Intrinsic Defects in Semiconductors

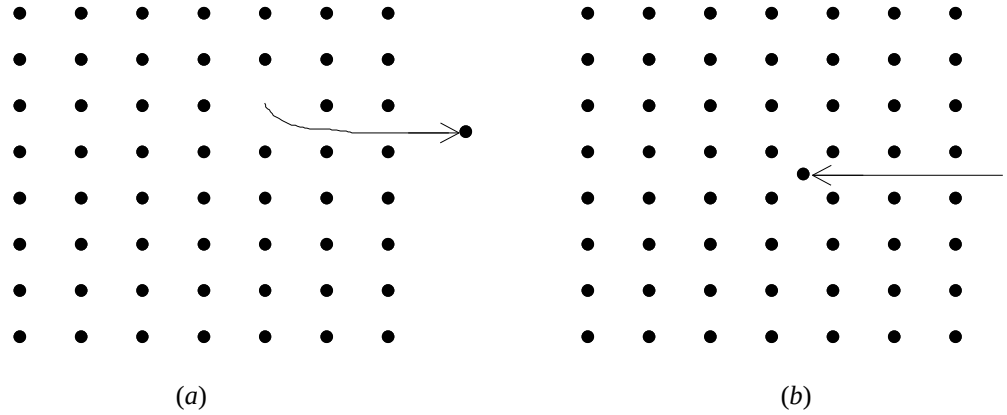
In all previous consideration of crystal structure and crystal growth, for simplicity it has been assumed that the silicon crystal lattice is entirely free of defects. Of course, in reality, this cannot be true since at any temperature greater than absolute zero, no crystal of finite size can be absolutely 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:



(*a*) (*b*)

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:

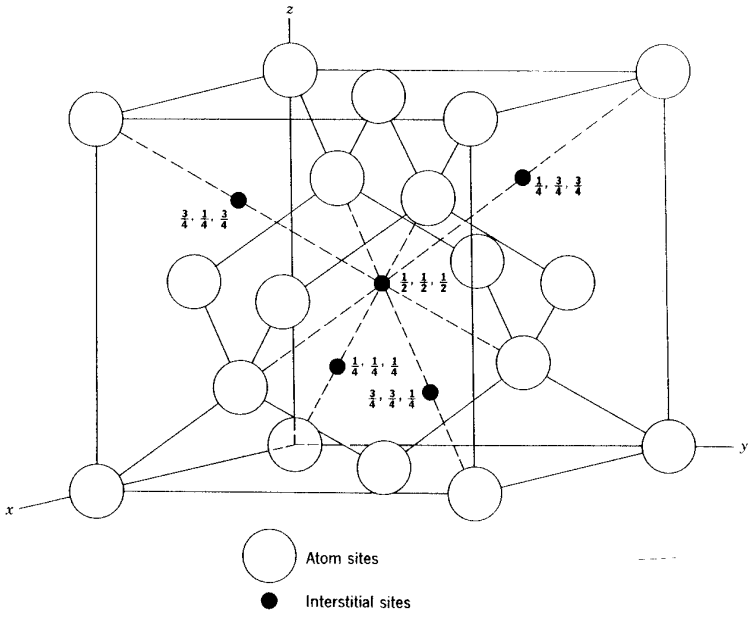


Fig. 19: Interstitial sites in the diamond cubic structure

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:

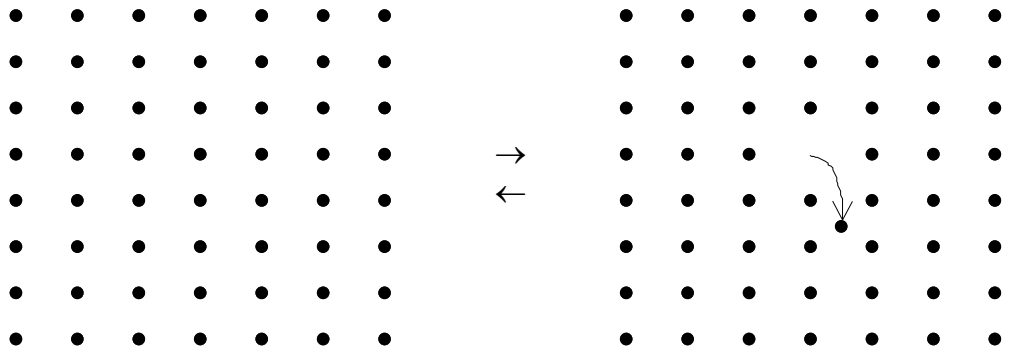


Fig. 20: Vacancy-interstitial "chemical" equilibrium

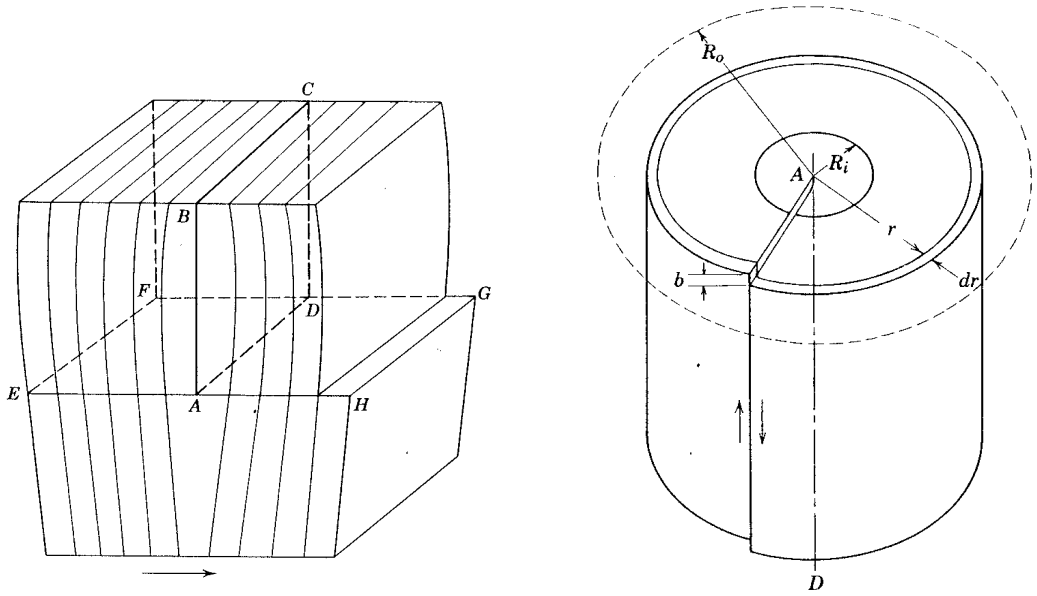
Clearly, the figure on the left represents an undefected lattice. The figure on the right represents "dissociation" of the lattice into a vacancy and an interstitial. It is further worthwhile to observe that vacancies and interstitials strongly interact and can also recombine, resulting in a significant 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:



(*a*) (*b*)

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.

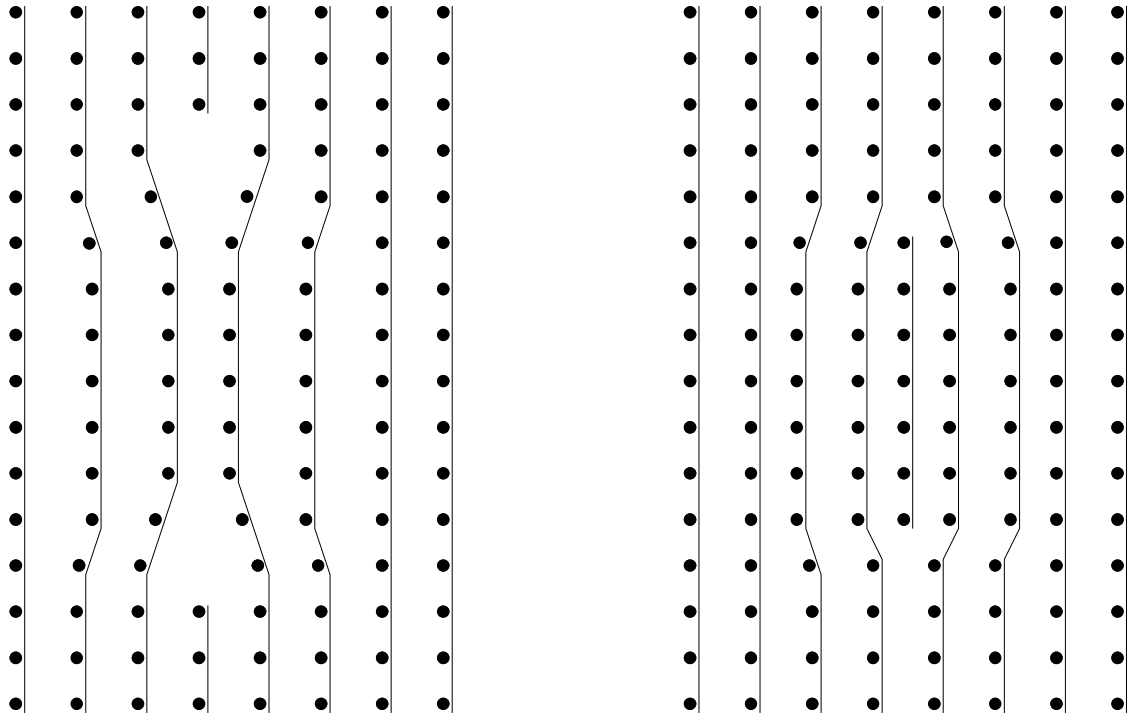
Line defects, *i.e.*, dislocations, can also interact with point defects naturally present within the lattice. This is most easily understood for the simple case of a pure edge dislocation. Suppose that due to the random thermal motion of the lattice, a vacancy migrates to the dislocation axis. Clearly, 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:



(*a*) (*b*)

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_{Mv}^{C}, "configurational" entropy and, ΔS_{Mv}^{X}, "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_{Mv}^{C}, as follows:

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

For simplicity, one can ignore excess entropy, ΔS_{Mv}^{X}, 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_{Mv}^{X} 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:

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

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, ΔE_{SiO_2} and ΔS_{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 stoichometric 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:

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

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:

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

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:

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

From this one immediately obtains:

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

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:

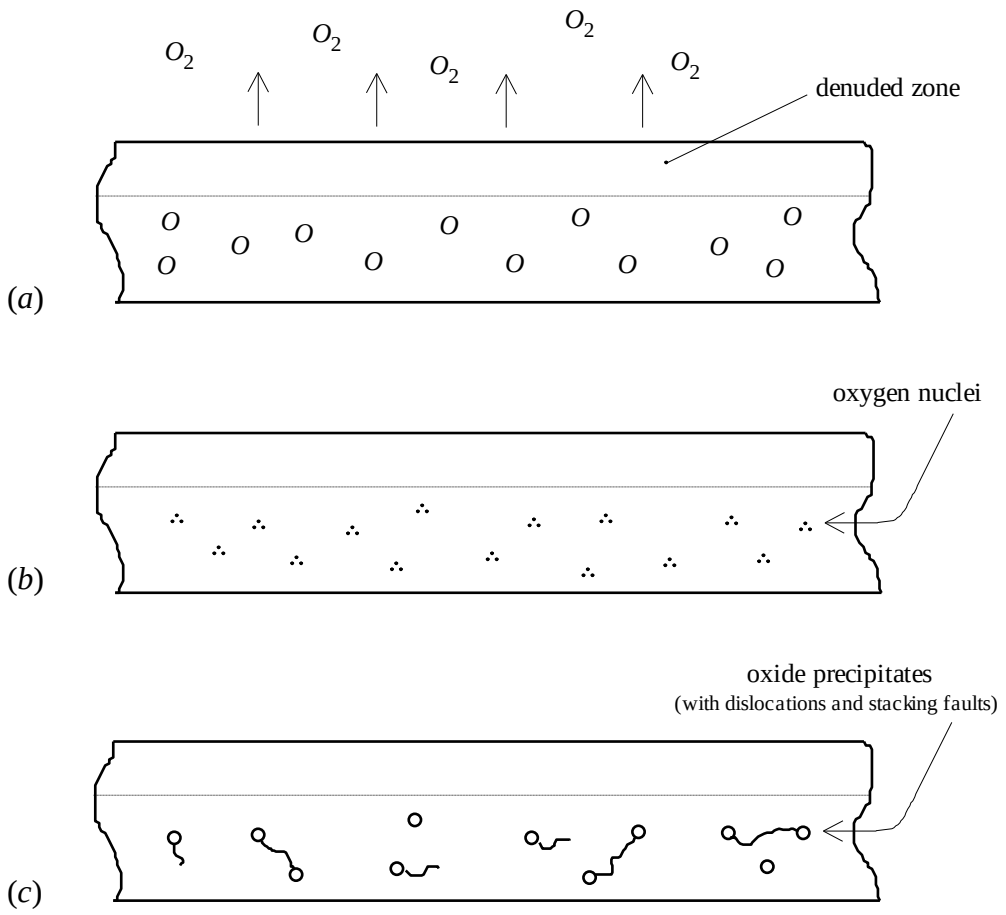


Fig. 23: Internal gettering process (*a*) denuding; (*b*) nucleation; (*c*) precipitate growth

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:

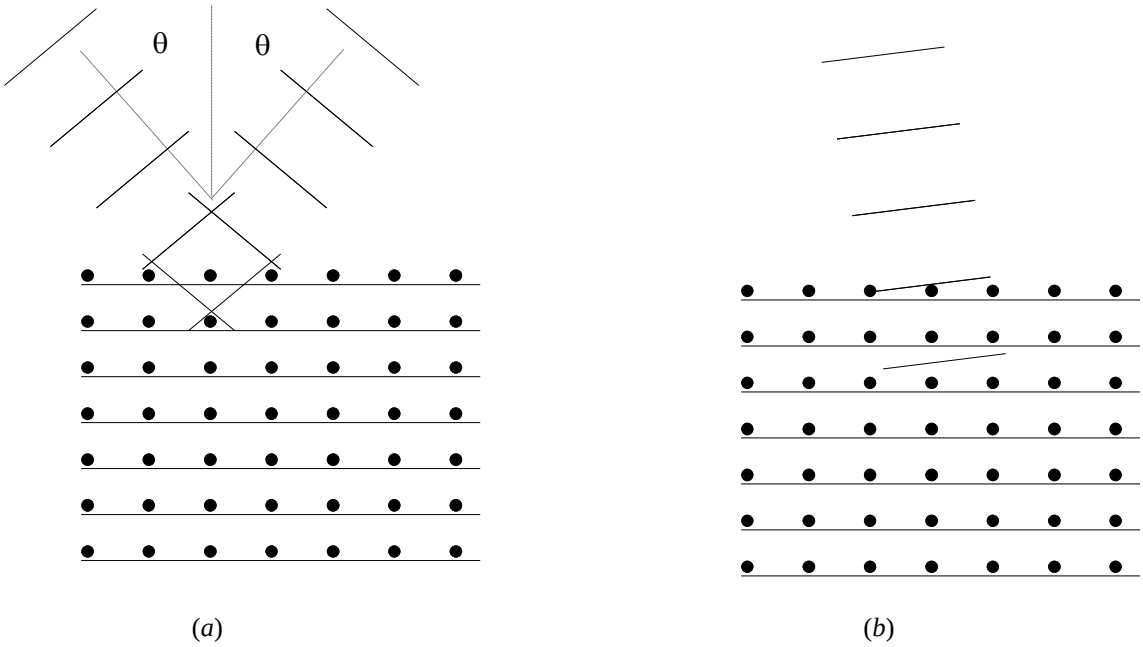


Fig. 24: X-ray diffraction (*a*) constructive interference; (*b*) destructive interference

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:

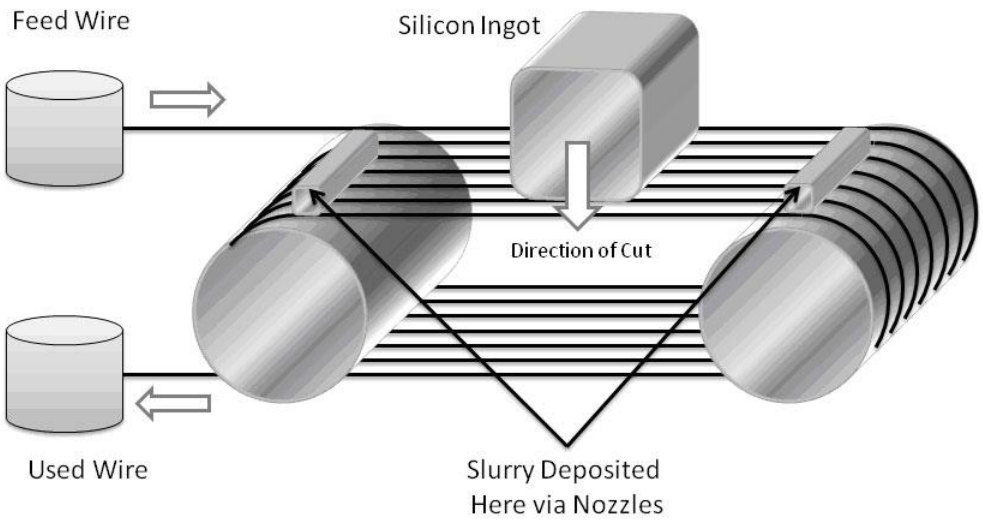


Fig. 25: Slicing of a silicon ingot (here shaped for solar cell fabrication)

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:

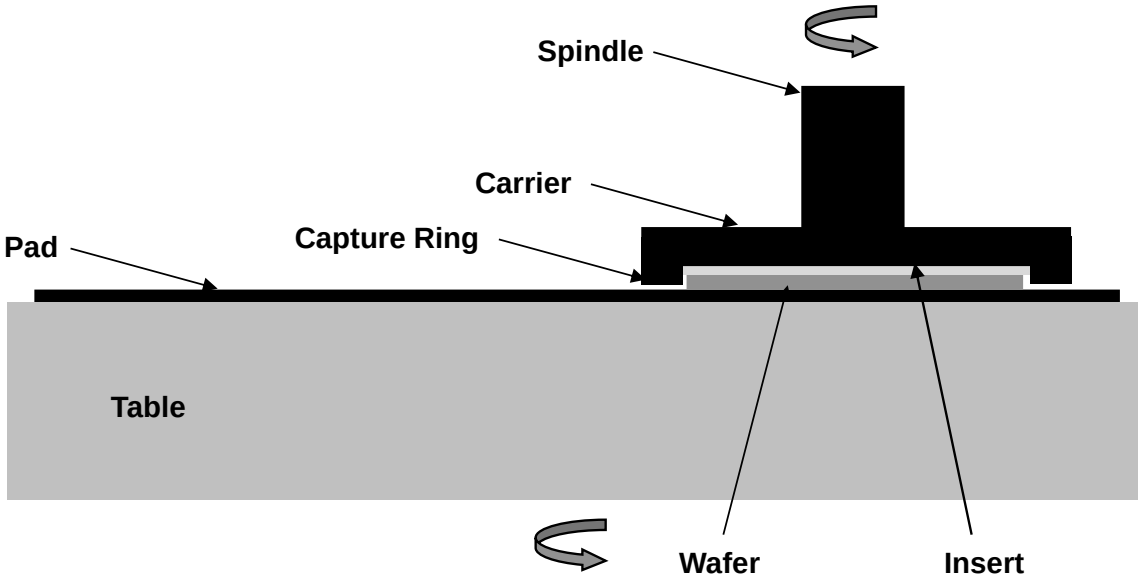


Fig. 26: Schematic of CMP machine

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:

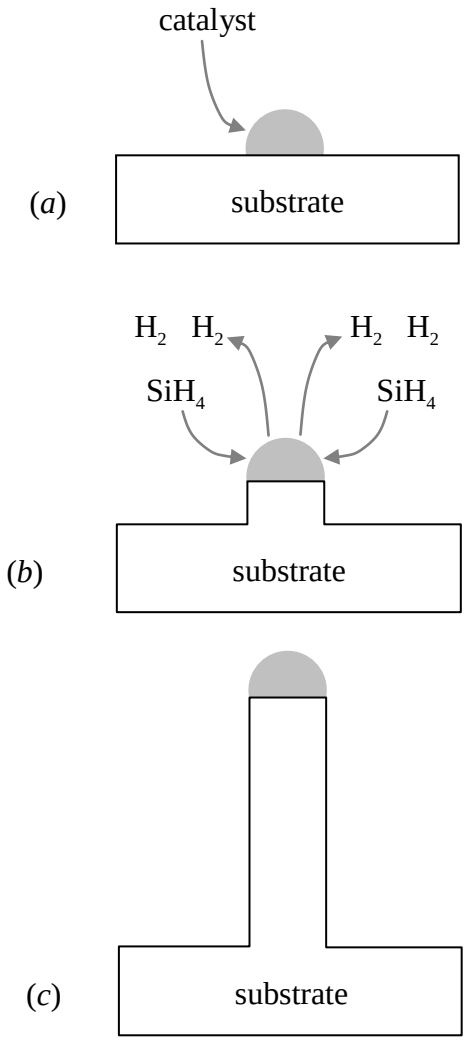


Fig. 27: Vapor-liquid-solid (VLS) process (*a*) catalyst on substrate; (*b*) growth; (*c*) finished nanowire

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:

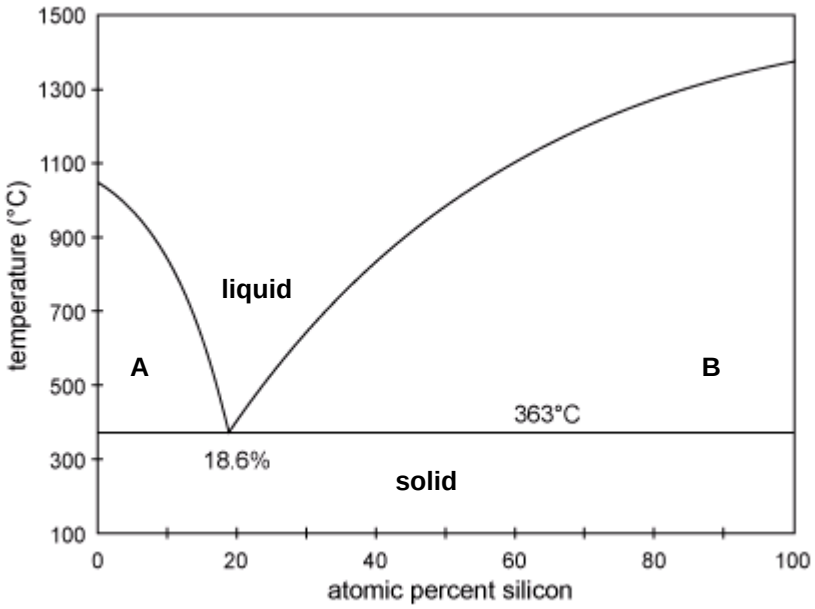


Fig. 28: Gold-silicon binary alloy phase diagram

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.