Since, the vapor pressure of a volatile liquid is a function only of temperature then, for some fixed flow rate of carrier gas, the concentration of dopant containing vapor in the gas stream remains constant. In all cases, control of the effective flow rate of dopant into the furnace is very important in order to achieve a consistent doping profile.

At this point, a number of practical observations can be made concerning various dopant diffusion processes. As asserted previously, at high concentrations (near the solubility limit) dopant diffusion may become non-linear, which is likely due to the formation of ionized impurity-vacancy complexes. Indeed, these complexes can be regarded as "chemical" species formed within the crystalline medium by the "reaction" of ionized dopant atoms and charged vacancies. In particular, ionized arsenic forms complexes with both singly and doubly charged negative vacancies. Corresponding equilibria can be written as follows:

$$2\mathrm{As}^{+} + V^{=} \rightleftarrows \mathrm{As}_2V$$

$$\mathrm{As}^{+} + V^{-} \rightleftarrows \mathrm{As}V$$

Here, AsV and AsV_2 denote uncharged arsenic-vacancy complexes. At lower arsenic concentration, arsenic diffusion is linear and is dominated by neutral and singly charged negative vacancies. The behavior of ionized phosphorus is similar, however only one kind of negatively charged phosphorus-vacancy complex is formed:

$$\mathrm{P}^{+} + V^{=} \rightleftarrows \mathrm{P}V^{-}$$

It is further found that at high concentration, although neutral and singly charged negative vacancies both contribute phosphorus diffusion is dominated by doubly charged negative vacancies. However, if phosphorus concentration is reduced, the diffusion mechanism changes to one dominated by singly charged negative vacancies. As one might expect, boron diffusion is mediated by neutral and positively charged vacancies. In general, complex formation is not important for boron and the dominant diffusion mechanism is dominated by singly charged positive vacancies.

Diffusion in Polycrystalline Silicon

Since doped polycrystalline silicon (*i.e.*, polysilicon) is commonly used as a dopant source for single crystal silicon, it is instructive to digress briefly to consider dopant diffusion in polysilicon itself. In general, due to the existence of grain boundaries in polycrystalline materials, *viz.*, polysilicon, diffusion fluxes must be divided into contributions from bulk and grain boundary diffusion. Indeed, shallow level dopant impurities in polysilicon, in particular phosphorus and arsenic, may preferentially segregate into grain boundaries. (Boron does not significantly segregate into grain boundaries.) Of course, in the specific case of polysilicon, bulk diffusivity can be generally expected to be of a similar magnitude (maybe somewhat larger depending on processing conditions) as the corresponding diffusivity in single crystal silicon.

However, as is often the case in many materials, grain boundary diffusivity is typically much, much larger. Indeed, grain boundary diffusivity can be as much as six orders of magnitude larger than bulk diffusivity. However, it is obvious that the amount of the material volume occupied by grain boundaries is a relatively small fraction of the total material volume (unless the grains are extremely small). Therefore, the total diffusion flux can be represented as follows:

$$F = -D_{bulk}(1 - f_{GB})\frac{\partial C}{\partial x} - D_{GB} f_{GB} k_{GB} \frac{\partial C}{\partial x}$$

Here, D_{bulk} and D_{GB} are bulk and grain boundary diffusivities, f_{GB} is the fraction of material occupied by grain boundaries, and k_{GB} is a grain boundary segregation coefficient. Naturally, diffusivities can be expressed as Arrhenius forms, therefore:

$$F = -D_{obulk} e^{-Q_{bulk}/kT}(1 - f_{GB})\frac{\partial C}{\partial x} - D_{oGB} e^{-Q_{GB}/kT} f_{GB} k_{GB} \frac{\partial C}{\partial x}$$

Typically, the activation energy for grain boundary diffusion, Q_{GB}, is much smaller than the activation energy for bulk diffusion, Q_{bulk}. This is easily understood since no vacancy formation energy is required for grain boundary diffusion. It is often found to be the case that bulk diffusion will dominate at high temperatures just due to geometrical factors and that grain boundary diffusion will dominate at low temperatures due to the lower activation energy. The temperature for which bulk and grain boundary diffusion fluxes are equal is called the *Tamman temperature*.

Interaction of Diffusion and Oxidation Processes

Before considering interactions between diffusion and oxidation processes, it is useful to recall that diffusion in amorphous dielectrics, *e.g.*, thermal oxide, is conceptually a much simpler process than diffusion in crystalline semiconductors. Specifically, there is no interaction with mobile carriers to be considered and many species just diffuse through existing voids and spaces within the network structure. This is similar to interstitial diffusion in silicon, and consequently the activation energy can be expected to be quite low. (In particular, molecular species such as water and oxygen diffuse in thermal oxide in this way.) Alternatively, "network formers" such as boron, phosphorus, and arsenic are generally oxidized along with silicon and, accordingly, are "dissolved" in thermal oxide, *i.e.*, become directly incorporated into the glassy network structure. As such, network formers are analogous to substitutional impurities, *i.e.*, dopants, in crystalline silicon and, thus, diffuse much more slowly. Of course, one important difference between dielectrics and semiconductors is that very high electric fields can be sustained within dielectrics. Therefore, field enhanced diffusion due to drift in an externally applied electric field can become very important. This is particularly true for metallic species, such as sodium, that easily become ionized in thermal oxide.

Of course, in a bulk silicon crystal, vacancies and interstitials exist in a well-defined equilibrium. Therefore, both interstitial and vacancy concentrations can usually be treated as constant throughout the bulk. However, one important exception to this situation occurs in the vicinity of a growing thermal oxide interface. As observed elsewhere, during oxidation as many as one out of a thousand silicon atoms at the interface fail to become incorporated into the growing oxide film and become silicon self-interstitials. Clearly, the growing oxide interface provides a source of interstitial defects that then can readily diffuse into the bulk. Thus, in addition to condensing into extrinsic stacking faults, these interstitials contribute to enhancement of dopant diffusivity in proximity of the thermal oxide interface through the interstitialcy mechanism. This is called *oxidation enhanced diffusion*. The degree of enhancement depends largely on process conditions (temperature, oxidant species, oxidant pressure, substrate orientation, *etc.*). Observed enhancement is largest for [100] substrates and effectively absent for [111] substrates. (Of course, the reason for this orientation dependence has to do with the detailed structure of [100] and [111] interfaces.)

In addition to oxidation enhanced diffusion, extrinsically doped silicon can interact with a growing silicon dioxide layer through *dopant segregation*. This provides another example of a thermodynamic distribution equilibrium; however, instead of determining the relation of solute concentrations between solid and liquid phases of silicon (as in the case of crystal growth), this equilibrium determines the relation between solute concentrations in crystalline silicon and amorphous silicon dioxide. Thus, the dopant segregation coefficient, m_I, at the Si/SiO$_2$ interface has the fundamental definition:

$$\mathrm{m}_I = \frac{C_I}{C_I^{ox}}$$

Here, C_I^{ox} is the concentration of impurity, *I*, in thermal oxide and, of course, C_I is the concentration of the same impurity in silicon. Of course, as a thermodynamic equilibrium constant m_I is temperature dependent. Moreover, in contrast to solid-melt distribution equilibria important in crystal growth, it has been found experimentally that depending on specific impurity atoms, m_I may be either greater or less than unity. In particular, phosphorus, arsenic, and antimony all have m_I's about 10 and, thus, segregate preferentially into the silicon substrate. In contrast, m_I for boron (*i.e.*, m_B) has a value of 0.1 to 0.3 depending on conditions. (Some researchers have reported that the value of m_B is a function of both temperature and oxidizing ambient.) In any case, boron preferentially segregates into the oxide layer.

Naturally, combined effects of dopant segregation and diffusion (irrespective of oxidation enhanced or not) can be expected to affect surface doping concentrations if a thermal oxide layer is grown on extrinsically doped silicon. Moreover, it would seem transparently obvious that the degree to which surface doping is affected should strongly depend on relative rates of diffusion and oxidation. Indeed, this is the case and it is found that boron surface concentration can be substantially reduced if a thermal oxide is grown on a boron diffusion. Dopant segregation is illustrated in the following figure:

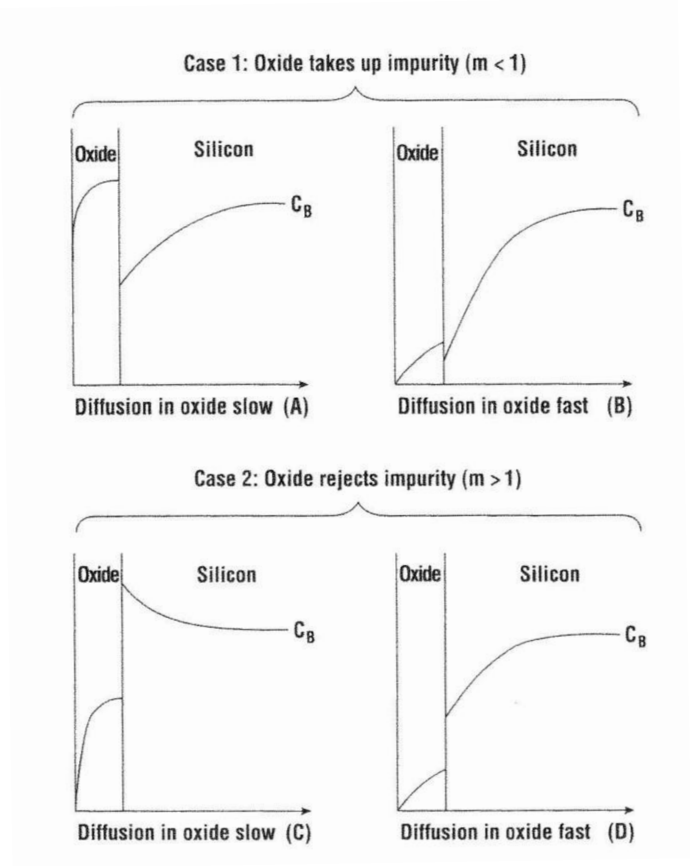

Case 1: Oxide takes up impurity (m < 1)
Oxide
Silicon
C_B
Diffusion in oxide slow (A)
Oxide
Silicon
C_B
Diffusion in oxide fast (B)
Case 2: Oxide rejects impurity (m > 1)
Oxide
Silicon
C_B
Diffusion in oxide slow (C)
Oxide
Silicon
C_B
Diffusion in oxide fast (D)

Fig. 64: Dopant redistribution during thermal oxidation

Conversely, because they tend to be rejected by the growing oxide, the surface concentration of *n*-type dopants such as phosphorus or arsenic can be greatly increased by thermal oxidation. In extreme cases, the surface concentration can exceed the solid solubility limit for the dopant species in which case precipitates form at the Si/SiO_2 interface. Of course, any precipitated dopant cannot be electrically active. In practice, such extreme effects must be avoided and it may become necessary to readjust the surface concentration by additional doping or diffusion.

Ion Implantation

Ion implantation is currently the method of choice for introduction of dopant species into semiconductor substrates for state-of-the-art integrated circuit fabrication. Indeed, this process affords much better control of the impurity concentration profile for shallow and/or low dose doping than is possible using classical non-implanted diffusion processes. To understand why this is so, one need only consider conventional furnace doping. In the case of a constant source, the surface concentration is effectively at the solubility limit during the entire process. Typically, this concentration is quite high which makes this method unsuitable for very low doses. Similarly, in the case of an instantaneous source generated by conventional pre-deposition, the dopant surface concentration is, again, effectively at the solubility limit. Furthermore, since dose cannot be controlled directly, *i.e.*, surface concentration, not dose, is determined by the solubility limit at a given temperature, small process variations, *e.g.*, in time or temperature, can cause relatively large variations in dose. Also, an additional complication may arise because after a pre-deposition, all of the dopant atoms are effectively on the surface of the substrate. In this situation, dose control can become difficult because dopant can evaporate from the substrate surface during a subsequent high temperature drive process. (This problem can be remedied by use of a “cover oxide” or other judicious modification of process conditions.) In contrast, ion implantation is not subject to any of these limitations. First of all, very small amounts of dopant can be accurately measured and precisely introduced into the silicon wafer. Moreover, dose rather than concentration is measured directly during ion implantation. Secondly, dopant atoms are deposited at some depth below the surface of the semiconductor and, therefore, they are much less easily lost during any subsequent diffusion drive (or other heat treatment).

Conceptually, ion implantation is extremely simple. Dopant containing source gases, *e.g.*, arsine (AsH_3), phosphine (PH_3), diborane (B_2H_6), *etc.*, are initially ionized in an electrical discharge within a vacuum chamber or “source”. A high voltage power supply is used to “extract” atomic ions from the discharge. These extracted ions are then sorted according to mass by dispersion in a strong magnetic field, *i.e.*, by magnetic mass spectrometry. The desired ions are then selected and electrostatically accelerated through a final high voltage stage to the desired energy. Typical acceleration energies range from a few thousand to a few hundred thousand electron-volts. The resulting pure, monoenergetic ion beam enters the “end station” and strikes the wafer (or *target*). Since the beam cross section is typically much smaller than the wafer diameter, to obtain a uniform dose, electrostatic deflection of the beam to “scan” the wafer is required. In addition, mechanical motion of the wafer stage can also be used to augment electrostatic scanning and to average out directional effects. Implanted dose is controlled by directly measuring beam current (using a “Faraday cup” or some other device) and integrating over time. Modern ion implanters can easily achieve doses below 10^{11} cm^{-2}. This is far lower than is achievable using conventional doping methods.

Of course, the nature of the interaction of an ion beam with a crystalline solid is strongly dependent on the incident ion kinetic energy. At very low energy, ions do not penetrate the target crystal and are just deposited on the surface, *i.e.*, so-called *ion beam deposition* or *ion beam plating*. At higher energy, the impinging ions deposit kinetic energy just in the first few surface layers. This causes *sputtering* which scatters atoms of the crystalline solid from the surface into the ambient and, thence onto surrounding surfaces. (Indeed, this is the underlying physical process for physical vapor deposition (PVD) of thin films.) At still higher energy, ions penetrate the surface and come to rest inside the crystal. This is, of course, ion implantation. At the highest energies, impinging ions penetrate very deeply and can be applied to sophisticated analytical techniques, *e.g.*, Rutherford back-scattering (RBS). Obviously, a collateral effect of penetration of ions into a crystalline solid is the creation of damage and disorder within the crystal lattice. Therefore, an important aspect of ion implantation processes requires control and/or elimination of these effects. (Indeed, ion implantation of species such as argon (Ar), silane (SiH_4), or germane (GeH_4), *etc.*, can be used intentionally to damage or *amorphize* the crystal for gettering or other purposes.)

Elementary Hard Sphere Collision Dynamics

As asserted previously, for ion kinetic energies in the range of a few thousand electron-volts to several hundred thousand (or even several million) electron-volts, ions impinging on a solid surface penetrate into the interior and then come to rest. (At kinetic energies above several million electron-volts, ions can completely pass through the solid and, also nuclear reactions may occur which induce radioactivity!) Obviously, when impinging ions penetrate a crystalline solid, they must collide with atoms inside the solid and lose energy. Furthermore, since kinetic energies of implanted ions are much higher than atomic bond energies of the solid, ion-atom collisions can generally be treated as isolated collisions between free particles, *i.e.*, atomic binding energy can be neglected. In the simplest picture, these collisions can be viewed as collisions between two hard spheres (*i.e.*, one treats ions and atoms as “billiard balls”). Although the actual ion-atom interaction is much more complicated, this simplistic approach provides at least a qualitative insight into the form of observed implant concentration profiles. (Of course, a hard sphere collision is only a simplified approximation to a real physical process; however, if collision energy is relatively large, a hard sphere approximation often gives quite reasonable results.)

To describe hard sphere collisions, one first defines a vector, $\mathbf{v}_i$, as incident velocity of an implanted ion and a second vector, $\mathbf{v}'_i$, as the resulting ion velocity after collision with a silicon atom. The initial silicon atom velocity vector, $\mathbf{v}_s$, is taken to be zero since it is reasonable to consider silicon atoms as initially at rest within the crystal lattice. The recoil velocity vector, $\mathbf{v}'_s$, of a silicon atom is, of course, its velocity after it is struck by an implanted ion. Other parameters which characterize a collision between an ion and a silicon atom are impact parameter, b, which is the “center-to-center miss distance” (by definition, if b vanishes the collision is “dead center”) and scattering angle, χ, which is the deflection angle of the incident ion due to the collision.

Of course, as in any two-body collision, total momentum must be conserved, therefore:

$$m_i \mathbf{v}_i = m_i \mathbf{v}'_i + m_s \mathbf{v}'_s$$

Here, m_i is ion mass and m_s is the mass of a silicon atom. (Of course, masses are unchanged by collision.) Likewise, kinetic energy must also be conserved, hence:

$$\frac{m_i v_i^2}{2} = \frac{m_i v_i'^2}{2} + \frac{m_s v_s'^2}{2}$$

(Here, v_i, v_i', and v_s' are just defined as the magnitudes of $\mathbf{v}_i$, $\mathbf{v}_i'$, and $\mathbf{v}_s'$, respectively.) From these two conditions, the characteristics of the collision are completely determined.

Within this context, the geometry of a collision of two hard spheres is illustrated by the following figure:

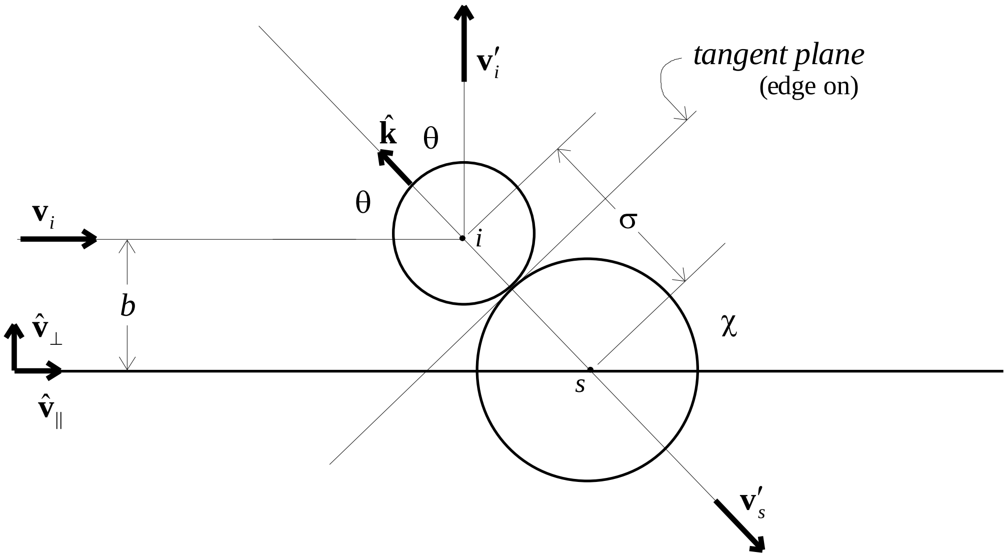


Fig. 65: Geometry of a collision between two hard spheres

Clearly, when two hard spheres collide, the entire interaction occurs as an instantaneous *impulse* just at the moment they "touch". Furthermore, this impulse is directed along a line connecting the centers of the two spheres. This direction defines the *apse* of the collision and is characterized by a unit vector, $\hat{\mathbf{k}}$, which by definition "points" outward from the center of the silicon atom at the moment of contact. Thus, if *G* is defined as a scalar quantity that corresponds to *impulse strength*, then incident, scattered, and recoil velocities are formally related as follows:

$$\mathbf{v}_i' = \mathbf{v}_i + \frac{G}{m_i}\hat{\mathbf{k}} \quad ; \quad \mathbf{v}_s' = -\frac{G}{m_s}\hat{\mathbf{k}}$$

Clearly, these expressions trivially satisfy momentum conservation. Thus, to determine *G*, one substitutes these two expressions into the formula for energy conservation to obtain the result:

$$m_i v_i^2 = m_i v_i^2 + 2G\mathbf{v}_i \cdot \hat{\mathbf{k}} + \frac{G^2}{m_i} + \frac{G^2}{m_s}$$

Obviously, $m_i v_i^2$ can be subtracted from both sides, hence:

$$-2\mathbf{v}_i \cdot \hat{\mathbf{k}} = G\left(\frac{1}{m_i} + \frac{1}{m_s}\right) = G\left(\frac{m_i + m_s}{m_i m_s}\right)$$

One explicitly solves for the impulse strength in terms of a dot product taken between the incident ion velocity and the apse unit vector:

$$G = -2\left(\frac{m_i m_s}{m_i + m_s}\right)\mathbf{v}_i \cdot \hat{\mathbf{k}}$$

As one intuitively might expect, *G* depends on atomic masses, incident ion velocity magnitude, and the *apse angle*, θ. (Clearly, $\cos\theta$ is just $\mathbf{v}_i \cdot \hat{\mathbf{k}}$ divided by v_i.) The quantity, $m_i m_s/(m_i+m_s)$, often appears in the classical mechanics of binary collisions and is called *reduced mass*. Accordingly, one can substitute for *G* to obtain expressions for the scattered ion velocity and the recoil velocity of a silicon atom:

$$\mathbf{v}_i' = \mathbf{v}_i - 2\left(\frac{m_s}{m_i + m_s}\right)\mathbf{v}_i \cdot \hat{\mathbf{k}}\hat{\mathbf{k}} \quad ; \quad \mathbf{v}_s' = 2\left(\frac{m_i}{m_i + m_s}\right)\mathbf{v}_i \cdot \hat{\mathbf{k}}\hat{\mathbf{k}}$$

Clearly, the dot product, $\mathbf{v}_i \cdot \hat{\mathbf{k}}$ is just determined by the geometry of the collision.

One observes from the collision geometry that $\hat{\mathbf{k}}$ can be decomposed into components parallel and perpendicular to incident ion velocity as follows:

$$\hat{\mathbf{k}} = \sin\theta\hat{\mathbf{v}}_\perp - \cos\theta\hat{\mathbf{v}}_{||}$$

Here, $\hat{\mathbf{v}}_{||}$ is a unit vector parallel to $\mathbf{v}_i$ and $\hat{\mathbf{v}}_\perp$ is a unit vector perpendicular to $\mathbf{v}_i$. The sine of the angle, θ, is just the ratio of impact parameter to the combined radii of the ion and silicon atom, σ. (Clearly, $\pi\sigma^2$ is *collision cross section*.) The negative cosine of θ appears in the above expression because, clearly, $\hat{\mathbf{v}}_{||} \cdot \hat{\mathbf{k}}$ must be formally negative, thus:

$$\sin\theta = \frac{b}{\sigma} \quad ; \quad \cos\theta = \sqrt{1 - \frac{b^2}{\sigma^2}}$$

These expressions are trivially substituted into the expression for $\hat{\mathbf{k}}$, to obtain:

$$\hat{\mathbf{k}} = \frac{b}{\sigma}\hat{\mathbf{v}}_\perp - \sqrt{1-\frac{b^2}{\sigma^2}}\hat{\mathbf{v}}_\parallel$$

Hence, it follows that:

$$\mathbf{v}_i \cdot \hat{\mathbf{k}} = \frac{b}{\sigma} v_i \hat{\mathbf{v}}_\parallel \cdot \hat{\mathbf{v}}_\perp - v_i \sqrt{1-\frac{b^2}{\sigma^2}}\hat{\mathbf{v}}_\parallel \cdot \hat{\mathbf{v}}_\parallel = -v_i \sqrt{1-\frac{b^2}{\sigma^2}}$$

This result is formally substituted back into the velocity expressions:

$$\mathbf{v}'_i = \mathbf{v}_i + 2\left(\frac{m_s}{m_i+m_s}\right) v_i \sqrt{1-\frac{b^2}{\sigma^2}}\hat{\mathbf{k}} \quad ; \quad \mathbf{v}'_s = -2\left(\frac{m_i}{m_i+m_s}\right) v_i \sqrt{1-\frac{b^2}{\sigma^2}}\hat{\mathbf{k}}$$

Obviously, the velocities can also be resolved in terms of parallel and perpendicular components:

$$\mathbf{v}'_i = v_i\hat{\mathbf{v}}_\parallel + 2\left(\frac{m_s}{m_i+m_s}\right) v_i \sqrt{1-\frac{b^2}{\sigma^2}}\left(\frac{b}{\sigma}\hat{\mathbf{v}}_\perp - \sqrt{1-\frac{b^2}{\sigma^2}}\hat{\mathbf{v}}_\parallel\right)$$

$$\mathbf{v}'_s = -2\left(\frac{m_i}{m_i+m_s}\right) v_i \sqrt{1-\frac{b^2}{\sigma^2}}\left(\frac{b}{\sigma}\hat{\mathbf{v}}_\perp - \sqrt{1-\frac{b^2}{\sigma^2}}\hat{\mathbf{v}}_\parallel\right)$$

It is convenient to collect terms as follows:

$$\mathbf{v}'_i = \left(1-2\left(\frac{m_s}{m_i+m_s}\right)\left(1-\frac{b^2}{\sigma^2}\right)\right) v_i\hat{\mathbf{v}}_\parallel + 2\left(\frac{m_s}{m_i+m_s}\right) v_i \frac{b}{\sigma}\sqrt{1-\frac{b^2}{\sigma^2}}\hat{\mathbf{v}}_\perp$$

$$\mathbf{v}'_s = 2\left(\frac{m_i}{m_i+m_s}\right)\left(1-\frac{b^2}{\sigma^2}\right) v_i\hat{\mathbf{v}}_\parallel - 2\left(\frac{m_i}{m_i+m_s}\right) v_i \frac{b}{\sigma}\sqrt{1-\frac{b^2}{\sigma^2}}\hat{\mathbf{v}}_\perp$$

In passing, one observes that the scattering angle, χ, is just $\pi-2\theta$.

If E_i is defined as the kinetic energy of the incident ion, *i.e.*, $m_i v_i^2/2$, kinetic energies of the scattered ion and recoil atom are easily determined from the preceding expressions:

$$E'_s = 4\left(\frac{m_i}{m_i+m_s}\right)^2\left[\left(1-\frac{2b^2}{\sigma^2}+\frac{b^4}{\sigma^4}\right)+\frac{b^2}{\sigma^2}\left(1-\frac{b^2}{\sigma^2}\right)\right]\frac{m_s v_i^2}{2} = \frac{4m_i m_s}{(m_i+m_s)^2}\left(1-\frac{b^2}{\sigma^2}\right)E_i$$

$$E'_i = E_i - E'_s = \left(1-\frac{4m_i m_s}{(m_i+m_s)^2}\left(1-\frac{b^2}{\sigma^2}\right)\right)E_i$$

Of particular interest is the energy transfer efficiency from the incident ion to the silicon atom. If one considers a collision for which m_i and m_s are equal:

$$E_s' = \left(1 - \frac{b^2}{\sigma^2}\right) E_i$$

However, if $m_i << m_s$, then:

$$E_s' \cong \frac{4m_i}{m_s}\left(1 - \frac{b^2}{\sigma^2}\right) E_i$$

Similarly, if $m_i >> m_s$, then:

$$E_s' \cong \frac{4m_s}{m_i}\left(1 - \frac{b^2}{\sigma^2}\right) E_i$$

Clearly, maximum energy transfer occurs for collisions for which the impact parameter vanishes. Furthermore, kinetic energy is most efficiently transferred if the mass of the ion is equal to the mass of a silicon atom. (For common dopants, this is most closely realized in the case of phosphorus.)

It is also instructive to consider the normal components of scattered ion velocity and recoil velocity of a silicon atom. It follows from the general result that:

$$\mathbf{v}_i' \cdot \hat{\mathbf{v}}_{||} = \left(1 - 2\left(\frac{m_s}{m_i + m_s}\right)\left(1 - \frac{b^2}{\sigma^2}\right)\right) v_i \quad ; \quad \mathbf{v}_s' \cdot \hat{\mathbf{v}}_{||} = 2\left(\frac{m_i}{m_i + m_s}\right)\left(1 - \frac{b^2}{\sigma^2}\right) v_i$$

Of course, ion implantation does not correspond to just a single collision, but many separate collisions. Naturally, the collisions that are most important in stopping an implanted ion are collisions with small impact parameter, *i.e.*, nearly centered collisions. Accordingly, for a perfectly centered collision, normal components of $\mathbf{v}_i'$ and $\mathbf{v}_s'$ are readily constructed as follows:

$$(\mathbf{v}_i' \cdot \hat{\mathbf{v}}_{||})_0 = \left(\frac{m_i - m_s}{m_i + m_s}\right) v_i \quad ; \quad (\mathbf{v}_s' \cdot \hat{\mathbf{v}}_{||})_0 = \left(\frac{2m_i}{m_i + m_s}\right) v_i$$

Clearly, if $m_i < m_s$, then $(\mathbf{v}_i' \cdot \hat{\mathbf{v}}_{||})_0$ is negative. This means that light implanted ions tend to be scattered back toward the surface, *i.e.*, back-scattered. Conversely, if $m_i > m_s$, then $(\mathbf{v}_i' \cdot \hat{\mathbf{v}}_{||})_0$ is positive and heavy ions tend to be scattered forward into the bulk, *i.e.*, forward-scattered. Obviously, if m_i equals m_s, then $(\mathbf{v}_i' \cdot \hat{\mathbf{v}}_{||})_0$ vanishes. This implies that in the case of equal ion and atom masses, a dead center collision stops the ion. In passing, it is worthwhile to observe that for a given ion kinetic energy, ion velocity magnitude has an inverse relationship with ion mass, *i.e.*, v_i equals $\sqrt{2E_i / m_i}$. Thus, the heavier the ion, the lower is its velocity for some equivalent kinetic energy. Therefore, one expects heavy ions to penetrate much less than light ions. This is indeed the case. Furthermore, since kinetic energy transfer is inefficient if the mass difference is large, for light ions, *e.g.*, boron, penetration is even more enhanced.

Implant Range and Straggle

The preceding formulation of hard sphere collisions provides a qualitative description of ion implantation; however an accurate description requires a more sophisticated model. Physically, there are two important scattering mechanisms for ions implanted into silicon. These are electronic and nuclear scattering. The terminology is self-explanatory. In general, electronic scattering corresponds to interaction of atomic electrons of impinging ions and silicon atoms. In descriptive terms, this process is somewhat similar to viscous drag in a liquid medium. Thus, the electrons can be considered to behave collectively much like a fluid. In contrast, nuclear scattering corresponds to direct interaction of ion and silicon atom nuclei. This process more closely resembles elementary hard sphere scattering. Of course, both of these scattering mechanisms ultimately bring the ion to rest, *i.e.*, stop the ion. Typically, it is found that *nuclear stopping* dominates at low energy and *electronic stopping* at high energy (although for light species, electronic stopping may dominate at all useful energies). Of course, exact details depend on substrate and implanted species. Extensive calculations have been made to describe both of these mechanisms within the context of dopant ion implantation into single crystal silicon.

In reality, the actual path that an ion follows during implantation may be quite convoluted (resembling a bolt of lightning). However, the average of the total integrated distance traveled by an ion is well-defined and is called *range*. Of course, what is of real interest in integrated circuit fabrication is not the total distance an ion travels, but is, rather, the average penetration depth normal to the silicon surface. This is the net distance an ion travels perpendicular to the substrate surface and is called *projected range*. Extensive tables and databases for projected range have been compiled for various dopant species implanted into silicon (as well as into silicon dioxide, silicon nitride, and photoresist). Additionally, there are several computational algorithms that use “Monte Carlo methods” to determine projected range.

Obviously, one does not expect that all of the ions will come to rest at exactly the projected range, but that there will be a distribution of penetration depths. The broadness of this distribution, measured normal to the surface (normal variance) is called *projected straggle*. If the implanted species distribution corresponds to a normal Gaussian, this is just the ordinary standard deviation. In addition, it is clear that implanted ions will have a lateral variance (not necessarily the same as the normal variance) as well. This is a measure of how far an ion moves laterally relative to the point of initial penetration of the substrate surface and is called *lateral straggle*. In general, the ratio of lateral straggle to projected straggle is greater than unity for light ions and is less than unity for heavy ions. Again, one can rationalize this behavior in terms of the propensity of light ions to back-scatter and heavy ions to forward-scatter. Typical values of range and straggle are summarized in the following figures:

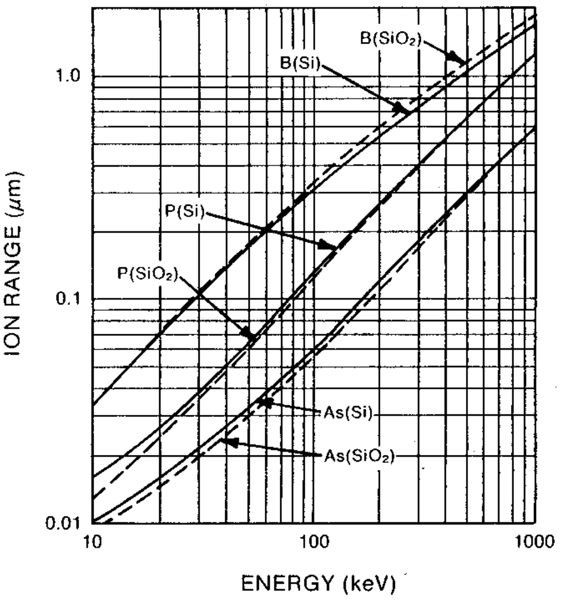


Fig. 66: Projected range in Si and SiO_2 for B, P, and As

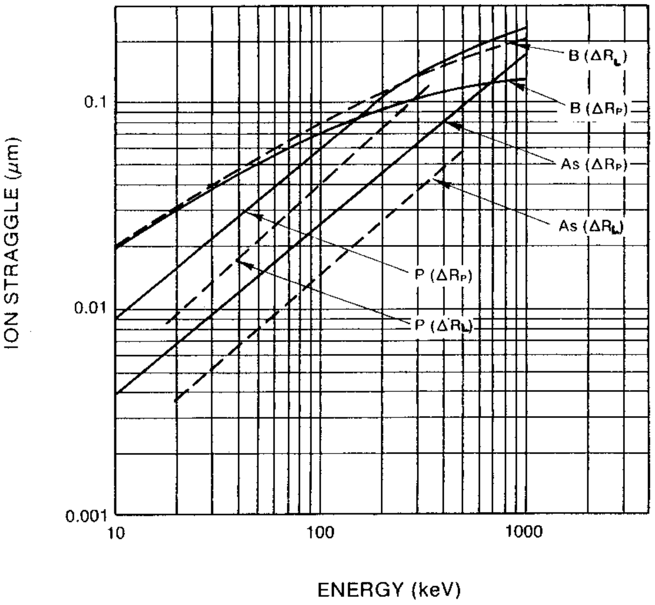


Fig. 67: Projected and later straggle in Si for B, P, and As

Obviously, for blanket implantation, lateral straggle is not an issue; however, for a patterned implant, it is important.

Summary of the LSS Model of Ion Implantation

Calculation of range and straggle for ions implanted into silicon is quite complicated and beyond the scope of the present course. However, Lindhard, Scharff, and Schiott have carried out numerical modeling of ion implantation (*i.e.*, the LSS model). Within the LSS model, dimensionless energy and range, ε and ρ, are defined as follows:

$$\varepsilon = \frac{E\alpha}{q^2}\left(\frac{M_s}{Z_i Z_s (M_i + M_s)}\right) \quad ; \quad \rho = R\pi\alpha^2 N \frac{M_i M_s}{(M_i + M_s)^2}$$

Here, *E* and *R* are incident ion kinetic energy and ion range, *N* is atomic density of the solid (for silicon this is eight divided by the lattice parameter cubed), M_i and Z_i are mass and atomic numbers of an implanted ion, M_s and Z_s are mass and atomic numbers of a substrate atom (for silicon these are 28 and 14, respectively), and α is a characteristic length given by the formula:

$$\alpha = \frac{0.8853 a_0}{\sqrt{Z_i^{2/3} + Z_s^{2/3}}}$$

The parameter, a_0, is the *Bohr radius, i.e.*, mean electronic orbital radius for a hydrogen atom, and has a value of 0.529 Å.

The functional relationship between ρ and ε has been calculated within the LSS model. The result is quite complicated, however, at low energy the relationship between ρ and ε can be approximated as a simple proportionality:

$$\rho = \kappa\varepsilon$$

Here, κ is a dimensionless constant. This expression can be rewritten in terms of *R* and *E* as follows:

$$R = \frac{1}{0.8853}\left(\frac{\kappa}{q^2 \pi a_0}\right)\frac{E}{N}\left(\frac{M_i + M_s}{M_i}\right)\left(\frac{\sqrt{Z_i^{2/3} + Z_s^{2/3}}}{Z_i Z_s}\right)$$

It is convenient to convert atom density, *N*, to substrate mass density, ρ_s, and collect remaining factors into an aggregate constant, *K*:

$$R = K\frac{E}{\rho_s}\left(\frac{M_s (M_i + M_s)}{M_i}\right)\left(\frac{\sqrt{Z_i^{2/3} + Z_s^{2/3}}}{Z_i Z_s}\right)$$

For silicon, *K* is found to have an approximate value of $6(10^{-7})$ g cm^{-2} keV^{-1}.

Projected range and projected straggle can also be approximated within the context of the LSS model. The following expressions are obtained:

$$R_P = \frac{3RM_i}{3M_i + M_s} \quad ; \quad \Delta R_P = \frac{2R_p\sqrt{M_iM_s}}{3(M_i + M_s)} = \frac{2RM_i\sqrt{M_iM_s}}{(M_i + M_s)(3M_i + M_s)}$$

These expressions are appropriate for estimation of implant range, projected range, and projected straggle, however if very precise values are needed they are inadequate. (More precise data is available in tabular, graphical, and/or electronic form.)

Implant Damage

It is clear from consideration of both simple hard sphere collision dynamics and the LSS model that as an implanted ion travels through a crystalline solid, a significant number of target atoms are displaced from lattice sites due to recoil. Furthermore, if the kinetic energy of the recoiling atom is sufficiently high, additional lattice atoms may be displaced due to recoil from secondary collisions. Thus, after ion implantation, the crystal lattice may be substantially damaged. Indeed, in the worst case, an implanted semiconductor crystal is completely amorphized, *i.e.*, all crystal ordering is destroyed. Within the context of the LSS model, nuclear and electronic stopping curves have been calculated and are shown below:

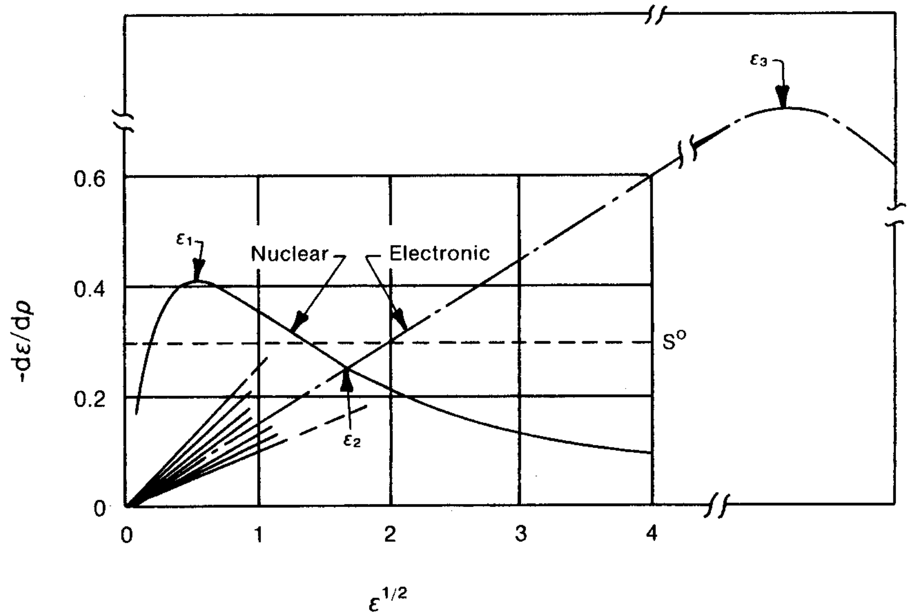


Fig. 68: Nuclear and electronic stopping as a function of dimensionless velocity

Of course, ρ and ε are identified as dimensionless range and energy, hence $\varepsilon^{1/2}$ corresponds to dimensionless velocity or momentum. In this figure, ε_1, ε_2, and ε_3 are dimensionless energies corresponding respectively to maximum nuclear stopping, equal nuclear and electronic stopping, and maximum electronic stopping. It is clear from the figure that, as previously asserted, nuclear stopping dominates at low ion energy (or velocity) and electronic stopping dominates at high ion energy (or velocity). Furthermore, since nuclear scattering directly implies displacement of atomic nuclei, most of the damage due to implantation can be attributed to the nuclear stopping mechanism. (Indeed, this is consistent with a view of nuclear scattering as resembling a simple two body collision.) In contrast, in electronic scattering, kinetic energy lost by an implanted ion causes electronic excitation rather than recoil. (In terms of a two body collision this corresponds to a large impact parameter or "grazing" collision.) Thus, most of the damage associated with implantation can be expected to occur at "end-of-range" (EOR) just before an implanted ion comes to rest. This further implies that the region of maximal damage due to implantation should occur at or near the maximum of the corresponding implanted concentration profile. This is indeed found to be the case. In particular, since light ions tend to back-scatter into the damaged region, the damage maximum and concentration maximum often closely coincide. For heavier ions, the damage maximum occurs closer to the surface than the concentration maximum due to forward-scattering. Furthermore, since the nuclear stopping is generally less dominant for light ions, implant damage is substantially less than in the case of heavy ions. Typically, even at moderate dose, ion implantation of a heavy species such as arsenic or antimony results in complete amorphization. In contrast, implantation of boron may result in severe lattice damage, but not in amorphization.

As indicated previously, sometimes it is desirable to damage the semiconductor crystal lattice intentionally without the introduction of dopant impurities. This may be for the purpose of gettering contamination, intentional pre-amorphization for the benefit later ion implantation or other processing, *etc.* In any case, this can be accomplished by implantation of silicon or germanium (or even tin). Clearly, these species do not act as either acceptor or donor impurities since they have a valence of four. Alternatively, argon ion implantation can be used to create damage. (Obviously, argon cannot substitute into the crystal lattice since it is a noble gas.)

Channeling

Range has been defined previously as the average integrated distance traveled by an implanted ion after it enters the solid and before it comes to rest. However, a hidden assumption in this simple definition is that the atoms of the crystalline solid are randomly distributed. Of course, this cannot be really true for a crystal since it is, by definition, characterized by a high degree of symmetry and order. Therefore, it comes as no surprise that if a crystallographic direction of the target solid, *i.e.*, a silicon wafer, coincides with the incident direction of incoming ions, then the range deviates from what is expected in an amorphous solid. In particular the range becomes much larger and the implanted concentration profile becomes extremely skewed into the bulk. This phenomenon is called *channeling* and is easily visualized as the result of an implanted ion following a relatively unimpeded path through the solid due to crystallographic order. Furthermore, since the diamond cubic structure is relatively open, one expects channeling to be a severe problem for implantation into single crystal silicon.

However, channeling can be greatly reduced by the simple expedient of misaligning implant and crystallographic directions. This is illustrated below for a high energy arsenic implant into [111] silicon:

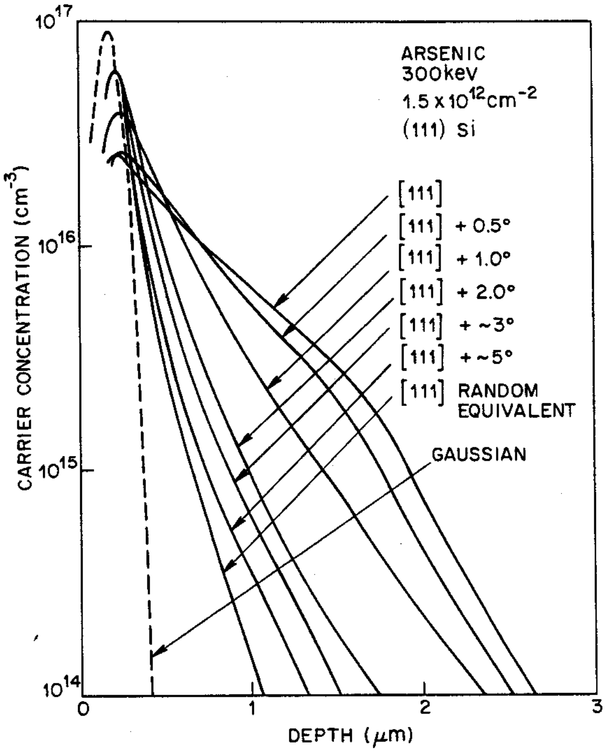


Fig. 69: Effect of channeling and tilt in a high energy arsenic implant

Several features are evident in this figure. First of all, as expected, in comparison to a purely Gaussian profile, the implanted arsenic concentration profile is naturally skewed into the bulk due to forward-scattering. Second, if the implant direction is exactly aligned normal to the [111] crystallographic direction, then channeling is severe. However, by misaligning implant and crystallographic directions by only 5°, channeling is almost entirely eliminated. Of course, implant misalignment can be achieved by simply tilting the substrate in the end station of the implanter. In addition, the substrates themselves can be sawed a few degrees off the exact crystallographic orientation during manufacture. Both of these are done in practice. Also, channeling can be reduced by use of an amorphous screening layer or, as indicated previously, a pre-amorphizing implant of silicon or germanium.

Practical Ion Implantation Processes

The general characteristics of a practical ion implantation process, including channeling and lattice damage are illustrated in the following figure for an antimony implant into silicon:

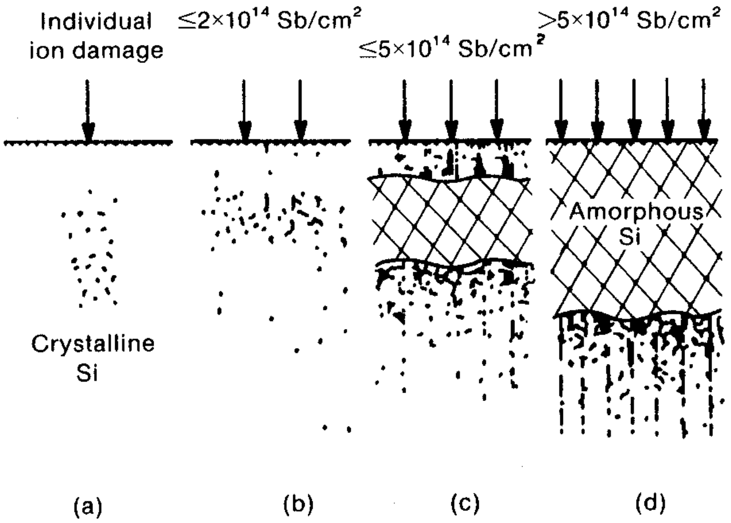


Fig. 70: Antimony implantation into single crystal silicon

If the dose is light, as in (a), then only isolated defects are generated within the silicon crystal. For a somewhat higher dose, (b), isolated defects begin to coalesce to form regions of damage. If the dose is made still higher, (c), then the damaged regions combine to form an amorphous layer below the substrate surface (relatively near the concentration maximum). Finally, for a very high dose, the amorphized region reaches all the way to the surface. In addition, the effect of channeling is shown in the above figure by the series of parallel damaged regions extending deep into the wafer. In some form, this behavior will be realized during implantation of any atomic species. (Of course, for a light species a fully amorphized region may not be formed unless the dose is extremely high.)

In practice, implantation processes are carried out at low temperature. Indeed, modern implanters have temperature controlled end stations that, among other things, allow photoresist to be used directly as an implant mask. (Formerly, implantation required the use of a silicon oxide or silicon nitride mask.) Of course, the thickness and stopping power of the masking material must be sufficient to block implanted species from reaching the interior of the wafer. In contrast, any masking material used in conventional diffusion processes must withstand the high ambient temperature of the furnace. Other problems that one may encounter in practice are "knock-on" and generation of oxide fixed charge or interface traps. Knock-on occurs when an implanted ion scatters an atom (usually oxygen) from a surface oxide layer, *e.g.*, a screen oxide) into the bulk. This creates interstitial defects, which tend to reduce carrier mobility. Typically, screen oxides have been used to collect impurities, *e.g.*, metal atoms, which can be deposited on the wafer surface due to secondary scattering within the implanter. More recently, improvements in equipment design have made screen oxides generally unnecessary. Likewise, implantation into a permanent oxide layer, *e.g.*, a field or other cover oxide, can create damage that ultimately appears as fixed charge or interface traps and, also, is not removed by subsequent heat treatment. Of course, these problems must be controlled by process and equipment design as well as by careful monitoring.

Equipment related issues can also create practical problems in an implantation process. Some of these include incorrect dose or energy, striping, or "blast defects". Obviously, incorrect dose or energy results if internal monitors either fail or become uncalibrated. In particular, measurement of dose is critical and requires integration of an effectively continuous measurement of "beam current" as a function of time. As observed previously, in conventional ion implanters the ion beam entering the end station has a much smaller radius than the wafer. Therefore, in order to implant species into the entire wafer, the beam must be "scanned" in a raster pattern across the wafer surface. This is accomplished by electrostatic deflection of the ion beam or a combination of electrostatic deflection and mechanical motion of the wafer itself. Striping results if the beam scan becomes misaligned. In such a case, either undoped or doubly doped implant stripes are formed. Blast defects are the result of accumulation of a large positive electrostatic charge on the wafer surface during implantation. These are formed when the accumulated charge catastrophically discharges as an arc. Blast defects quite literally appear as small pits or craters in the wafer surface. Obviously, some provision must be made to neutralize excess electrostatic charge built up on the wafer during implantation. This can be done by incorporating a neutralizing grid or "flood gun" within the end station.

Implantation of Molecular Species

In addition to atomic species, it is sometimes advantageous to implant molecular species. The most common example of this is implantation into silicon of the molecular ion, BF_2^+, derived from the source gas boron trifluoride (BF_3). Obviously, to control dose and energy, dissociation of the molecular ion in the beam itself must be avoided. However, once the ion enters the substrate it is quickly dissociated into boron and fluorine atoms. The primary reason for the use of BF_2^+ implantation is to allow the formation of shallow *p*-type regions. Naturally, since it is very light, even at low energy an atomic boron ion penetrates deeply into the silicon substrate. In addition, boron ions are also very prone to channeling. Therefore, it is quite difficult to form a shallow *p*-type region using conventional boron implantation. In contrast, for the purpose of ion implantation into a silicon substrate, BF_2^+ behaves as a heavy species and gives a much shallower implanted concentration profile. Furthermore, the associated implanted fluorine atoms generally do not degrade device performance. Fluorine is generally lost from the wafer during subsequent high temperature processing since silicon fluorides are quite volatile. However, care must be taken if the dose is very high. In this case, there is some evidence of complex (or compound) formation between silicon, boron, and fluorine, which degrades the quality of the silicon crystal. Obviously, due to these additional complexities, BF_2^+ implantation into silicon should not be used unless it is found to be absolutely necessary.

Implanted Concentration Profiles and Subsequent Diffusion

When a pure, monoenergetic bean of ions is implanted into a silicon substrate, as one might expect, the concentration has well-defined depth dependence. To first order, this concentration depth profile can be considered as a Gaussian of the form:

$$C(x) = C_{\max} e^{-(x-R_p)^2 / 2\Delta R_P^2}$$

Clearly, this concentration profile is similar to an instantaneous source diffusion profile except that the maximum concentration occurs at a depth or distance, R_P, below the surface, *i.e.*, at the projected range. The maximum concentration is unknown, but may be determined in terms of the *total dose*, N_{tot}. Of course, N_{tot} is obtained by integrating over the concentration profile as follows:

$$N_{\text{tot}} = \int_0^\infty dx C(x) \cong C_{\max} \int_{-\infty}^{\infty} dx e^{-(x-R_p)^2 / 2\Delta R_P^2} = C_{\max} \Delta R_P \sqrt{2\pi}$$

Here, it has been assumed that maximum of the implanted concentration profile is significantly below the surface (thus, allowing the limits of the integral to be uncritically regarded as $-\infty$ to ∞ instead of 0 to ∞).

$$C(x) = \frac{N_{\text{tot}}}{\Delta R_P \sqrt{2\pi}} e^{-(x-R_p)^2 / 2\Delta R_P^2}$$

Projected range and straggle are obtained either by calculation or from tabulated data.

Skewed Profiles

Of course, a Gaussian form for an implanted concentration profile is just a first order approximation to the actual profile. In practice, implanted profiles generally exhibit *skewness* due to the scattering characteristics of ions by substrate atoms. By definition, skewness is a measure of the asymmetry of some distribution about the distribution average. (Clearly, an implanted concentration profile can be regarded as the distribution of ion depths.) As was found in the simple hard sphere model, lighter ions tend to back-scatter toward the substrate surface, while heavy ions tend to forward-scatter into the bulk. This causes implant profiles for light ions, *e.g.*, boron, to be skewed toward the surface and implant profiles for heavy ions, *e.g.*, arsenic, to be skewed toward the bulk. If skewness is not too extreme, implanted concentration profiles may be constructed by joining two Gaussian profiles at some “model range”, R_M:

$$\frac{C(x)}{\int dxC(x)}=\begin{cases}\dfrac{2}{(\Delta R_{<}+\Delta R_{>})\sqrt{2\pi}}e^{-(x-R_M)^2 / 2\Delta R_{>}^2} & x\geq R_M\\[2ex] \dfrac{2}{(\Delta R_{<}+\Delta R_{>})\sqrt{2\pi}}e^{-(x-R_M)^2 / 2\Delta R_{<}^2} & x< R_M\end{cases}$$

Here, the model range and projected range are related by the empirical expression:

$$R_M = R_P - 0.8\left|\Delta R_{>} - \Delta R_{<}\right|$$

Similarly, the projected straggle corresponds to just the average straggle:

$$\Delta R_P = \frac{\Delta R_{>} + \Delta R_{<}}{2}$$

Typically, profile parameters are obtained from experimental data.

Four Moment Profiles

Of course, implanted profiles may deviate significantly from a simple Gaussian and in addition to skewness, they may exhibit *kurtosis*. By definition, kurtosis is a measure of "peak sharpness" of a distribution or profile. If skewness and kurtosis are both significant, a four moment distribution is required to describe implanted concentration profiles adequately. The *Pearson distribution function*, $f(x)$, is often used for this purpose, which by definition, satisfies the first order differential equation:

$$\frac{df}{dx}=\frac{(x-a)f}{b_0+b_1x+b_2x^2}$$

Naturally, the Pearson distribution is unit-normalized:

$$1=\int_{-\infty}^{\infty}dxf(x)$$

Accordingly, projected range, R_P, projected straggle, ΔR_P, skewness, γ, and kurtosis, β, are defined by the first four moments of the distribution as follows:

$$R_P=\int_{-\infty}^{\infty}xdxf(x)$$

$$\Delta R_P=\int_{-\infty}^{\infty}dx(x-R_P)^2 f(x)$$

$$\gamma = \frac{1}{\Delta R_P^3} \int_{-\infty}^{\infty} dx (x - R_P)^3 f(x)$$

$$\beta = \frac{1}{\Delta R_P^4} \int_{-\infty}^{\infty} dx (x - R_P)^4 f(x)$$

These are related to the four parameters, a, b_0, b_1, and b_2, by the expressions:

$$a = b_1 = -\frac{\gamma \Delta R_P^2 (\beta + 3)}{10\beta - 12\gamma^2 - 18}$$

$$b_0 = -\frac{\Delta R_P^4 (3\gamma^2 - 4\beta)}{10\beta - 12\gamma^2 - 18}$$

$$b_2 = -\frac{6 + 3\gamma^2 - 2\beta}{10\beta - 12\gamma^2 - 18}$$

Moreover, if the "b-parameters" satisfy the inequality, $0 < b_1^2 / 4b_0 b_2 < 1$, then the Pearson distribution can be represented as an analytical form. In this case, an explicit formula may be written for the implanted concentration profile as follows:

$$C(x) = C_o \exp\left\{ \frac{1}{2b_2} \ln(b_2 x^2 + b_1 x + b_0) - \frac{2b_2 a + b_1}{b_2 \sqrt{4b_2 b_0 - b_1^2}} \arctan\left(\frac{2b_2 x + b_1}{\sqrt{4b_2 b_0 - b_1^2}} \right) \right\}$$

This formula can be used to fit very accurately experimental implanted concentration profiles. (Of course, C_o, is determined by integrating the distribution to obtain the total dose.)

Two-Dimensional Profiles

Although "blanket" implantation into unpatterned wafers is sometimes applicable, it is more common to implant wafers covered with some thin film, *e.g.*, thermal oxide, patterned with open "windows" etched down to the underlying silicon wafer surface. In practice, this overlying film is sufficiently thick so that it serves as a mask for implanted ions allowing them to penetrate into the silicon substrate in some areas, but not in others. In this case, it is important to understand the distribution of implanted ions at pattern edges. Accordingly, if one considers a hypothetical case in which all impinging ions enter the surface of the substrate at a single definite point, it is clear that the "vertical" depth profile, $C(x)$, as is required can still be represented as a simple Gaussian, skewed Gaussian, or four moment form. However, since scattering is a random process, it is further reasonable to assume that a "horizontal" or lateral profile can be represented as a simple Gaussian. Thus, the corresponding two-dimensional profile is easily constructed as the formal product of depth and lateral profiles:

$$C_{2\mathrm{D}}(x,y) = \frac{C(x)}{\sqrt{2\pi\Delta R_L}} e^{-y^2 / 2\Delta R_L^2}$$

Here, y is horizontal distance from the "penetration point" and, of course, ΔR_L is lateral straggle. Clearly, the lateral concentration profile must be radially symmetric about the penetration point, *i.e.*, y equal to zero.

To describe a realistic situation, this hypothetical distribution must be combined with some appropriate "masking" function that describes the detailed window shapes in the overlying thin film mask. This can be very complicated, however, if one considers the case of an indefinitely long "stripe" having edges at ±*a*, the masking function is a simple "step function" form, hence:

$$C(x,y) = \int_{-a}^{a} dy' C_{2\mathrm{D}}(x, y-y') = \frac{C(x)}{\sqrt{2\pi\Delta R_L}} \int_{-a}^{a} dy' e^{-(y-y')^2 / 2\Delta R_L^2}$$

Here, $C(x,y)$ is not a "true" two-dimensional concentration profile, but is a depth profile taken at a distance, y, measured perpendicularly from the "centerline" of the stripe. Although the integral cannot be constructed analytically, it can be represented in terms of complementary error functions, thus:

$$C(x,y) = \frac{C(x)}{\sqrt{2\pi\Delta R_L}} \int_{y-a}^{y+a} dy'' e^{-y''^2 / 2\Delta R_L^2} = \frac{C(x)}{\sqrt{2\pi\Delta R_L}} \left(\int_{y-a}^{\infty} dy'' e^{-y''^2 / 2\Delta R_L^2} - \int_{y+a}^{\infty} dy'' e^{-y''^2 / 2\Delta R_L^2} \right)$$

$$C(x,y) = \frac{C(x)}{2} \left(\mathrm{erfc}\left(\frac{y-a}{\Delta R_L \sqrt{2}} \right) - \mathrm{erfc}\left(\frac{y+a}{\Delta R_L \sqrt{2}} \right) \right)$$

Clearly, this functional form illustrates that the concentration of implanted species does not abruptly cut off at feature edges, but instead "feathers out" due to lateral straggle.

Penetration of Masking and Screening Layers

Of course, characteristics of an implanted concentration profile for an ionic species of a given incident kinetic energy is dependent on the composition of the substrate due to differences in stopping mechanisms. Although not an exact relationship, range and straggle both typically exhibit an inverse relationship to substrate density. Thus, range and straggle in low density materials such as photoresist, are large. In contrast, range and straggle in high density materials, *e.g.*, heavy metals, are small. This is important for practical ion implant processes, because various materials are used as masking layers for ion implantation. In addition, materials such as silicon dioxide or polycrystalline silicon are also used as screening layers. Clearly, the only difference between screening and masking layers is that a screening layer is intended to allow a significant portion of the implant to penetrate into the underlying silicon substrate, but a masking layer is not. As a practical matter, implant profiles within screening or masking layers can be estimated by scaling vertical distances relative to silicon using the ratio of projected ranges.

An accurate description of the implant profile near a feature edge requires detailed knowledge of the masking function, *M*, which along with masking material thickness, x_t, takes into account mask composition, wall angle, implant angle, *etc.* An appropriate expression might be constructed as follows:

$$C(x,y_1,y_2)=\frac{C(x)}{\sqrt{2\pi\Delta R_L}}\int_{-\infty}^{\infty}dy_1'\int_{-\infty}^{\infty}dy_2'M(y_1',y_2';x_t,\theta,\cdots)e^{-(y_1-y_1')^2\big/2\Delta R_L^2}e^{-(y_2-y_2')^2\big/2\Delta R_L^2}$$

Here, the radial distance, *y*, has been resolved into two mutually perpendicular cartesian distances, y_1 and y_2. (Clearly, this is required to treat rectangular shaped features.)

Diffusion of a Gaussian Implant Profile

In general, an implanted concentration profile serves as the initial condition for subsequent diffused concentration profiles. It is quite straightforward to apply the general solution of the diffusion equation constructed previously:

$$C(x,t)=\frac{1}{2\sqrt{\pi Dt}}\int_{-\infty}^{\infty}dx'C(x')e^{-(x-x')^2\big/4Dt}$$

Of course, $C(x)$ is just the implanted diffusion profile. For a general implanted profile, the diffused concentration profile, $C(x,t)$, can only be obtained by numerical evaluation of the above expression. However, in the case that $C(x)$ is substantially Gaussian, the integral can be constructed explicitly if one substitutes as follows:

$$C(x,t)=\frac{N_{\text{tot}}}{2\pi\Delta R_P\sqrt{2Dt}}\int_{-\infty}^{\infty}dx'e^{-(x'-R_p)^2\big/2\Delta R_P^2}e^{-(x-x')^2\big/4Dt}$$

One proceeds by combining exponential functions and completing the square in the resulting argument. The detailed manipulations of the exponents are as follows:

$$\frac{(x'-R_P)^2}{2\Delta R_P^2}+\frac{(x-x')^2}{4Dt}=\frac{x'^2-2x'+R_P^2}{2\Delta R_P^2}+\frac{x^2-2xx'+x'^2}{4Dt}=$$

$$\frac{1}{2}\left(\frac{2Dt(x'^2-2x'+R_P^2)+\Delta R_P^2(x^2-2xx'+x'^2)}{2\Delta R_P^2 Dt}\right)=$$

$$\frac{1}{2}\left(\frac{x'^2(\Delta R_P^2+2Dt)-2x'(2R_PDt+\Delta R_P^2x)+(2R_P^2Dt+\Delta R_P^2x^2)}{2\Delta R_P^2Dt}\right)$$

Next, one defines a new integration variable, *u*:

$$u=\frac{x'}{2\Delta R_P}\sqrt{\frac{\Delta R_P^2+2Dt}{Dt}} \qquad ; \qquad x'=2u\Delta R_P\sqrt{\frac{Dt}{\Delta R_P^2+2Dt}}$$

Upon substitution of *u*, the exponents take the form:

$$\frac{(x'-R_P)^2}{2\Delta R_P^2}+\frac{(x-x')^2}{4Dt}=u^2-2u\left(\frac{2DtR_P+\Delta R_P^2x}{2\Delta R_P\sqrt{Dt(\Delta R_P^2+2Dt)}}\right)+\frac{2R_P^2Dt+\Delta R_P^2x^2}{4\Delta R_P^2Dt}$$

At this point, one completes the square in the *u* terms as follows:

$$\frac{(x'-R_P)^2}{2\Delta R_P^2}+\frac{(x-x')^2}{4Dt}=\left(u-\frac{2DtR_P+\Delta R_P^2x}{2\Delta R_P\sqrt{(\Delta R_P^2+2Dt)Dt}}\right)^2-$$

$$\frac{4D^2t^2R_P^2+4DtR_P\Delta R_P^2x+\Delta R_P^4x^2}{4Dt(\Delta R_P^2+2Dt)}+\frac{2DtR_P^2+\Delta R_P^2x^2}{4\Delta R_P^2Dt}\left(\frac{\Delta R_P^2+2Dt}{\Delta R_P^2+2Dt}\right)$$

The last two terms can be formally combined to yield the result:

$$\frac{(x'-R_P)^2}{2\Delta R_P^2}+\frac{(x-x')^2}{4Dt}=\left(u-\frac{2DtR_P+\Delta R_P^2x}{2\Delta R_P\sqrt{(\Delta R_P^2+2Dt)Dt}}\right)^2-\frac{x^2-2xR_P+R_P^2}{2(\Delta R_P^2+2Dt)}$$

Clearly, the last term has the form of a perfect square. Naturally, one substitutes back into the original expression to obtain:

$$C(x,t)=\frac{N_{\text{tot}}}{\pi\sqrt{2(\Delta R_P^2+2Dt)}}\exp\left(-\frac{(x-R_P)^2}{2(\Delta R_P^2+2Dt)}\right)\int_{-\infty}^{\infty}du\exp\left(-\left(u-\frac{2DtR_P+\Delta R_P^2x}{2\Delta R_P\sqrt{(\Delta R_P^2+2Dt)Dt}}\right)^2\right)$$

The integral can be constructed explicitly and has a value of exactly $\sqrt{\pi}$. Therefore, it follows that:

$$C(x,t)=\frac{N_{\text{tot}}}{\sqrt{2\pi(\Delta R_P^2+2Dt)}}\exp\left(-\frac{(x-R_P)^2}{2(\Delta R_P^2+2Dt)}\right)$$

If one compares this result with the original expression for an instantaneous source profile, it is clear that the result of diffusion of a Gaussian implant profile is also a Gaussian profile. However, this profile differs from an instantaneous source profile in two aspects. First, the origin of the Gaussian is not at the wafer surface, *i.e.*, $x=0$, but rather is at $x=R_P$. Second, the standard deviation is not simply $\sqrt{2Dt}$ as expected from diffusion only, but corresponds to the combined variances due to both implant and diffusion, $\sqrt{\Delta R_P^2 + 2Dt}$. This is easily understood if one again considers an instantaneous source. Clearly, a Gaussian profile generated by some simple linear diffusion process remains Gaussian during any subsequent diffusion. This is true even if the diffusivity changes (due to changing the temperature, *etc.*) provided that no non-linearities arise. Naturally, the standard deviation of the profile increases during subsequent diffusion. Thus, diffusion of a Gaussian implant profile is very similar to diffusion of a Gaussian diffusion profile, *i.e.*, an instantaneous source diffusion.

Thermal Budget

The preceding observations are quite general and allow estimation of the total variance of any doping profile due to subsequent thermal processing. Such estimates are typically cast in terms of total "*Dt*". Simply stated, total *Dt* is obtained by considering all high temperature individual process steps (oxidations, diffusions, *etc.*), multiplying dopant diffusivity at the temperature of each step by the associated process time, and adding all of these results together, thus:

$$(Dt)_{\text{tot}} = \sum_i D(T_i)t_i$$

As asserted previously, the square root of $2(Dt)_{\text{tot}}$ is the expected average displacement of an arbitrary dopant atom due to all subsequent thermal processing. Therefore, for classical diffusion processes, the total variance of the dopant concentration profile is estimated by the expression:

$$\sigma_{\text{tot}} = \sqrt{2(Dt)_{\text{tot}}}$$

Of course, for combined implant-diffusion processes, projected straggle must also be included:

$$\sigma_{\text{tot}} = \sqrt{\Delta R_P^2 + 2(Dt)_{\text{tot}}}$$

Clearly, σ_{tot} can be used to estimate how much particular junction depths will change due to thermal processing.

Obviously, control of thermal processing is a critical issue for any practical chip fabrication process. Within this context, it is typical for total allowable *Dt* or "thermal budget" to be set by the most critical diffusion profile (or junction depth). Of course, this is ultimately determined by the desired performance of devices in the finished circuit.

For example, junction depths of source/drain diffusions in CMOS devices cannot be allowed to become too large, otherwise transistors will suffer from unacceptable “short channel effect”. Similarly, adequate separation must be maintained between device diffusions and isolation diffusions. If this separation becomes too small or vanishes altogether (*i.e.*, diffusions overlap), then large “parasitic” junction capacitances will be introduced into the circuit resulting in an overall lowering of speed and collateral loss of performance. Indeed, there are many other critical criteria that can serve to set thermal budget. In general, the prevailing trend is toward progressively smaller and smaller thermal budgets.

Implant Activation and Removal of Damage

In contrast to classical non-implant, diffusion processes, implanted dopant atoms are not initially substituted into the crystal lattice. Of course, a dopant atom cannot be *electrically active*, *i.e.*, act as an acceptor or donor impurity, unless it occupies a silicon atom site within the diamond cubic crystal lattice. Therefore, additional, post-implant heat treatment is required for *implant activation*, *i.e.*, to cause dopant atoms to substitute into the crystal lattice. In addition, after ion implantation, the crystal lattice is very likely to be substantially damaged. Indeed, in the case of complete amorphization all crystal ordering is lost. This might seem at first to represent a catastrophic situation, however it has been found that subsequent, moderate heat treatment results in efficient recrystallization of amorphous regions as well as substitution of dopant atoms into lattice sites. Unfortunately, repair of crystal damage outside of amorphized regions is more difficult. Thus, some unrepaired, EOR defects often remain after post-implant annealing. Accordingly, it is critical to design the overall implantation process so that residual EOR defects do not degrade finished device performance.

In the case of implantation of a light species such as boron into silicon, damage repair is particularly difficult since amorphous regions are likely not to be present at all. Indeed, in severe cases, recrystallization of a heavily boron implanted region can result in polycrystalline rather than single crystalline silicon. Obviously, if any grain boundaries extend through a junction, then the corresponding device will fail due to reverse leakage through the junction. Typically, this problem is avoided by a pre-amorphizing implant of silicon or germanium followed by boron implantation. (In addition, pre-amorphization should reduce any channeling.)

Furnace Annealing and Transient Enhanced Diffusion

The usual method for post implantation anneals has been the use of a conventional quartz tube furnace. In this case, ion implanted wafers are annealed in an inert ambient such as nitrogen or argon. Clearly, such a process is very similar in principle to a classical diffusion drive. Indeed, as described previously, out-diffusion from an implanted region is often desirable and within this context, ion implantation takes the place of a classical pre-deposition. However, as observed previously, in contrast to dopant pre-deposition, ion implantation always results in some damage to the crystal lattice. Moreover, it is found that, depending on the location and severity of the damaged region, dopant diffusion during subsequent annealing can be significantly enhanced. However, this enhancement is temporary and persists only until recrystallization is substantially complete. Even so, the resulting concentration profile can be substantially different than one would expect for an undamaged lattice. This phenomenon is called *transient enhanced diffusion* or TED. Physically, TED arises because the activation energy for diffusion in damaged silicon is much lower than in undamaged silicon. This is easily understood within the context of the vacancy mechanism. Obviously, the number of vacancies per unit volume will be much higher in regions of damage. (In this sense, TED is similar to diffusion in polycrystalline silicon.) The practical result is that implanted species initially diffuse much faster than expected and resulting concentration profiles and junction depths become, respectively broader and deeper.

Physically, activation energy for solid phase epitaxy (SPE), *i.e.*, recrystallization of an amorphized layer overlying a crystalline layer, is about 2.3 eV for silicon. In contrast, activation energy for generation and diffusion of point defects is about 5 eV. Thus, if implant conditions are sufficient to create a fully amorphized layer, recrystallization occurs readily at moderate temperature by SPE. Indeed, annealing temperatures should be kept low to promote SPE since repair of the lattice by this mechanism is essentially complete. At higher temperatures, point defect generation and diffusion compete with SPE and damage repair is not as efficient. However, once amorphous regions have been recrystallized, if possible the temperature should be raised to above 900°C to "anneal out" any remaining defects. Clearly, if an amorphous region is not produced by implantation, higher temperatures are required for damage repair since SPE does not occur. In this case, annealing a partially damaged layer at low temperature is undesirable since stable extended defects such as dislocation loops can be formed by condensation of point defects. These stable defects are very difficult to remove even in a subsequent high temperature anneal. In the absence of an amorphized region, full activation of implanted dopants, particularly boron, typically requires a temperature in excess of 950°C, although, partial activation occurs as low as 450°C. If an amorphous layer exists, activation occurs along with recrystallization since dopant atoms are readily incorporated into the lattice.

Rapid Thermal Annealing

Clearly, furnace annealing requires a reasonably long process time, typically a least a fraction of an hour to, perhaps, several hours. This is particularly troublesome if TED as well as ordinary diffusion is to be minimized. This limitation can be overcome by rapid thermal annealing or RTA. Practical RTA systems utilize high intensity infrared/optical radiation which heats the substrate to a very high temperature in a very short period of time, typically a few seconds. As a consequence, the most important characteristic of rapid thermal annealing is that many physical processes do not have time to come to equilibrium. This is especially important in the case of implant annealing and activation. As observed previously, activation energy of recrystallization of amorphized material, *i.e.*, solid phase epitaxy or SPE, is quite low in comparison to the activation energy of diffusion processes. Of course, low activation energy implies that the rate constant of the process is large. Therefore, recrystallization occurs much more rapidly than diffusion and amorphized regions are rapidly recrystallized with substantially less dopant diffusion than can be achieved using conventional furnace annealing. For this reason, RTA is widely used for implant anneals.

To consider RTA in more detail, one observes that there are three natural mechanisms of heat transfer. These are *conduction*, *convection*, and *radiation*. Of course, conduction is a linear transport process analogous to diffusion and as asserted previously, the rate that heat energy is transported is proportional to the gradient in temperature, *i.e.*, Fourier's Law. This can be recast in a form corresponding to the usual expression of Ohm's Law:

$$\Delta T = R_{th} W$$

Here, ΔT is temperature difference (or drop) between the substrate and, perhaps, a hot chuck on which the substrate rests. Clearly, ΔT is analogous to a voltage drop in the ordinary electrical case of Ohm's Law. Accordingly, W is "thermal current" and R_{th} is thermal resistance. (Naturally, the relationship between thermal resistance and thermal conductivity is analogous to the relationship between ordinary electrical resistance and electrical conductivity, *i.e.*, reciprocal resistivity.) Obviously, to achieve rapid heating, thermal resistance must be made as low as possible. In practical systems this is achieved by various clamping mechanisms and/or introduction of gas with a high thermal conductivity, *e.g.*, helium, between the hot chuck and the back of the substrate. Indeed, these techniques are not limited to RTA systems, but are useful in any situation for which rapid thermal equilibration is desirable.

Convection can be identified with macroscopic flow of a working fluid between some heat source and a cold object. As a consequence of the flow, heat energy is transported. Obviously, convection cannot occur in a vacuum. Natural examples of convection can be observed on many different scales, which range from a pot of water boiling on a kitchen stove to thunderstorms, tectonic movement of the continental plates, and transport of heat from the core to the surface of the sun. In all cases, the basic phenomenon is the same: the temperature of some fluid is raised by contact with a heat source. This results in a change in density of the fluid and a corresponding mechanical disequilibrium, usually due to the influence of gravity. The disequilibrium causes the fluid to move toward the cold object, which becomes heated as a consequence of heat transfer from the fluid. The density of the cooled fluid is again changed, which results in macroscopic motion of the working fluid back toward the heat source. Again, a linear relationship can be used to describe natural convection:

$$Q = h_{con}(T_{hot} - T_{cold}) = h_{con}\Delta T$$

Here, Q represents the rate of heat transfer, h_{con} is some convective heat transfer coefficient, and, of course, ΔT is, again, temperature difference. Obviously, this expression is analogous to ordinary expressions for mass transfer. This comes as no surprise since convective heat transfer requires a collateral mass transfer of the working fluid. Convection can also be augmented by the forced flow of a hot fluid. In this case, a mechanical pump is used to enhance or even replace convective flow. This might be considered as a fourth form of heat transport.

Neither conduction nor convection, irrespective of whether natural or forced, is generally able to deliver sufficient heat energy to a small substrate in the short time required for RTA processes. Therefore, radiation is the primary mechanism of heat transfer for rapid thermal annealing. Of course, the total *exitance* of a perfectly "black body" is described by the well-known *Stefan-Boltzmann equation*:

$$M(T) = \varepsilon\sigma T^4$$

By definition, total exitance, $M(T)$, is the thermal power per unit area radiated by a black body with absolute temperature, T, ε is emissivity, which for a black body is independent of radiation wavelength, and σ is the Stefan-Boltzmann constant, which has a value of $5.6697(10^{-8})$ W/m^2 °K^4.

As one might expect, emissivity of real material objects is dependent on wavelength. In this case, spectral exitance, $M_\lambda(T)$, corresponds to *Planck's radiation law*:

$$M_\lambda(T) = \varepsilon(\lambda, T)\frac{\kappa_1}{\lambda^5\left(e^{\kappa_2/\lambda T} - 1\right)}$$

Here, κ_1 and κ_2 are the first and second radiation constants, which have values of $3.7142(10^{-16})$ Wm^2 and $1.4388(10^{-2})$ m°K, respectively. Obviously, total exitance is obtained by integration of $M_\lambda(T)$ over all wavelengths. (If emissivity is independent of wavelength, *i.e.*, the body is "black", and, naturally, the result of this integration is just the Stefan-Boltzmann equation.) Of course, $\varepsilon(\lambda,T)$ is emissivity at a specified wavelength, λ, and, moreover, emissivity may also generally depend on temperature. Therefore, when radiation falls on a surface, it may be reflected, absorbed, or transmitted and fundamental energy conservation requires that:

$$\varepsilon(\lambda, T) = 1 - \rho(\lambda, T) - \tau(\lambda, T)$$

By definition, $\rho(\lambda,T)$ and $\tau(\lambda,T)$ are reflectance and transmittance. Obviously, for opaque materials, transmittance vanishes.

To estimate the rate of heat transfer by radiation, it is not necessary to know the detailed dependence of emissivity on wavelength, but rather to replace $\varepsilon(\lambda,T)$ with an average value. In this case, one obtains:

$$Q = \sigma(\varepsilon_s T_s^4 - \varepsilon T^4)AF$$

Here, A represents heated substrate area and F is some configuration coefficient or "view factor". Again, Q represents the rate of heat transfer and σ is the Stefan-Boltzmann constant. Clearly, ε_s and ε are average emissivities of the illumination source and the substrate, respectively, and T_s and T are corresponding temperatures. Consequently, it is clear that the power density radiated by a hot object is proportional to the fourth power of its temperature. Therefore, the rate of radiative heat transfer is proportional to the difference of the temperature of the illumination source and the substrate each raised to the fourth power. In contrast, the rate of heat transfer due to either conduction or convection is proportional to the simple temperature difference, *i.e.*, heat transfer is a linear function of the temperature of the heat source. Physically, this implies that conduction and convection will be important at low temperatures, but radiation will dominate heat transfer at high temperatures.

A fundamental practical difficulty with RTA is the measurement and control of temperature. In the most primitive systems, this is done "open loop" by just setting the overall radiative power of the illumination source. However, this does not compensate for any differences in emissivity of the substrate, which often are quite significant.

Indeed, the same mechanism which causes transparent thin films to exhibit coloration corresponds to a fundamental change in effective emissivity of the substrate. Therefore, considerable effort has been expended to develop techniques for rapid and accurate temperature measurement. The most common devices for temperature measurement are optical pyrometers. In general, these determine the spectral distribution of radiation emitted by the substrate and determine the corresponding radiant or black body temperature. Pyrometers are attractive for this application because they do not require direct contact with the substrate and have relatively fast response times. However, they are subject to errors due to emissivity variations. These can be corrected by using two or more pyrometers with different spectral responses or by careful calibration using thermocouples and an appropriate test substrate. In the first case, complicated algorithms are used to determine the correction. In the second case, a calibration, which corresponds closely to the actual substrate, is determined. Other non-pyrometric techniques, which rely on mechanical or acoustical techniques have been suggested. However, pyrometry remains the dominant method of temperature measurement and control in RTA. Other issues associated with RTA are uniformity of heating and thermoplastic stresses. Obviously, non-uniformities in heating result in temperature non-uniformity of the substrate and can arise due to non-uniform illumination and/or losses at edges or contact points of the substrate. These problems are generally correctable by equipment design such as well-designed shielding or the use of susceptors made out of a highly conductive material, *e.g.*, silicon carbide or graphite.

Appendix A: The Thermodynamics of Oxygen in Silicon

The formation of oxide precipitates within a silicon crystal lattice has been treated in general; however it is instructive to consider actual quantitative data. For convenience, an oxide precipitate is assumed to be a small, spherical particle of SiO_2 embedded within an otherwise perfect silicon crystal. Of course, the standard formation reaction for silicon dioxide is as follows:

$$\mathrm{Si}(s) + \mathrm{O}_2(g) \rightarrow \mathrm{SiO}_2$$

As a matter of chemistry, a standard formation reaction defines a process for which one mole of some material (in this case, SiO_2) is formed from corresponding elements in standard state. (Here, this is crystalline silicon and oxygen gas.) Extensive tabulations of thermodynamic data for formation reactions have been compiled and are summarized for silicon, oxygen, and silicon dioxide in the following table:

Thermodynamic Potentials of Silicon, Oxygen, and Silicon Dioxide:

	ΔH_f°*	ΔG_f°*	S°**
Si(*s*)	0	0	18.81 *4.50*
O_2(*g*)	0	0	205.152 *49.03*
SiO_2(*quartz*)	-910.7 *-217.7*	-856.3 *-204.7*	41.46 *9.909*
SiO_2(*cristobalite*)†	-909.5 *-217.37*	-855.5 *-204.46*	42.68 *10.20*
SiO_2(*tridymite*)†	-909.1 *-217.27*	-855.3 *-204.42*	43.5 *10.4*
SiO_2(*quartz glass*)†	-903.5 *-215.94*	-850.7 *-203.33*	46.9 *11.2*

* kJ/mole (italics: kcal/mole); ** J/mole°K (italics: cal/mole°K)

† data taken from Handbook of Chemistry and Physics-1st Student Ed. (1988), all other data taken from CODATA Key Values for Thermodynamics

Of course, standard conditions are defined as 298.15°K and an ambient pressure of one atmosphere. Clearly, the thermodynamic potentials for all forms of silicon dioxide (quartz, cristobalite, tridymite, and glass) are quite similar.

However, within the silicon lattice, oxygen is not in gaseous diatomic form. Therefore, to be applicable to oxide precipitation, the formation reaction must be modified as follows:

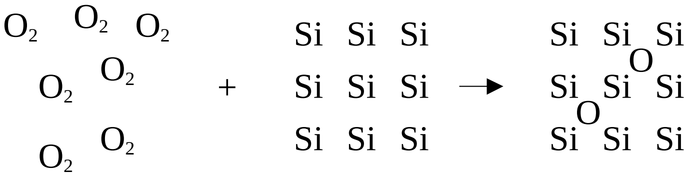
$$\mathrm{Si}(s) + 2\mathrm{O}(int) \rightarrow \mathrm{SiO}_2$$

Here, O(*int*) denotes oxygen atoms occupying interstitial sites within the silicon crystal lattice. Clearly, this reaction and the standard formation reaction are related by a third reaction that represents dissolution of oxygen gas in the silicon lattice and which can be formally written as follows:

$$O_2(g) \rightarrow 2O(int)$$

If the standard free energy of this reaction can be found, then the standard free energy of the previous reaction is easily determined.

For this purpose, it is useful to consider the dissolution of oxygen in solid silicon as a microscopic process. Obviously, oxygen molecules must react with the silicon lattice to form oxygen interstitials. This is represented schematically below:



Of course, this is just an alternative representation of the dissolution reaction appearing above, however, the silicon lattice is included explicitly. Clearly, the overall enthalpy change for this process must include contributions from strain energy associated with an oxygen interstitial, binding energy of an oxygen molecule, and binding energy of an oxygen atom within the silicon crystal. One expects the first two of these contributions to be positive and the last one to be negative. However, with the exception of the binding energy of molecular oxygen, these contributions are not readily determined. In contrast, the entropy change can be represented as the difference of the configurational entropy change due to random distribution of oxygen atoms in interstitial sites, $\Delta S_{\mathrm{O}}^{C}$, and the standard entropy of oxygen gas, $S_{\mathrm{O}_2}(T)$:

$$\Delta S = \Delta S_{\mathrm{O}}^{C} - \frac{N_{\mathrm{O}}}{2N_A} S_{\mathrm{O}_2}(T)$$

Here, N_{O} is the number of oxygen interstitials and N_A is Avogadro's number. Of course, a standard entropy is also associated with the silicon lattice itself; however, if the lattice is not disrupted by oxygen interstitials, this entropy can be taken to be unchanged when oxygen interstitials are introduced into the lattice and therefore makes no contribution to ΔS. (One should note here that ΔS is defined as the entropy change associated with the formation of N_{O} oxygen interstitials.)

Naturally, if N is defined as the number of interstitial sites in the crystal, then the configurational entropy change is easily represented as a binomial coefficient:

$$\Delta S = k \ln\left(\frac{N!}{(N - N_{\mathrm{O}})!N_{\mathrm{O}}!} \right) - \frac{N_{\mathrm{O}}}{2N_A} S_{\mathrm{O}_2}(T)$$

As is usual, one applies Stirling's approximation to obtain:

$$\Delta S = k(N\ln N - (N - N_{\mathrm{O}})\ln(N - N_{\mathrm{O}}) - N_{\mathrm{O}}\ln N_{\mathrm{O}}) - \frac{N_{\mathrm{O}}}{2N_A} S_{\mathrm{O}_2}(T)$$

One formally adds and subtracts $N_{\mathrm{O}}\ln N$ to the quantity within the parenthesis, from which it follows that:

$$\Delta S = -k\left(N_{\mathrm{O}}\ln\frac{N_{\mathrm{O}}}{N} + (N - N_{\mathrm{O}})\ln\left(1 - \frac{N_{\mathrm{O}}}{N}\right)\right) - \frac{N_{\mathrm{O}}}{2N_A} S_{\mathrm{O}_2}(T)$$

Clearly, one expects that N will be much larger than N_{O}, hence the second logarithmic term can be ignored, thus:

$$\Delta S = -kN_{\mathrm{O}}\ln\frac{N_{\mathrm{O}}}{N} - \frac{N_{\mathrm{O}}}{2N_A} S_{\mathrm{O}_2}(T)$$

Conceptually, it is convenient to replace absolute numbers of oxygen interstitials and interstitial sites, N_{O} and N, by corresponding concentrations, C_{O} and C. Furthermore, ΔS can be recast as a molar quantity if one rescales the right hand side by the ratio, N_A/N_{O}. Thus, it follows that:

$$\Delta S = -kN_A\ln\frac{C_{\mathrm{O}}}{C} - \frac{S_{\mathrm{O}_2}(T)}{2}$$

Since there are five interstitial sites per diamond cubic unit cell, it follows that C is just $5/a^3$, such that a is just the lattice parameter for silicon. Therefore, one finds that C is approximately $3.123(10^{22})$ cm^{-3}. Furthermore, kN_A is the ordinary ideal gas constant, R, which has a nominal value of 8.31441 J/mole°K.

The standard entropy of oxygen gas at any temperature and one atmosphere pressure can be obtained from the standard entropy at 298°K by means of the integral formula:

$$S_{\mathrm{O}_2}(T) = \int_{298}^{T} \frac{C_p}{T}\,dT + S_{\mathrm{O}_2}(298°\mathrm{K})$$

Here, C_p is the molar heat capacity at a constant pressure of one atmosphere. If one assumes that oxygen is an ideal diatomic gas, then C_p has the value of $7R/2$, hence it follows that:

$$S_{\mathrm{O}_2}(T) = \frac{7R}{2}(\ln T - \ln 298) + S_{\mathrm{O}_2}(298°\mathrm{K})$$

Alternatively, $S_{O_2}(T)$ can be determined more accurately from published curve fits for the temperature dependence of constant pressure heat capacity, $C_p(T)$.

$$C_p(T) = a + bT + cT^2 + \frac{d}{T^2}$$

Here, *a*, *b*, *c*, and *d*, are empirical coefficients. Thus, one obtains:

$$S_{O_2}(T) = \int_{298}^{T} \left(\frac{a}{T} + b + cT + \frac{d}{T^3} \right) dT + S_{O_2}(298°\mathrm{K})$$

For convenience, an aggregate coefficient, *B*, can be defined in terms of *a*, *b*, *c*, *d*, and the standard entropy of oxygen gas:

$$B = a \ln 298 + 298b + (298)^2 \frac{c}{2} - \frac{d}{2(298)^2} - S_{O_2}(298°\mathrm{K})$$

Hence, $S_{O_2}(T)$ has the following form:

$$S_{O_2}(T) = a \ln T + bT + \frac{cT^2}{2} - \frac{d}{2T^2} - B$$

For oxygen gas, published values for *a*, *b*, *c*, and *d* are 34.602 J/mole°K, 1.0795(10^{-3}) J/mole°K^2, 0 J/mole°K^3, and −785377 J°K/mole, respectively. From these values, one finds *B* equal to −3.2777 J/mole°K.

Obviously, it follows from the fundamental definition of Gibbs free energy that for the oxygen dissolution reaction:

$$\Delta G(T) = \Delta H(T) + RT \ln \frac{C_O}{C} + \frac{TS_{O_2}(T)}{2}$$

Obviously, ΔH is unknown. However, an method for the determination of ΔH is afforded by the oxygen solubility equilibrium. Of course, ΔG vanishes if dissolved oxygen is in equilibrium with ambient oxygen gas. Therefore, it follows that:

$$\Delta H(T) = -RT \ln \frac{C_O^{sat}(T)}{C} - \frac{TS_{O_2}(T)}{2}$$

Here, $C_O^{sat}(T)$ is saturated interstitial oxygen concentration at an absolute temperature, *T*. As shown in the following figure, this quantity has been determined experimentally over the temperature range 1000-1300°C:

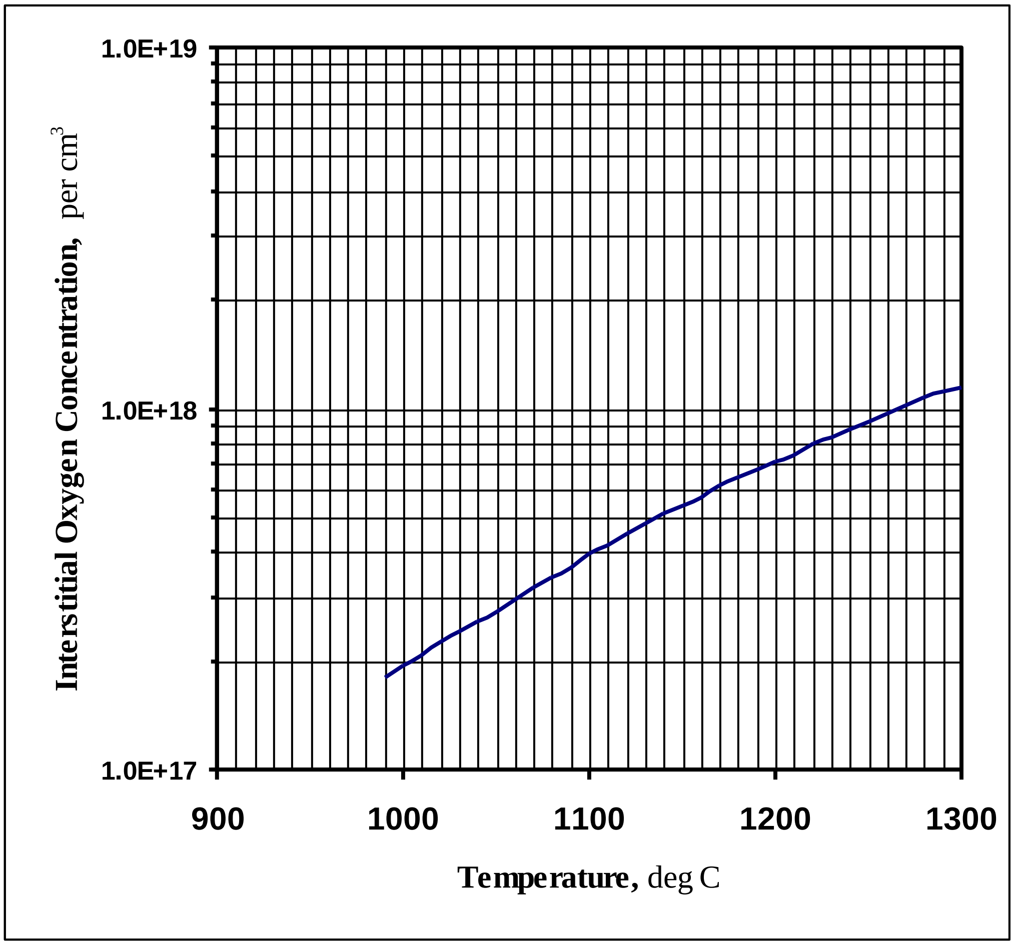


Fig. A1: Interstital oxygen solubility as a function of temperature

Clearly, the saturated oxygen interstitial concentration varies from 10^{17} to 10^{18} cm^{-3} over the given temperature range. This is consistent with typical oxygen concentrations observed in CZ silicon wafers. Furthermore, it seems clear from the trend, that the solid solubility of oxygen in silicon should further decrease when extrapolated to lower temperatures. This effect is likely the result of an increased strain energy contribution to enthalpy due to increased lattice rigidity at lower temperatures. Accordingly, if these concentrations are used to determine $\Delta H(T)$, one finds that resulting values are negative, but relatively small. Of course, negative values imply that energy is released when oxygen dissolves in a silicon crystal. This can be rationalized if one considers experimentally measured binding energies. In particular, Si-Si and O-O binding energies are observed to be 326.8 kJ/mole (78.1 kcal/mole) and 498.34 kJ/mole (119.106 kcal/mole), respectively. These can be compared to the Si-O binding energy, which is found to be 798.7 kJ/mole (190.9 kcal/mole). Clearly, the formation of Si-O bonds from Si-Si and O-O bonds is strongly exothermic. (This is also clear just from the large, negative formation enthalpy of silicon dioxide.) However, lattice strain largely offsets this so that the magnitude of the enthalpy of formation for oxygen interstitials is fairly small. Calculated values for the enthalpy of formation for oxygen interstitials is given in the following figure:

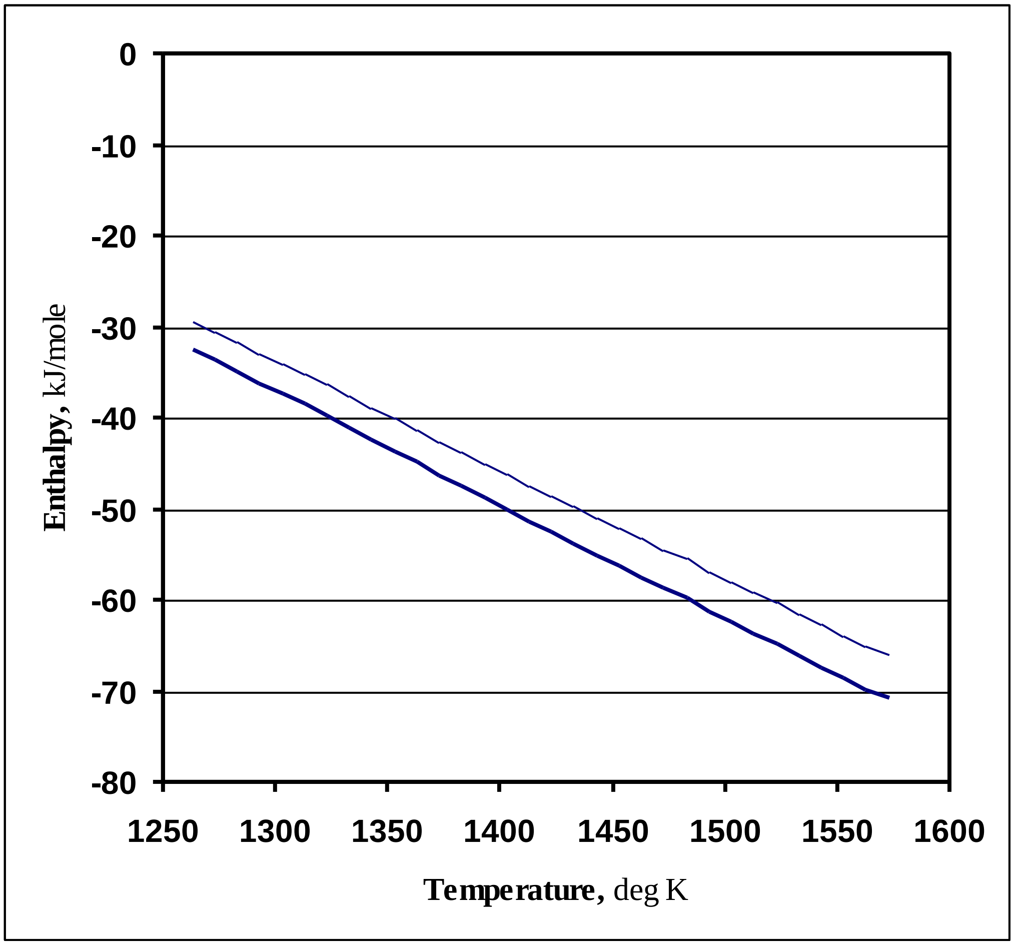


Fig. A2: Enthalpy of formation of interstitial oxygen as a function of temperature
(Heavy line: C_p obtained from empirical curve fit; Light line: C_p taken as $7R/2$)

Here, C_p has been estimated both from an empirical curve fit (heavy line) or as a constant, $7R/2$ (light line). The difference is found to be only about 3 kJ/mole and a simple linear fit is quite sufficient to describe the temperature dependence of both results. Hence, $\Delta H(T)$ can be represented by the empirical linear expression:

$$\Delta H(T) = c_p T + \Delta H_0$$

From the curve fit data, the parameters, c_p and ΔH_0 are found to be –0.12459 kJ/mole°K and 124.87 kJ/mole, respectively. Similarly, for C_p taken as $7R/2$, c_p and ΔH_0 are found to be –0.11873 kJ/mole°K and 120.52 kJ/mole, respectively. This expression may be substituted into the expression for ΔG to obtain the empirical result:

$$\Delta G(T) = \Delta H_0 + T\left(c_p + R\ln\frac{C_{\mathrm{O}}}{C} + \frac{S_{\mathrm{O}_2}(T)}{2} \right)$$

This is the Gibbs free energy of formation per mole of oxygen interstitials for an elemental silicon crystal having an oxygen interstitial concentration of C_{O}.

Conventional enthalpy of formation of SiO_2 as a function of temperature is readily obtained by integrating over the heat capacity for SiO_2 as follows:

$$\Delta H_{f\,\mathrm{SiO}_2}(T) = \int_{298}^{T} C_p(T)dT + \Delta H_{f\,\mathrm{SiO}_2}(298°\mathrm{K})$$

Again, as in the case of elemental oxygen, $C_p(T)$ for SiO_2 as is expressed as an empirical curve fit. This result is then combined with $\Delta H(T)$ for oxygen dissolution to obtain the quantity, $\Delta H_{\mathrm{SiO}_2}$, hence:

$$\Delta H_{\mathrm{SiO}_2}(T) = \Delta H_{f\,\mathrm{SiO}_2}(T) - 2(c_p T + \Delta H_0)$$

This is represented by the figure:

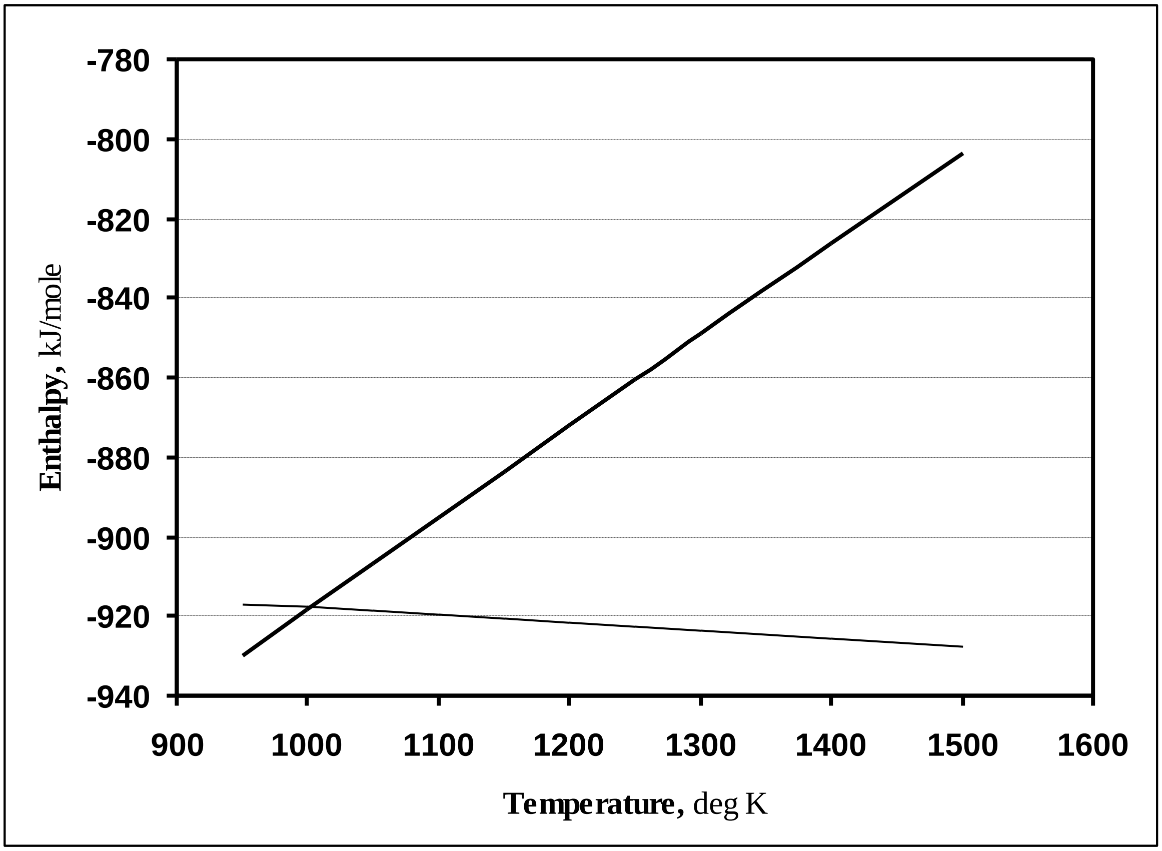


Fig. A3: Enthalpy of formation of bulk silicon dioxide in silicon as a function of temperature

Here, $\Delta H_{\mathrm{SiO_2}}$, defined as enthalpy of formation of bulk silicon dioxide from elemental silicon and oxygen interstitials (ignoring surface and strain energies associated with precipitates), corresponds to the heavy plot. This is contrasted with ordinary enthalpy of formation of SiO_2 (corresponding to the light plot). Clearly, over the temperature range 1000-1300°K, $\Delta H_{\mathrm{SiO_2}}$ varies only by about 120 kJ/mole.

Of course, entropies for silicon, oxygen, and silicon dioxide can be determined as a function of temperature in an entirely analogous fashion. These quantities can be combined with the enthalpy of formation obtained previously to obtain the conventional Gibbs free energy of formation of SiO_2 as a function of temperature. Naturally, the resulting Gibbs free energy of formation is then combined with $\Delta G(T)$ for oxygen dissolution to obtain the quantity, $\Delta G_{\mathrm{SiO_2}}$, hence:

$$\Delta G_{\mathrm{SiO_2}}(T) = \Delta G_{f\,\mathrm{SiO_2}}(T) - 2\Delta H_0 - T\left(2c_p + 2R\ln\frac{C_{\mathrm{O}}}{C} + S_{\mathrm{O_2}}(T)\right)$$

As before, this relation is represented figuratively as follows:

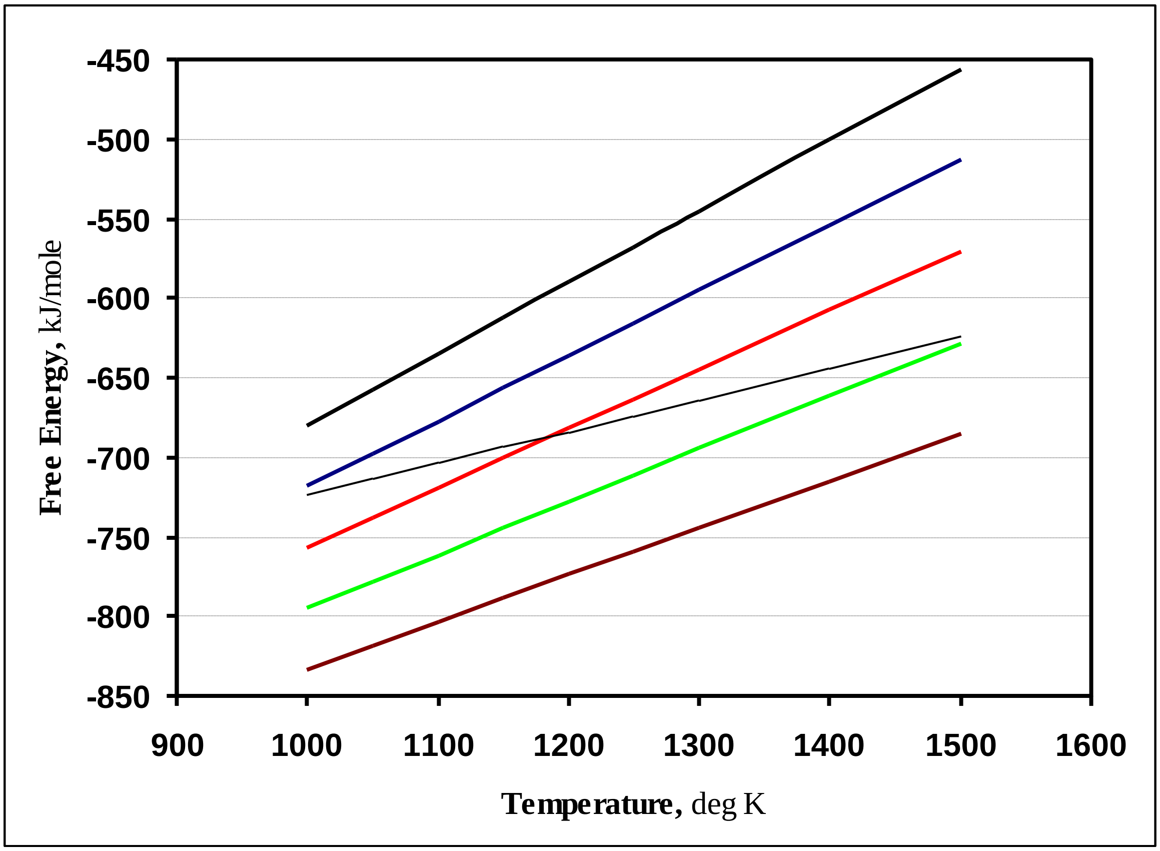


Fig. A4: Gibbs free energy of formation of bulk silicon dioxide in silicon as a function of temperature black: C_{O}=10^{14} cm^{-3}; blue: C_{O}=10^{15} cm^{-3}; red: C_{O}=10^{16} cm^{-3}; green: C_{O}=10^{17} cm^{-3}; brown: C_{O}=10^{18} cm^{-3}

Of course, $\Delta G_{\mathrm{SiO_2}}$ is Gibbs free energy of formation of bulk silicon dioxide from elemental silicon and oxygen interstitials (again, ignoring any precipitate surface and strain energies). Clearly, the Gibbs free energy has a much larger temperature variation than enthalpy. Furthermore, unlike enthalpy, Gibbs free energy is a function of interstitial oxygen concentration. This is illustrated by the heavy plots of various colors in the preceding figure. As one would expect, when interstitial oxygen concentration decreases, $\Delta G_{\mathrm{SiO_2}}$ becomes more positive (*i.e.*, oxide formation from interstitial oxygen is less favored.) As for enthalpy, the light plot corresponds to the conventional Gibbs free energy of formation of SiO_2. Clearly, if at some temperature the ordinary Gibbs free energy of formation and $\Delta G_{\mathrm{SiO_2}}$ are equal (*i.e.*, corresponding plots intersect), then the associated concentration of oxygen interstitials, C_{O}, can be identified with the solubility limit, *i.e.*, $\Delta G(T)$, as defined previously, exactly vanishes. Furthermore, since elemental oxygen gas is no longer a formal reactant for formation of bulk silicon dioxide from interstitial oxygen, all reactant and product phases can be considered condensed. Therefore, enthalpy, $\Delta H_{\mathrm{SiO_2}}$, is equivalent to internal energy, $\Delta E_{\mathrm{SiO_2}}$, and, likewise Gibbs free energy, $\Delta G_{\mathrm{SiO_2}}$, is equivalent to Helmholtz free energy, $\Delta A_{\mathrm{SiO_2}}$. Along with appropriate expressions for surface and strain energies, these quantities can be used to describe oxygen precipitation in silicon.

Appendix B: Transistors

Of course, the transistor is the most important semiconductor device and has enabled essentially all of modern solid-state electronics. However, as a matter of history, electronics began with vacuum tubes. As indicated previously, in 1880 Thomas Edison discovered that a two-element vacuum tube exhibits asymmetric conduction of current, *viz.*, the so-called “Edison effect”. Even so, it was not until the early twentieth century that the British physicist, John Ambrose Fleming, discovered that the Edison effect could be used to detect radio waves. Accordingly, Fleming developed and patented a two-element vacuum tube, which came to be known as the “diode”. Subsequently, three-element vacuum tubes or “triodes” were developed in the first decade of the twentieth century by Lee de Forest and others. Most importantly, triodes enabled development of the first true electronic amplifiers resulting in great improvement of telephony as well as radio transmitters and receivers. Physically, a vacuum tube operates by biasing two electrodes, conventionally called the “filament” and the “plate”. In operation, the filament is heated to a high temperature such that electrons can easily be thermionically emitted into the vacuum. Accordingly, if the plate is positive with respect to the filament, electrons are extracted and current flows. Conversely, if the plate is negative with respect to the filament, current does not flow. This is essentially the Edison effect. Moreover, as asserted above, a two-element tube is a diode; however, if a third electrode, called the “grid” (since it is usually a wire mesh or screen) is installed between the filament and plate, a bias voltage placed on the grid can modulate current flow. It is this “field effect” that allows a triode to operate as an electronic amplifier.

As early as 1925 it was recognized that field effect might also occur within crystalline solids. Indeed, the first patent for a *field effect transistor* (or FET) was filed in Canada by Austrian-Hungarian physicist, Julius Edgar Lilienfeld. Subsequently, in 1934 German physicist, Oskar Heil, also patented a field effect transistor of different design. Even so, no practical devices were ever built or tested. This is most likely a consequence of the lack of high purity semiconductor materials at that time. The development of high quality semiconductor crystals was motivated by the need during the Second World War for fast diodes, which were used in radars as a frequency mixer element in microwave receivers. Vacuum tubes were found to be too slow, but solid-state diodes fabricated from extremely pure germanium were found to be suitable. After the war, John Bardeen and Walter Brattain working under William Shockley at Bell Telephone Laboratories, succeeded in building the first operational “crystal triode”, *i.e.*, transistor, in 1947. However, this was not a FET, but rather a *point-contact transistor*. Although commercialized by the Western Electric Company, point-contact transistors were unfortunately found to be too fragile and were soon replaced by the *junction transistor* invented by Shockley in 1948. Subsequently, the *junction field effect transistor* or JFET was also successfully fabricated. (Actually, Bardeen, Brattain, Shockley, and others were trying to fabricate a JFET when the point-contact transistor was discovered since the JFET more closely resembles Lilienfeld’s original conception.)

Solid-state electronics was dominated by junction devices until the 1960’s, but these subsequently have been supplanted by the *metal-oxide-semiconductor field effect transistor* or MOSFET. The first practical MOSFET was invented and patented in 1959 by Dawon Kahng and Martin M. (John) Atalla, again, at Bell Telephone Laboratories.

As the name suggests, this device combines an MOS capacitor with a *pn*-junction. Accordingly, MOSFET's are both structurally and operationally different from junction transistors. Within this context, devices are made by fabricating an insulating layer on the surface of a semiconductor containing a *pn*-junction which defines the conductive "channel". As for a simple MOS capacitor, a gate electrode is formed on the top of the insulator. Typically, the semiconductor is crystalline silicon and the insulator is thermally oxidized silica. As noted previously, for this material combination the density of localized electron traps at the Si/SiO$_2$ interface can be quite low. Consequently, well-made silicon MOSFET's are inherently free from trapping and scattering of carriers.

Bipolar Transistors

Both point-contact and junction transistors are *bipolar*, which means that electrons and holes are majority carriers in different parts of the device; hence, current flowing through the device is carried by both electrons and holes. Physically, a bipolar junction transistor or BJT consists of two "back-to-back" *pn*-junctions that are in close proximity such that depletion regions interact. Accordingly, there are two kinds of bipolar transistors, *viz.*, *npn* and *pnp*. These designations specify the arrangement of the interacting junctions. A schematic representation of an *npn* BJT is shown in the following figure:

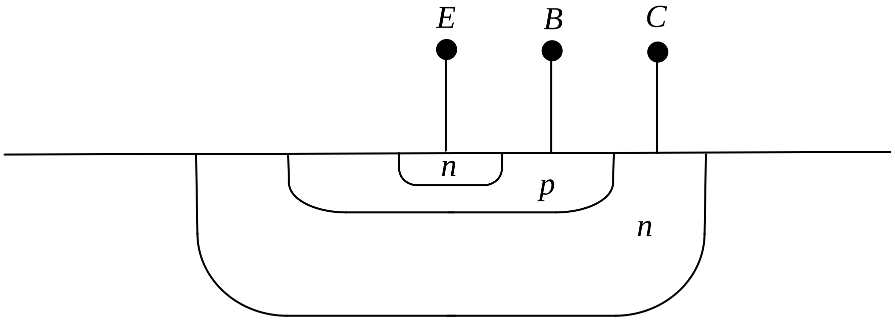


Fig. 71: Schematic of an *npn* BJT

Of course, a *pnp* is identical except that donor and acceptor doping is inverted. Here, *E, B,* and *C* denote *emitter, base,* and *collector* connections. (These designations descend from the original point-contact transistor for which in particular the "base" was a slab of semiconductor, *viz.*, germanium, to which "emitter" and "collector" point contacts were made.) Concomitantly, the two junctions are called "base-emitter" and "base-collector" junctions. In normal operation, for an *npn* transistor the base-emitter junction is forward biased (*i.e.*, base is positive with respect to the emitter) and the base-collector junction is reverse biased (*i.e.*, base is negative with respect to the emitter). A BJT is a current controlled device and in normal operation is governed by the equations:

$$I_E = I_0\left(e^{qV_{BE}/kT} - 1\right)$$

$$I_C = \alpha I_E = \left(\frac{\beta}{\beta+1}\right) I_E$$

$$I_B = (1-\alpha) I_E = \left(\frac{1}{\beta+1}\right) I_E$$

Here, α and β are identified as "common base" and "common emitter" current gains, respectively. Clearly, these two parameters are not independent and, as such, are determined by the physical characteristics of the transistor structure. The value of α is usually close to unity, *e.g.*, 0.980 to 0.998, which implies that β has a value between 49 and 499. Within this context, it is clear that I_E is determined by the diode equation for which V_{BE} is to be interpreted as the potential difference across the base-emitter junction. Typically, this is just the diffusion potential and, hence, is 0.5 to 0.7 volts in normal operation. (Obviously, I_0 remains defined as reverse saturation current just as in the case of a simple diode.) For completeness, if both junctions are forward biased, the transistor is said to be "saturated" and if both junctions are reversed biased the transistor is said to be "cut-off". In saturation, current flowing through the device is large and essentially independent of base current or V_{BE}. Conversely, in cut-off only very small leakage currents flow. It is possible to operate the device in inverted mode in which the base-collector junction is forward biased and the base-emitter junction is reverse biased. In principle if the device is exactly symmetric normal and inverted operation would have identical characteristics; however, as is evident from the figure, BJT's are generally asymmetric and inverted operation exhibits inferior characteristics.

Unipolar Transistors

In contrast to bipolar transistors all kinds of FET's are *unipolar*, which means that current is carried through the device either by electrons, *viz.*, *n*-channel, or holes, *viz.*, *p*-channel. In this regard, as asserted previously FET's resemble a vacuum tube triode, which are also unipolar devices, *i.e.*, current is carried by electrons. A schematic of a *p*-channel JFET is shown below:

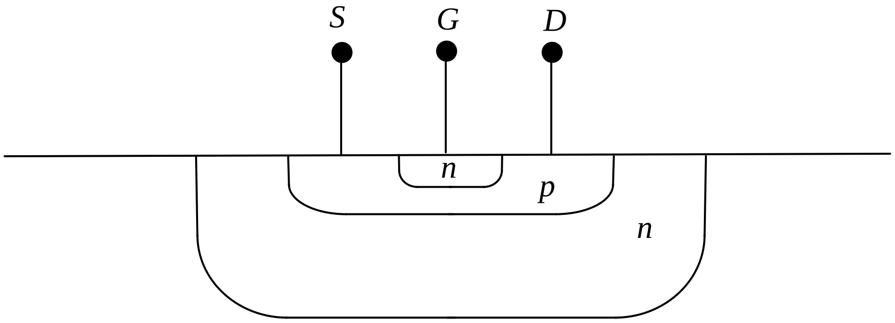


Fig. 72: Schematic of a *p*-channel JFET

Clearly, with the exception of external connections this figure is essentially identical to previous figure of an *npn* transistor and, thus, also consists of back-to-back junctions; however, operation is quite different. (In practice, the deep *n*-type region is reverse biased and, as such, serves to confine carriers, *viz.*, holes, to the channel.) Of course, in analogy to a BJT an *n*-channel JFET corresponds to inversion of donor and acceptor doping. As a matter of convention, *S*, *G*, and *D* denote *source*, *gate*, and *drain* connections. Concomitantly, if the gate is relatively unbiased and if the source is more positive than the drain, current can be expected to flow through the channel between source and drain. However, if the gate is biased positively with respect to the channel, *i.e.*, the gate-channel junction is reverse biased, then the depletion region expands and reduces effective conductivity of the channel. Indeed, if the channel is sufficiently thin, the depletion region can extend all the way across the channel, thus, effectively "pinching" it off so that very little current flows. In this case, the JFET is said to be "off". In general, this kind of operation is characteristic of a "depletion mode" FET.

Clearly, if the gate-channel junction is forward biased, an undesirably large current flows through the gate into the channel. Therefore, JFET's are generally not operated in this mode. However, if the gate is insulated as in a MOSFET, operation in "enhancement mode" is possible. Such a device is illustrated as follows:

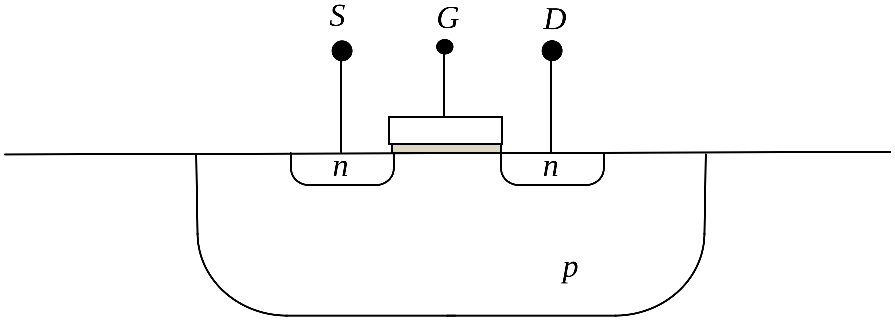


Fig. 73: Schematic of an enhancement mode *n*-channel MOSFET

Here, the light gray region is an insulator, typically formed of thermal oxide, although for very small devices other types of dielectrics may be used. Obviously, as asserted previously, a MOSFET is the combination of an MOS capacitor and *pn*-junctions. In any case, if the surface is accumulated or depleted under the MOS structure, very little current, *i.e.*, only leakage current, can flow between the source and drain. However, as the surface becomes inverted the surface becomes *n*-type and substantial current can flow. Indeed, in strong inversion the device becomes saturated in analogy to a BJT. Physically, source-drain current, I_D is determined by the bias voltage applied to the gate. Therefore, a FET is a voltage controlled device. In the case that the channel is accumulated or depleted, I_D is approximated by the expression:

$$I_D \cong I_{D0} e^{q(\mathrm{V}_{GS}-\mathrm{V}_t)/nkT}$$

Here, V_{GS} is the potential difference between gate and source connections and V_t is *threshold voltage*. In practice, V_t is determined by device structure and broadly corresponds to the potential associated with minimum capacitance of the MOS structure. Concomitantly, the ideality factor, n, corresponds to the formula:

$$n = 1 + \frac{C_s}{C_{ox}}$$

Of course, C_{ox} are C_s are oxide and depletion capacitances per unit area defined just as for a simple MOS capacitor. Naturally, the preceding expression is characteristic of cut-off and corresponds to $V_{GS} < V_t$. Accordingly, in normal operation, *i.e.*, $V_{GS} > V_t$ and $V_{DS} < (V_{GS} - V_t)$, I_D can be represented as follows:

$$I_D = \mu C_{ox} \frac{w}{l}\left((V_{GS} - V_t)V_{DS} - \frac{V_{DS}^2}{2} \right)$$

As might be expected, V_{DS} is the potential difference between drain and source connections, μ is effective carrier mobility, and w and l are surface width and length dimensions of the channel. Naturally, in saturation, *i.e.*, $V_{GS} > V_t$ and $V_{DS} \geq (V_{GS} - V_t)$, I_D becomes independent of V_{DS}, hence:

$$I_D = \mu C_{ox} \frac{w}{l}\left(\frac{(V_{GS} - V_t)^2}{2} \right)$$

This expression does not account for anomalies and is applicable only to ideal devices for which l is much greater than w.

For completeness, a depletion mode MOSFET is illustrated by the following figure, thus:

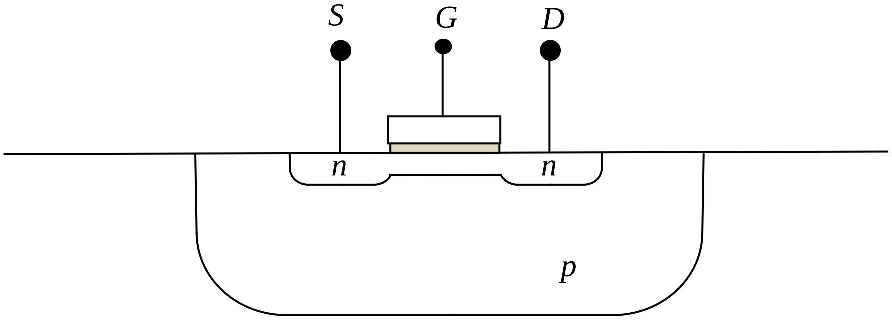


Fig. 74: Schematic of a depletion mode *n*-channel MOSFET

Clearly, operation of this device can be expected to be similar to an *n*-channel JFET, that is to say, if the gate is sufficiently negative with respect to the channel, the semiconductor surface is depleted and the channel becomes “pinched off”.

State-of-the-Art Devices

In conclusion, transistor structure has progressed far beyond the simple ideal structures described previously. The current state-of-the-art is embodied by devices with channel lengths of 32 to 22 nm or less. The former is still of planar design, but includes modifications such as a stressed channel and high-k gate dielectric. The reason for using stressed materials is to increase carrier mobility. Indeed, it has long been known that stress causes "splitting" of the band structure, which is characterized by sub-bands in which carriers have lower effective mass and, thus, higher mobility. As a practical matter, stress can be induced by epitaxial deposition of SiGe alloy on a pure silicon substrate. In this case, the SiGe alloy has a slightly larger lattice parameter than pure silicon, which results in biaxial stress in the deposited layer. In addition, stress may be modified and controlled by deposition of intrinsically stressed dielectric layers over the channel. As noted elsewhere, high-k gate dielectric is needed to reduce gate to channel leakage current which is inherent in ordinary thermal silica.

Reduction of channel length to 22 nm or less requires even more radical modification of transistor structure. In this case, a three dimensional "FinFET" structure has been introduced. Such a structure is characterized by a thin "blade" or "fin" of silicon, which is almost completely isolated from the substrate. The gate can then "wrap around" the fin on three sides, a so-called "tri-gate" structure, thus, resulting in more effective control of the channel.