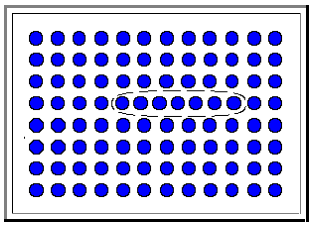
The Crowdion

Advanced

- A crowdion is a (postulated) special low temperature configuration of interstitials in **fcc** metals.
  - For making a crowdion, image the row of atoms along a densely packed **<111>** direction. Now take a number of atoms - say **10** - and "crowd in" one more.
  - A kind of elongated interstitial along a **<111>** direction is obtained - a crowdion.
- There was a big scientific controversy over the question if crowdions actually exist as a metastable interstitial configuration in the seventies. This controversy has never been finally resolved; however, most researchers in the field believe that it does not exist.
  - The advocates of crowdions came mainly from the Max-Planck-Institut für Metallforschung; Institut für Physik in Stuttgart; the opponents clusters around the research Center Jülich.
  - The following schematic drawing illustrates the crowdion configuration (for sake of clarity, the crowdion here is along **<100 >** in a cubic primitive crystal instead of **<111>**).



Open Question to Point Defects

Von G.F. Cerofolini (1989):

Having recently written a book on physical chemistry of, in and on silicon [1], I have considered some of the most obvious queries which could be raised when considering such a topic, viz.:

1. which is the atomic configuration of point defects? (e.g., is the self- interstitial quasi free or does it have a dumb-bell configuration?)
2. has each defect only one configuration or are several configurations possible?
3. which is the electronic structure?
4. which charge states are associated with each defect and where are they located in the gap?
5. can the defect be actually considered pointlike (i.e., do the remaining atoms remain on their lattice location) or does the deformation extend to long range?
6. does an entropic or enthalpic barrier exist for Frenkel-pair recombination?
7. which are the defect diffusivities in relation to their charge states?
8. is the surface an effective generation-recombination centre for point defects?
9. to which extent does this generation-recombination rate depend on surface conditions (free, oxidated, nitridated, etc.)?

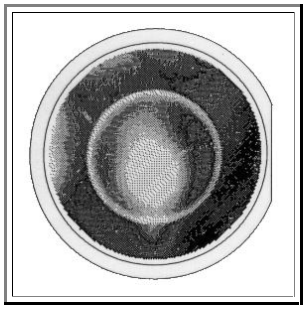
Most of them, however, remained unanswered.

Advanced

D-Defects Detected by ELYMAT Technique

Advanced

- With the ELYMAT (a special technique to map minority carrier lifetime in **Si**; see the publications in the link), **D**-defects and other microdefects in **Si** can be "seen" in some cases because they decrease the minority carrier life time (they act as recombination centers).
  - The pictures obtained monitor the local photo current (induced by a scanned Laser beam) in special electrolytic junctions. It is a direct measure of the minority carrier life time. A typical picture of state-of-the-art as-grown **150 mm Si** wafers from around **1990** is shown below. Bright areas correspond to decreased life times.



  - The most outstanding feature is the well-defined ring. It is due to small defects incorporating SiO_2.
  - With hindsight gained by much research in the nineties, the situation is as follows: Inside the oxygen-precipitate ring, small vacancy agglomerates (in the form of octahedral little voids) dominate; outside the ring, interstitials agglomerates (probably in the form of small stacking faults and dislocation loops (the old "classical" swirl defects)) were formed.
- This rather unique defect pattern is the result of the complicates interaction of three main point defects: vacancies, **Si**-interstitials and **O**-interstitials. Whereas the above interpretation is now universally accepted, the details about the primary defects are not yet known beyond reasonable doubt.
  - For a recent review read the paper of Bob **Falster** and V.V. **Voronkov**.

Lattice and Crystal

Basics

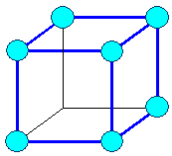
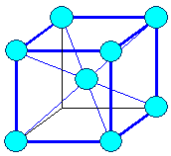
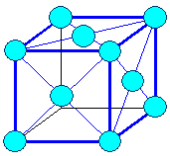
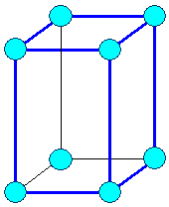
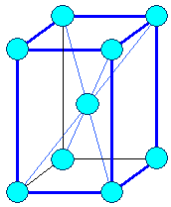
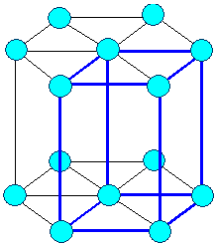
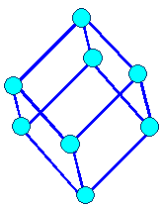
An ideal crystal is a repetition of identical structural units in three dimensional space. The periodicity is described by a mathematical **lattice** (which are mathematical points at specific coordinates in space), the identical structural units (or **base** of the crystal) are the atoms in some specific arrangement which are unambiguously placed at every lattice point. Note that *a lattice is not a crystal*, even so the two words are often used synonymously in colloquial language, especially in the case of elemental crystals where the base consists of one atom only.

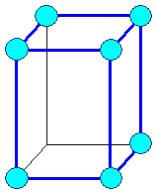
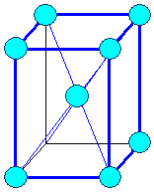
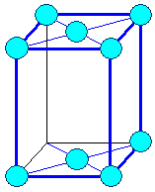
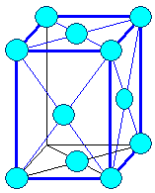
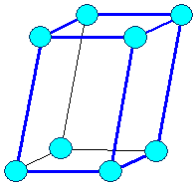
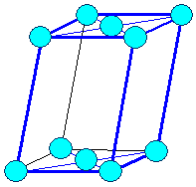
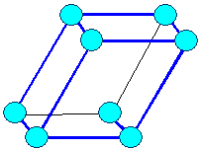
- All possible *lattices* can be described by a set of three linearly independent vectors $\underline{a}_1$, $\underline{a}_2$, and $\underline{a}_3$, the unit vectors of the lattice. Each lattice point than can be reached by a translation vector $\underline{T}$ of the lattice given by

$$\underline{T} = (u \cdot \underline{a}_1, v \cdot \underline{a}_1, w \cdot \underline{a}_3)$$

- With u, v, w = integers.

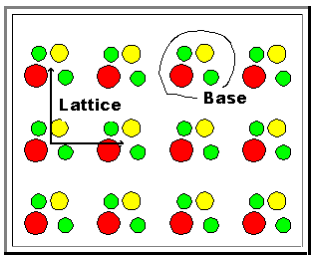
It is convenient, to classify lattices according to some basic symmetry groups. This yields the **14 Bravais lattices**, which are commonly used to describe lattice types. Their basic features are shown below (For sake of clarity, the lattice points are shown as little spheres and occasionally only "visible" lattice points are shown. These are *not* atoms, however!)

Name of crystal system Length of Base vectors	Angles between axes	Bravais Lattices		
Cubic $a_1 = a_2 = a_3$	$\alpha = \beta = \gamma = 90^0$	 **cubic primitive**	 **cubic body centered** (***bcc***)	 **cubic face centered** (***fcc***)
Tetragonal $a_1 = a_2 \neq a_3$	$\alpha = \beta = \gamma = 90^0$	 Tetragonal primitive	 Tetragonal body centered	
Hexagonal $a_1 = a_2 \neq a_3$	$\alpha = \beta = 90^0$, $\gamma = 120^0$	 Hexagonal (elementary cell continued to show hex. symmetry)		
Rhombohedral $a_1 = a_2 = a_3$	$\alpha = \beta = \gamma \neq 90^0$	 Rhombohedral		
Orthorhombic $a_1 \neq a_2 \neq a_3$	$\alpha = \beta = \gamma \neq 90^0$			

Orthorhombic $a_1 \neq a_2 \neq a_3$	$\alpha = \beta = \gamma \neq 90^0$	 Orthorhombic primitive	 Orthorhombic body centered	 Orthorhombic base face centered  Orthorhombic face centered
Monoclinic $a_1 \neq a_2 \neq a_3$	$\alpha = \beta = 90^0, \gamma \neq 90^0$	 Monocline primitive		 Monocline base face centered
Tricline $a_1 \neq a_2 \neq a_3$	$\alpha \neq \beta \neq \gamma \neq 90^0$	 Tricline		

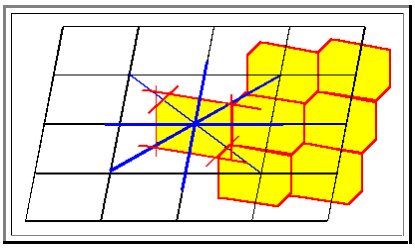
A *crystal* now is obtained by taking a **Bravais lattice** *and* adding a **base**!

- The base can just be one atom (as in the case of many elemental crystals, most noteworthy the metals), two identical atoms (e.g. **Si**, **Ge**, **C**(diamond)), two different atoms (**NaCl**, **GaAs**, ...) three atoms, ... up to huge complex molecules as in the case of protein crystals.
- An arbitrary example is shown below



For certain applications, a Bravais lattice may not be the best choice. Whereas, for example, it shows best the cubic symmetry of the cubic lattices, its **elementary cell** is not a **primitive unit cell** of the lattice, i.e. there are unit cells with a smaller volume (but without the cubic symmetry). For other cases (especially if working in reciprocal lattices) the choice of a **Wigner-Seitz** cell may be appropriate, which is obtained by intersecting all lines from one lattice point to neighboring points at half the distance with planes at right angles to the lines

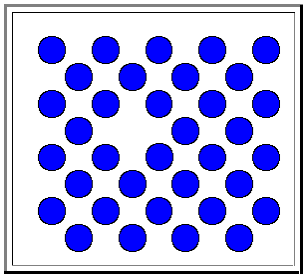
- This is shown schematically below: The blue lines connect lattice points, the red lines denote the intersection at right angles. The resulting Wigner-Seitz cell and its use in constructing the lattice are shown in yellow.



In practical work, one oftens refers to **crystal types** instead of lattices by using the name of prominent crystals, crystallographers or minerals etc.; e.g. "diamond type, Perovskites, "Zinkblende" structure and so on. A few examples are given in the link.

Vacancy

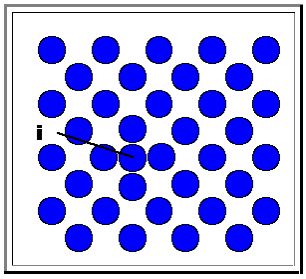
A vacancy is most simple crystal lattice defect imaginable. It is simply a missing atom in the base of a crystal; below a schematic drawing for an ***fcc*** crystal with one atom in the base.



Vacancies play an important role for the diffusion of atoms in crystals and thus for all of metal, semiconductor and ceramic technology. Atoms move by jumping into neighboring vacancies; this leads to a net flux of atoms and a (opposite) flux of vacancies.

Interstitials

- Interstitials are all atoms sitting not on their regular place, but between other atoms. The picture shows the simple case of a self interstitial atom in an elemental **fcc** crystal.



  - If the crystal is viewed as periodic arrangement of hard spheres, interstitials sit in the interstices between the spheres. For the most prominent simple crystals there are two kinds of interstices: Octahedra- and Tetrahedra interstices or gaps.

- There are two basic kinds of interstitials: Intrinsic and extrinsic interstitials:

  - **Intrinsic interstitials** are interstitials atoms of the same kind as the atoms of the crystal "**self-interstitials**"). They are practically non existent in elemental crystals (i.e. in all metals) with the big exception of **Si**, where intrinsic interstitials play an important role in diffusion and microdefect formation.
  - **Extrinsic interstitials** are interstitial atoms of a foreign (extrinsic) type, e.g. **C** in **Fe** or **O** in **Si**. They may diffuse directly through the lattice (i.e. without the help of vacancies) and play an important role in many technically relevant materials.

Thermal Equilibrium

- *Thermal equilibrium* is a central concept in thermodynamics. It describes the unique state of an ensemble of particles (i.e. the atoms of an crystal) that the system assumes by itself sooner or later (and later can mean really, really late) for a given set of intrinsic parameters (e.g., temperature, pressure, chemical potential) and extrinsic parameters (e.g., volume, entropy, number of particles).
  - The state of the system is unambiguously described by a state function which is called a **thermodynamic potential** and there are several thermodynamic potentials that can be used for a system description.
  - Whereas in principle *any* thermodynamic potential can be used for *any* situation (because they are related by a so-called **Legendre transformation**); it is useful to use specific thermodynamic potentials for specific systems.
- Depending on the kind of "contact" between the system under consideration and the environment (e.g. totally isolated, energy flow permitted, particle flow permitted, and so on), typical situations are:
- *Constant volume **V**, temperature **T**, and number of particles **N***.
  - The proper thermodynamic potential is the **free energy *F*(*V, T, N*)** (sometimes called **Helmholtz energy**).
- *Constant pressure **p**, constant temperature **T**, and constant particle (= atom) number **N***
  - This is the situation typical for a crystal. The appropriate thermodynamic potential is the **free enthalpy *G(p, T, N)*** (sometimes called **Gibbs energy**).
- The free enthalpy (defined as ***G = H – TS***) with ***H*** = enthalpy of the system and ***S*** = entropy is thus the most important thermodynamic potential when considering defects.
  - Thermal equilibrium for this case then simply means a state with an (absolute) minimum of the free enthalpy of the crystal.

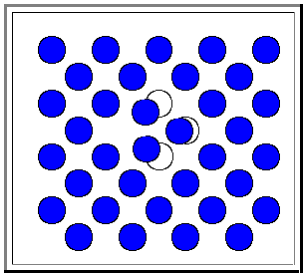
Band Gap

- This is just a quick reminder of the most fundamental electronic property of crystals, the *band gap* in the energy states of the electrons in a crystal.
  - If atoms condense into solids, the overlap of the energy states of the outermost electrons means that according to the **Pauli principle**, the energy levels must split into many level because no two electrons can be in the same (energy) state.
  - These levels may have energy gaps, i.e. within a band gap there are no allowed states for electrons. Bands need not be fully filled; the **Fermi energy** marks the energy state where just half of the available levels are occupied. At finite temperatures, the distribution function around the Fermi energy is "soft" and symmetric.
- Crystals can be classified according to their band structure. Materials without a bandgap or a very small bandgap are conductors, materials with a very wide bandgap are insulators.
  - Materials with a bandgap of about **0,5 eV - 2,5 eV** are semiconductors with especially remarkable electronic properties. This is due to the fact that there carrier densities can be influenced dramatically by introducing additional states in the bandgap via defects
  - This is usually done with substitutional impurity atoms (dopants), but crystal lattice defects in general also introduce states in the bandgap and thus influence the electronic properties of semiconductors.
- Semiconductors like **Si** or **GaAs** are the mainstay of modern technology only because it proved to be possible to control their crystal lattice defects to an unprecedented level of accuracy. But we should always bear in mind that among the huge number of semiconducting crystals most are technically useless because we cannot control their defects!
  - More about band gaps and semiconductors can be found in the *hyperscript* "Semiconductors".

Extended Vacancy

Illustration

- The question "How extended (how large) is a vacancy?", addresses the following concept:
  - The atoms next to a vacancy move to some extent into the free areas, their neighbors do the same. Eventually, the vacancy is kind of smeared out over a relatively large volume.
  - The vacancy would then appear like a little amorphous region or a little droplet of liquid material - we may have **26** atoms in a volume usually occupied by **27**.
- The vacancy then would be hard to notice on a schematic drawing, this is shown below for the case of four atoms removed, three put back in.

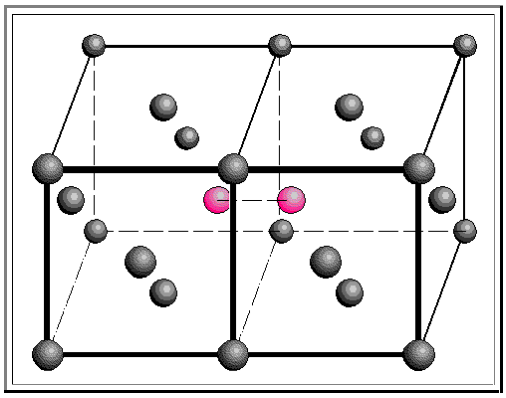


  - If the three blue atoms would assume the position shown and not the position of the empty circles, the vacancy would be hardly noticeable
- Does this happen? As a rule: No! - at least as far as we know.
  - Vacancies and interstitials are (for entropic reasons) sharply localized, and we know that from measurements. The reason is that extended vacancies change the vibrations frequencies of many atoms and thus add more entropy than necessary for thermal equilibrium.
  - There is, however, one exception to this rule - it is, like always **Si**. There is reason to believe that at high temperature both vacancies and interstitials are extended to some degree. But the last word on this is not yet in.

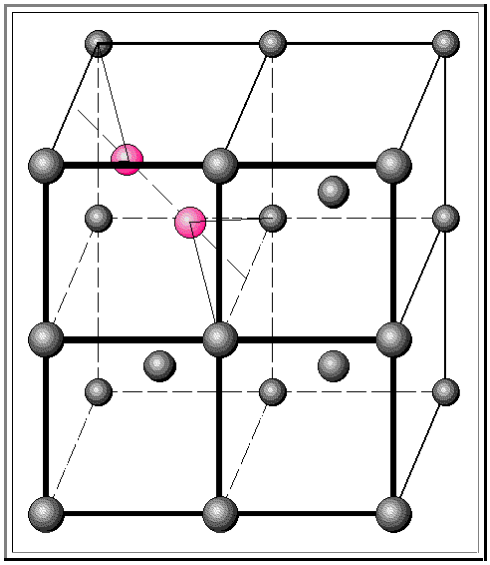
Dumbbells

Illustration

- A more symmetrical position for an interstitial atom may be obtained if two atoms share the space of one.
  - This is illustrated below for a **fcc** crystal; a configuration like this is called a dumbbell. In the dumbbell configuration, the interstitial now has a preferred direction; it is no longer isotropic.



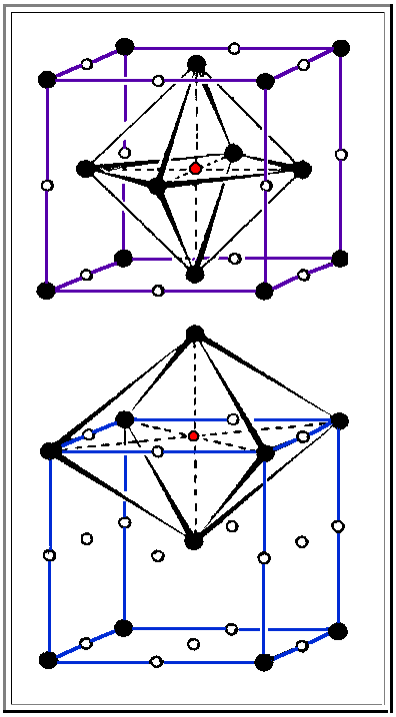
- In **bcc** crystals, a dumbbell configuration exists, too; it is shown below. Again, the interstitial now has a preferred direction.



Octahedral Sites

Illustration

- An octahedral position for an (interstitial) atom is the space in the interstices between **6** regular atoms that form an octahedra.
  - Four regular atoms are positioned in a plane, the other two are in a symmetrical position just above or below. All spheres can be considered to be hard and touching each other.
  - The six spheres define a regular octahedra, in its interior there is a defined space for an interstitial atom, bordered by six spheres.
- Octahedral sites exists in **fcc** and **bcc** crystals. The other prominent geometric environment for interstitials is the tetrahedral site.

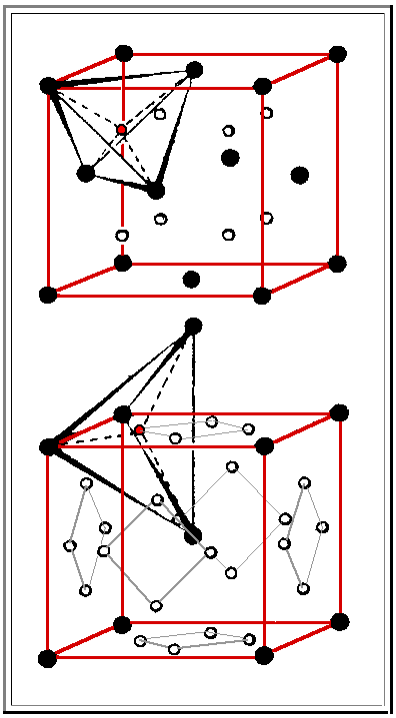


- This illustration shows the octahedral site in an **fcc** lattice bottom. We have **12/4 +1 = 4** positions per unit cell.
- Here we have octahedral sites in the **bcc** lattice. We have **12/4 + 6/2 = 6** positions per unit cell.

Tetrahedral Sites

- In a tetrahedral site the interstitial is in the center of a tetrahedra forms by four lattice atoms. Three atoms, touching each other, are in plane; the fourth atom sits in the symmetrical position on top.
  - Again, the tetrahedral site has a defined geometry and offers space for an interstitial atom.

Illustration

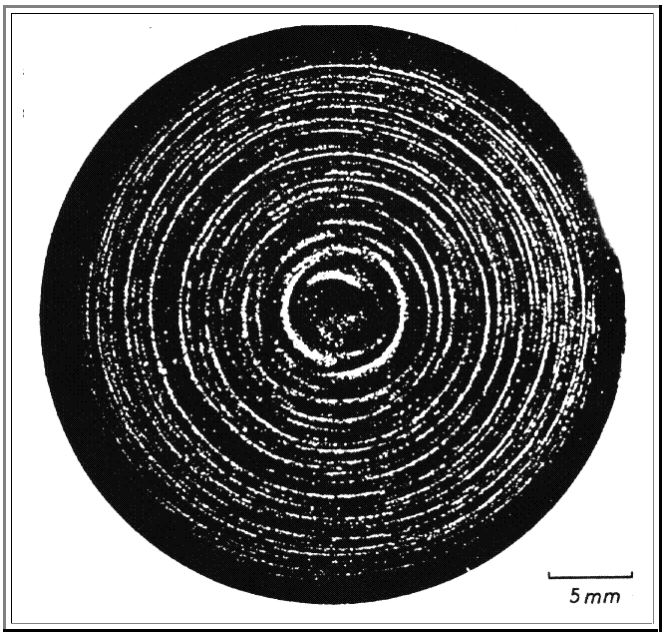


- The configuration on top is the tetrahedral position in the **fcc** lattice. The black circles denote lattice points, the red circle marks one of the **8** the tetrahedral position.
- The picture on the bottom shows the tetrahedral configuration for the **bcc** lattice. We have **(6 · 4)/2 = 12** positions per unit cell.

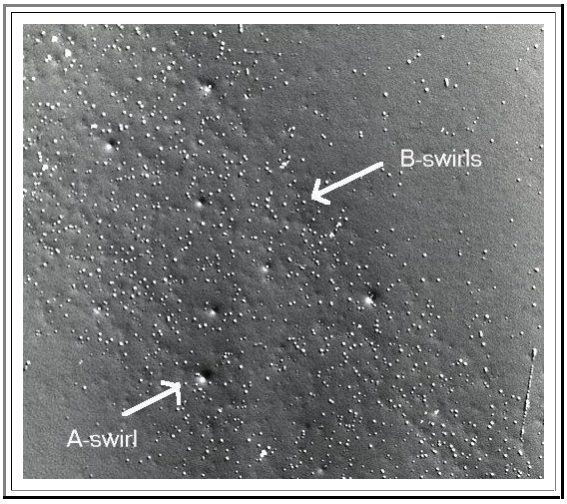
Swirl Defects

Illustration

- Swirl defects were discovered in the seventies in large dislocation-free Si crystals grown for micro electronic applications. They occur in two variants, the so-called **A**-swirl and **B**-swirl defects. The following picture shows a photography of a Si wafer that was preferentially etched to delineate the defects obtained by illuminating from the side (so that only light scattered at the defects enters the lens of the camera).
  - The typical spiral or swirl-like pattern explains the name of the defects.



  - Looking at the etch pattern with a microscope at high magnifications shows that there are a lot of small defects (the B-swirls; white dots) and a much smaller number of larger defects (**A**-swirls; the black-white contrasts). Quantitative evaluation of the micrograph shows that the **B**-swirls are delineated as small and shallow pits whereas the **A**-swirls are delineated as hillocks.



- Swirl defects are generated by the agglomeration of point defects while the crystals cools. The Si crystal growing industry soon learned how to grow crystals without swirl defects - without ever understanding precisely what they were.
  - But that did not mean that the crystals were defect free - it only meant that the methods employed then did not detect what was there. With new methods, defects reappeared, now called D-defects and bother the chip industry.
  - More about swirls can be found in a original research paper (in German) in the link.

Quantum Mechanical Concept of Entropy

Advanced

- As we have seen in the basic module, all kinds of definitions for the entropy are equivalent as long as some undetermined constant is allowed. Using quantum theory however, an absolute definition of the entropy, or an absolute zero point for entropy emerges.
- Without going into details, what happens is:
  - The "phase space volume" definition for ***P*** is the most general choice. Since we have to have a pure number in the **ln**, the volume that the system under consideration occupies in phase space must be divided by an appropriate elementary unit of phase space. In classical physics, there is no way of uniquely defining that unit; you are left with the ambiguity as discussed above.
  - Quantum theory, however, leaves only *one* choice for the elementary unit Π_0 of phase space volume for a system with ***N*** particles:

$$\Pi_0 = h^{3N}$$

  - With **h =** Plancks constant.
- Entropy, it turns out, is a well defined quantity after all. But again, for most applications, especially concerning defects, you do not have to worry about the finer points highlighted here.

Yakov Ilich Frenkel

Advanced

*Feb. 10th 1894 in Russia

† 1952 Soviet Union

In giving a short biography of Yakov Ilich Frenkel, one can't do better than **Serguey L. Lopatniko** from the Center for Composite Materials, University of Delaware, and **Alexander H.-D. Cheng** from the Department of Civil Engineering, University of Mississippi. What follows are excerpts from their article in the *Journal of Engineering Mechanics*, ASCE, 2005.

- The complete article may be found in this Link

If you ask a *physicist* from any country: “Have you heard about vacancies in crystals, quantum theory of conductivity, excitons, exchange interaction leading to spontaneous magnetization of the ferromagnetics and their domain structure?” He definitely will say: “Yes! These are basic physics. Everybody knows.”

- If you ask a *material scientist*: “Do you know that apparent stiffness of a metal is many orders lower than its theoretical value?” You will get the same answer: “Of course I know. It is a common knowledge.”
  If you ask a *chemist*: “What do you think about the definition of temperature for a single molecule?” He will answer: “Oh, it is one of the most important ideas in the theory of reaction of gases.”
  If you ask an *astronomer*: “Do you know that if a star has a mass slightly larger than our Sun, it can become unstable and collapse into a neutron star?” The astronomer will tell you: “Of course! It is basic astronomy.”
  If you ask a *geophysicist*: “Do you know that the Earth’s magnetic field is mostly generated by the movement of electrically conducting liquid in the melted part of Earth mantle?” He will definitely say: “Sure, we all know that as the Earth’s Dynamo.”

However, if you ask a *western* scientist: “Who introduced all these ideas in science?” You will get, perhaps, many great names such as Dirac, Heisenberg, Pauli, Chandrasekhar, Bullard, among others.

- It is improbable that somebody will give you the answer: one person introduced all these ideas—the brilliant Russian scientist *Yakov Il’ich Frenkel*.

Yakov Frenkel was born on *February 10, 1894* in the southern Russian city *Rostov-on-Don* (to Jewish parents). Since his early years he showed a talent in music, fine arts, and science. Being a student in the May Gymnasium at St. Petersburg, Ya. Frenkel wrote a 100-page mathematical paper, which was sent to Jacov Viktorovich Uspenskii, then a student of the famous Andrei Andreyevich Markov, for comment. Uspenskii found that the young Frenkel had rediscovered many results of the calculus of finite differences, which was not a part of his Gymnasium education.

.....

- Right after the Revolution, in 1918, Frenkel left St. Petersburg and took part in the organization of Tavrichesky University in Yalta, Crimea.

.....

- Living conditions in Crimea at that time were terrible. Excellent climate of Crimea could not compensate for the deprivation of war and hunger. Professor of the University was rationed 200 grams (less then ½ pound) of bread per day and a “free lunch”—one plate of “kasha.” Frenkel was jailed for two months during that time for political reasons. His younger brother Sergei, also a brilliant scientist, was drafted by the army and was killed in an accident. It was devastating for Frenkel and his family because of the five Frenkel siblings, only two remained alive.

.....

- In **1926** Frenkel introduced the key idea of defects of crystalline structure. He showed that the “evaporation” of atoms (or ions) from their equilibrium states occurred under finite temperature and introduced the idea of moving holes that could propagate though crystals independent of the movement of the atom that left it. These defects are known as **Frenkel defects**. On the base of this idea he calculated the electric conductivity of ion crystal and developed the theory of vibrational-translational movement of molecules in liquids and amorphous bodies, and particularly the theory of diffusion and viscosity of liquids and amorphous bodies.

.....

- In **1945**, the 220th anniversary of the Academy of Sciences was held, which was also an occasion to celebrate the returning to peace and the reconnection of the Russian scientific community with the international community.
  - Frenkel was able to meet a number of old friends, including F. Joliot-Curie, I. Langmuir, and M. Born. During the sessions Frenkel was honored, along with other scientists, with the Labor Red Banner Order. Two years later, his Kinetic Theory of Liquids was awarded the First Grade State Prize.
  - However, even at that time there existed the first hint of a change in the socialism policy; and the first gust of cold wind reached Frenkel soon after the anniversary. The ensuing political persecution affected not only Frenkel, but also many other prominent scientists. Frenkel's work was criticized for not contributing to the construction of the society of great socialism. His contributions in quantum mechanics and the theory of relativity were labeled as servility to Western science. His publications in the Western journals were unpatriotic. Several of his best books were published in German and English before they became available in Russian. To his accuser, these testified that Frenkel was in a hurry "to help the Americans use the achievements of Soviet Science in the interest of monopolistic capitalism."
  - He was even accused by his colleagues and the director of the Institute that his use of terms like "forced collectivization of electrons" and "collectivization under pressure" was a derision of soviet collective farms.
  - Frenkel's work was greatly affected and his health deteriorated toward the end of his life. He died in **1952**, not quite **58** years old.

- Use the Link and read the whole article! It's worth it - not just for learning about Frenkel's achievements, but also for learning soemthing about the not-so-remote past.

Walter Schottky

Advanced

*July 23rd 1886, in Zürich, Switzerland
†March 4th1976, in Pretzfeld, Germany

I apologize to whoever wrote this text (in German). I forgot where I found it. here is (an occasionally embellished) translation

Walter Schottky was a German physicist.

- After he finished his education, he taught Physics as a Professor at the University of Rostock (Germany) from **1923** to **1927**.
- After that he switched to Siemens & Halske, where he worked in Berlin and Pretzfeld (obscure town in Bavaria, where Siemens kept a Research center). He conducted basic research in semiconductor physics (better known by then as "dirty physics" and with no products to speak of) and Electronics (meaning whatever one did with vacuum tubes). The Schottky effect was named after him (meaning a special mode for electron emission from hot filaments), the Schottky diode, Schottky defects and the Schottky equation (also know as Schottky-Langmuir law of space charges)
- He conducted important research towards the "Schrot" effect (how does noise come about in electron currents?), space charge topics (not only in semiconductors, but also in vacuum tubes, etc.) and about blocking behavior of semiconductors (then still a kind of puzzle).
- **1915** he invented the tetrode (a special vacuum tube of large importance) and **1918** the "Superhet" principle for radio receivers (not much radio in **1918** yet!).

Equilibrium Concentration of the Vacancy-Impurity Atom Complex

Advanced

- We consider a complex of one vacancy and one foreign atom (a Johnson complex) in thermodynamic equilibrium.
  - We start with a *number* n_F of foreign atoms that is given by external circumstances. Some, but not necessarily all of these atoms will form a complex with a vacancy. The number of these complexes we call n_C.
- We can calculate the equilibrium number of Johnson complexes exactly analogous to the equilibrium number of vacancies by simply defining a formation enthalpy G_C and doing the counting of arrangements and minimization of the free enthalpy procedure.
  - This obviously will give us

$$n_C = (N_C - n_C) \cdot \exp - \frac{G_C}{kT}$$

  - With N_C = number of sites in the crystal where a vacancy could sit in order to form a Johnson complex.
  - We take $N_C - n_C$ in full generality because the places already occupied (= n_C) are no longer available, and we do *not* assume at this point that $n_C << N_C$ applies as in the case of vacancies.
- N_C, of course, is *not* the number of atoms of the crystal as in the case of vacancies, but *roughly* the number of foreign atoms - after all, only where we have a foreign atom, can we form a complex.
  - If we don't look at the situation *roughly* but in detail, we need to consider that there are as many possibilities to form a foreign atom - vacancy pair as there are nearest neighbors. We thus find

$$N_C = n_F \cdot z - n_C \cdot z$$

  - z is the coordination number of the lattice considered, i.e. the number of nearest neighbors. Again we do not neglect the places already taken, i.e. we subtract $n_C \cdot z$ from the total number of places.
- If we look at *concentrations*, we refer the numbers to the number of lattice atoms N which gives us for the concentration c_C of impurity atom - vacancy complexes

$$c_C = \frac{n_C}{N}$$

$$\frac{c_C}{c_F - c_C} = z \cdot \exp - \frac{G_C}{kT}$$

- We are essentially done. We have the concentration of Johnson complexes as a function of the concentration of the foreign atoms, the lattice type (defining z) and their formation enthalpy.
  - However, we would feel happier, if we could base the equation on material parameters which we already know - in particular on the equilibrium concentration of vacancies in the given material.
  - This needs a closer view on the formation enthalpy of the complex.
- As in the case of double vacancies, we may simply assume that there is a binding enthalpy between a vacancy and a foreign atom (otherwise there would be no driving force to form a complex in the first place).
  - We thus can write for G_C

$$G_C = G^V_F - (H_C - T \cdot \Delta S_C)$$

  - G^V_F is the free enthalpy of vacancy formation, H_C is the **binding enthalpy** of a Johnson complex, and $T \cdot \Delta S_C$ is the "**association entropy**" of the complex, accounting for the entropy change of the crystal upon the formation of a complex.
- Inserting this in the equation above gives for the concentration of Johnson complexes in terms of vacancy parameters and binding energies:

$$\frac{c_C}{c_F - c_C} = z \cdot \exp - \frac{G^V_F}{kT} \cdot \exp \frac{H_C}{kT} \cdot \exp \frac{\Delta S_C}{k}$$

$$c_C = (c_F - c_C) \cdot c_V \cdot z \cdot \exp \frac{H_C}{kT} \cdot \exp \frac{\Delta S_C}{k}$$

$$= c'_F \cdot c_V \cdot z \cdot \exp \frac{H_C}{kT} \cdot \exp \frac{\Delta S_C}{k}$$

- We used the familiar equation $c_V = \exp - (G^V_F / kT)$ to get this result.

We abbreviated the difference of the total concentration of foreign atoms and the concentration of Johnson complexes by c'_F; i.e. $c'_F = (c_F - c_C)$ because this allows a simple interpretation of the equation.

- The point now is to recognize that c'_F is nothing but the concentration of foreign atoms which are still available for a reaction with a vacancy, and that the last equation therefore is nothing but the mass action law written out for the reaction

$$1F + 1V \Leftrightarrow 1C$$

- With **F** = (available) foreign atom; **V** = vacany, and **C** = Johnson complex.

Looking closely (= thinking hard) you will notice that we now have a certain inconsistency in our book keeping:

- We always took into account that Johnson complexes already formed can *not* be neglected in counting possibilites, and we always corrected for that by using $c_F - c_C$ and so on - but we did *not* correct for the now more limited possibilities for positioning a single vacancy. We must ask ourselves if the presence of foreign atoms will change the equilibrium concentration of free vacancies.
- In other words, while we took the number of available positions for a vacancy in a complex to be $n_F \cdot z - n_C \cdot z$, we implicitly took the number of available positions for a free vacancy in the crystal to be simply $N =$ number of lattice atoms.
- Being more precise, we have to subtract $n_F \cdot z$ from N because $n_F \cdot z$ positions are, after all, *not* available for *free* vacancies. We thus have to replace N by $N' = N - n_F \cdot z$ when we consider the number of free vacancies.
- The concentration of the free vacancies thus becomes $c_V = (1 - z \cdot c_F) \cdot \exp - (G^V_F / kT)$, or $\exp - (G^V_F / kT) = c_V / (1 - z \cdot c_F)$
- Using this in the equation for the concentration yields

$$c_C = \frac{(c_F - c_C) \cdot c_V \cdot z}{(1 - z \cdot c_F)} \cdot \exp \frac{H_C}{kT} \cdot \exp \frac{\Delta S_C}{k}$$

$$\approx \frac{c_F \cdot c_V \cdot z}{(1 - z \cdot c_F)} \cdot \exp \frac{H_C}{kT} \cdot \exp \frac{\Delta S_C}{k}$$

- The last approximation is, of course, attainable if $c_C << c_F$, and that is the equation given in the backbone text.

A vague discomfort at the thought of the chemical potential is still characteristic of a physics education. This intellectual gap is due to the obscurity of the writings of J. Willard Gibbs who discovered and understood the matter 100 years ago.

C. Kittel; Preface to his book: Introduction to Solid State Physics

Advanced

The chemical potential

Names and Meanings

- This module is registered in the "advanced" part, despite the fact that the chemical potential belongs to basic thermodynamics. The reason is that people with a mostly *physical background* (like me) may often have learned exciting things like Bose-Einstein condensations and the Liouville theorem in their thermodynamics courses, but not overly much about chemical potentials and chemical equilibrium.
- First we will address, somewhat whimsically, a certain problem related to the name *"Chemical potential"*. It is, in the view of many (including professors and students), a slightly unfortunate name for the quantity $\partial G / \partial n_i$; meaning the partial derivative of the free enthalpy with respect to the particle sort ***i*** and all other variables kept constant (See a pure thermodynamic script as well).
  - In other words, the "chemical potential **μ**" is a measure of how much the free enthalpy (or the free energy) of a system *changes* (by dG_i) if you add or remove a number dn_i particles of the particle species **i** while keeping the number of the other particles (and the temperature ***T*** and the pressure ***p***) constant:

$$dG_i = \frac{\partial G}{\partial n_i} \cdot dn_i$$

  - Since particle numbers are pure numbers free of dimensions, the *unit of the chemical potential is that of an energy*, which justifies the name somewhat.
  - However, the particles considered in the context of general thermodynamics do not have to be only atoms or molecules (i.e. the objects of chemistry). They can be electrons, holes, or anything else that can be identified and numbered. In considering e.g., the equilibrium between electrons and holes in semiconductors, physically minded people do not feel that this involves chemistry. Moreover, they feel since electrons and holes are Fermions, classical thermodynamics as expressed in the chemical potential or the mass actions law, might not be the right way to go at it. The "chemical potential" of the electrons, however, is still a major parameter of the system (to the annoyance of the solid state physicists - they therefore usually call it "**Fermi energy**").
- A better name, perhaps, would help. How about "particle potential"? But such a name would not be too good either. Because now there is the danger of mixing-up the *thermodynamic Potential* ***G*** of the particles, and the "*Particle Potential*", which is a partial derivative of ***G*** – not to mention the common electrostatic or gravitational potential. Now, what exactly is a potential? Use the link to refresh your memory!
  - The Gibbs energy ***G***, e.g., may be viewed as a thermodynamic potential because it really is a "true" potential. Not only does it satisfy the basic conditions that its value is independent of the integration path (i.e. it does not matter how you got there), but it is also measured in units of energy and its minima (i.e. $dG = 0$) denote stable (or metastable) equilibrium.
  - The chemical potential meets the first two criteria, albeit the second one only barely. This is so because if you define it relative to the particle *concentration* and not the number (which would be equally valid), you end up with an energy *density* and not an energy.
  - The last condition, however, *is not true* for the chemical potential. Its minima do not necessarily signify equilibrium; the equilibrium conditions if several particles are involved are rather

$$\Sigma_i \mu_i = 0$$

  - Belowis a detailed derivation for this.

- Lets try a different approach. In a formal way, the particle numbers are *general coordinates* of the free enthalpy for the system under consideration. Since the partial derivatives of thermodynamic potentials with respect to the generalized coordinates can be viewed as **generalized forces** (in direct and meaningful analogy to the gravitational potential), the chemical potentials could just as well be seen as *chemical forces*.
  - The equilibrium conditions are then immediately clear: The sum of the forces must be zero. If there is only one particle in the system (e.g. vacancies in a crystal), equilibrium exists if there is no "chemical force", i.e. $\mu_{vac}=\partial G/\partial n_V=0$. If there are more particles that are coupled by some reaction equation, the left-hand sum of the chemical potentials (times the number of particles involved) must be equal to the right hand sum. An example:

Reaction

$$SiO_2 + 2CO \Leftrightarrow Si + 2CO_2$$

Equilibrium condition:

$$\mu_{SiO_2} + 2\mu_{CO} = \mu_{Si} + 2\mu_{CO_2}$$

  - Think of a beam balance and you get the drift.
- This suggests yet another name: "**Particle force**" or "**Particle change force**". Of course, now we would have a force being measured in terms of energy - not too nice either, but maybe something has to give?
  - Unfortunately, there is another drawback. If we look at currents (electrical or otherwise), i.e. at *non-equilibrium* conditions, the driving forces for currents very generally can be identified with the gradients of the chemical potentials (which still may be defined even under *global* non-equilibrium as long as we have *local* equilibrium). Now we would have a force being the derivative of a force - and that is not too clear either. In this context a potential would be a much better name.
- So - *forget it*! $\partial G/\partial n_i$ is called, and will be called "*chemical potential of the particle sort i*". But by now, you know what it means. Still, if you feel uncomfortable with the name "Chemical Potential" in the context of looking at non-chemical stuff, e.g. the behavior of electrons, use your own name while thinking about it, keep in mind what it means, but do write down "*chemical potential*".

The Burden of History: Gases and Fugacity

- The good part about the chemical potential is its simplicity - after you have dug through the usual thermodynamical calculations. It is especially easy to obtain for (ideal) gases.
  - An ideal gas is a system of particles *of any kind whatsoever* that obeys the equation $p\cdot V = N\cdot R\cdot T$ with N = Number of mols in the system; or $p\cdot V = n\cdot k\cdot T$ with n= Number of particles in the system.
  - Lets go through this quickly (haha), because we are not really interested in gases, but only want to remember the nomenclature and the way to go at it.
- From regular thermodynamics we get a lot of relations between the partial derivatives of state functions and therefore also for the chemical potential, e.g.

$$\frac{\partial \mu_i}{\partial p} = V_i$$

$$\frac{\partial \mu_i}{\partial T} = -S_i$$

  - *with the proper quantities kept constant and with care as to the use of absolute or molar values*
  - From these equations we obtain for the chemical potential of a pure ideal gas, i.e. a system consisting only of one kind of component - a bunch of O_2 molecules in a container, or a *bunch of vacancies in a crystal*:

$$\mu_{\text{ideal gas}}(p,T) = \mu^0_{\text{ideal gas}} + RT\cdot \ln \frac{p}{p^0}$$

- *Now wait a minute!* In the case of vacancies, we seem to have *two* components - the vacancies and the crystal, not to mention that considering vacancies as an ideal gas seems to be stretching the concept a bit.
  - Well - yes, there is the crystal, but for the real gas there is the vacuum in which the particles move. As long as the "container" of the ideal gas particles does not do anything, we may ignore it (if we don't, math will do it for us as as soon as we write down equations like the mass action law or others that tell us what happen inside the "container").
  - *So get used to the idea of treating point defects like an ideal gas for a start!*
- What is $\mu^0_{\text{ideal gas}}$? It is called something like "*the standard chemical potential for the pure phase*". Lets look at what it means from *two* points of view.
  - *First*, if we stay with the vacancy example, i.e. we consider an ideal gas of vacancies, the pressure is given by $pV=n\cdot kT$ with n=number of vacancies in the crystal, or $p=n\cdot kT/V$. Likewise, p^0, the pressure at some reference state, can be written as $p^0=NkT/V^0$ with N= number of vacancies at the reference state and V^0 volume of the system at the reference state.
  - Rewriting the chemical potential of our vacancies for n gives (in **3** easy steps)

$$\frac{p}{p^0} = \frac{n\cdot k\cdot T\cdot V^0}{N\cdot k\cdot T\cdot V} = \exp\frac{\mu_V - \mu_V^0}{RT}$$

  - Since the volume of the crystal will not change much no matter at what state you look, we have $(V^0/V)\approx 1$. Moreover, in equilibrium we demand $\mu_V=0$. This leaves us with

$$\frac{n}{N} = \exp -\frac{\mu_V^0}{RT}$$

  - And this looks very familiar! If we chose the standard state to be N= number of atoms of the crystal=number of sites for vacancies, μ_V^0 must be the energy of forming one mol of vacancies and that is simply the formation energy measured in **kJ/mol**. If you like electron volts, simply replace **R** by **k**.
  - In other words, the standard reference state is very important, but also a bit trivial. You can chose whatever you like, but there are *smart* choices and *not so smart choices*. Best to stick with the conventions - they usually are smart choices and you can use the numbers given in books and tables without conversion to some other system.
- Now the *second* point of view.
  - Since the chemical potential is an energy (with many properties very similar to the better known gravitational or electrostatic potential energy), there is no unique choice of its zero point. All hat counts are *changes*, i.e. μ_i(state x) – μ_i(state 0).
  - For μ_i**(state 0)** we write μ_i^0 and call it standard potential.
- So far so good. But what about the chemical potential of some stuff (always particles) in a *mixture* with other particles? To start easy, lets take a mixture of ideal gases - O_2 with N_2, vacancies and interstitials (both uncharged, so there is negligible interaction).
  - We want the chemical potential $\mu_i^{mix}(p,T)$ of the component i in a mixture of ideal gases as a function of the temperature and the (total) pressure. We first need the quantities "*mole fraction*" and "*partial pressure*" to describe a mixture.
  - The **mole fraction** x_i is simply the amount of phase i (measured in mols or particle numbers) divided by the sum of the amounts of all phases.
  - The **partial pressure** p_i of gas number i in a mixture of gases is simply the pressure that gas number **i** would have if you take all the other gases away and let it occupy the available volume. It follows that the total pressure $p=\Sigma_i\, p_i$ and $p_i/p=x_i$ (for ideal gases).
- With that we obtain for the chemical potential μ_i of the component i in a mixture of ideal gases

$$\mu_i^{mix}(p,T) = \mu_i^{pure}(p,T) + RT \cdot \ln \frac{p_i}{p}$$

- With p_i=partial pressure of component ***i*** and ***p*= actual pressure=Σp_i**
- *In words*: The chemical potential of gas number ***i*** in a mixture of gases at a certain temperature ***T*** and pressure ***p*** is equal to the chemical potential of this gas in the pure phase at ***p*** and ***T*** *plus* **RT· lnx_i**. *But note that* ***x_i < 1*** *for all cases* and thus **RT · lnx_i < 0**.
- Gases like to mix! It lowers their chemical potentials and thus their free enthalpy.

Now comes a big (and, to the eye of a physicist), somewhat confusing trick:

- We call $\mu_i^{pure}(p,T)$ now the *standard state* and write it μ_i^0 which is only the same thing as our old μ_i^0 as long as $p=p^0$, or, in the vacancy example above, $N=N_0$=**Lohschmidts** number (=number of particles in a mol). Again, you are free in your choices oft standard states - use it wisely!

Considering this, we obtain a kind of "*master equation*" for the chemical potential of the component ***i*** in some mixture of ideal gases:

$$\mu_i^{id}(p,T) = \mu_i^0 + RT \cdot \ln \frac{p_i}{p}$$

The **ln** term simply contains the entropy of mixing; otherwise, when we mix two gases, we would only add up the enthalpy/energy contained in the two pure components before the mixing.

- This is one way of writing down the chemical potential *for a mixture of gases*. Again note that whenever we see the Gas constant ***R*** instead of the Boltzmann constant **k**, you know that you are dealing with amounts that are taken *per mol* of a substance instead of per particle.

Again, what exactly is μ_i^0 now? Nothing but the reference for the energy scale, but nevertheless a quantity of prime importance, called the "*standard potential of component **i***" (the superscript "^{0}" always refers to the "standard" reference frame; in the case of gases mostly to atmospheric pressure and room temperature). It is also called **standard reaction enthalpy** and gives the change in the total free enthalpy at *standard conditions* if you wiggle the concentration of particle i a bit via

$$\Delta G^0 = \mu_i^0 \cdot \Delta n_i$$

In other words: $\mu_i^0 = \Delta G^0 / \Delta n_i$ or μ_i^0= the increase in enthalpy (or sloppily, energy) if you add a unit of the particles under consideration to the particles already in place.

- What do the equations mean? If we use the unit "particle", μ^0 is exactly the amount of free enthalpy needed to add (or subtract) one particle; usually given in **[eV/particle]** which is **[eV]**. If we use the unit "**mol**", it is the free enthalpy needed to add (or subtract) one mol, usually given in **[kJ/mol]**.

So far we have considered rather straight-forward thermodynamics; the difficulties arise if we use the concept of the chemical potential for **non-ideal gases**, for liquids and solids, for mixtures gases liquids and solids, or, as we do, for things like vacancies which are not usually described in those terms anyway. The first step is to consider non-ideal gases:

- If the gas is non-ideal, which means that it has some kind of interaction between its particles, it will obey some **virial equation** (any equation replacing $p \cdot V = N \cdot RT$). The simplest possible virial equation is $V = R \cdot T/p + B$ and for this we obtain

$$\mu^{non\text{-}id}(p) = \mu^0 + RT \cdot \ln \frac{p}{p^0} + B \cdot p$$

- For any other virial equation we can derive the corresponding formula for the chemical potential of that particular non-ideal gas. It will always have some extra terms containing the pressure.

However, to make things easy, chemists like to *keep the simple equation* for μ^{id} even in the case of non-ideal gases by *substituting the real pressure **p** by a quantity called* **fugacity *f*** chosen in such a way that the correct value for $\mu^{non\text{-}id}$ results.

- Fugacity and pressure thus are necessarily related and we define

$$f := \varphi \cdot p$$

- The dimensionless numberφ can always be calculated from the virial equation applicable to the situation. In our example we have

$$\ln\varphi = \frac{B \cdot p}{RT}$$

As long as we look at gases, there is no problem. Fugacity is a well defined concept, even if needs getting used to. The next step, however, is a bit more problematic.

Solids and Activities

Now we will turn to solids (and in one fell swoop we also include liquids in this). The good news is that the equation for a mix of ideal gases is equally valid for a mix of **ideal condensed** phases, i.e. **ideal solids**. The bad news is: An ideal solid in analogy to gases, i.e. without any interaction between the atoms, is an ***oxymoron*** (i.e. a contradiction in itself).

- What then are ideal solids supposed to be? Since we need interactions between the atoms or molecules, we must mean something different from gases. What is meant by "ideal" in this cases is that the interactions between the constituents of the solid are the same, regardless of their nature.
- Now that is certainly not a good approximation for most solids. So we use the same trick as in gases, we replace the mole fraction (which is a concentration) x_i of the component ***i*** by a quantity that contains the deviation from ideality; that quantity is called "**activity**" a_i.
- Again, we define the activity a_i of component **i** by

$$a_i := \varphi_i \cdot x_i$$

- With φ_i now carrying the burden of non-ideality.
- In contrast to gases, φ_i is *not* all that easily calculated, in fact it is almost quite hopeless. You may have to resort to an experiment and measure it.

In any case, if we use activities instead of concentrations or fugacities (which we treat as special case of activities), we are totally general and obtain for the chemical potentials of whatever component in any mixture:

$$\mu_i = \mu_i^0 + RT \cdot \ln a_i$$

Now, in looking at simple vacancies we already had the formula for the chemical potential of a vacancy; it read (if you put the various equations given in the link together):

$$\frac{\partial G}{\partial n_V} = 0 = G_F - kT \cdot \ln \frac{N}{n} = \mu_V$$

- with $n/N=n_V$, the equilibrium concentration of vacancies which we now also may call a_V, the activity of vacancies, if we want to be totally general.
- Wehave k instead of R, so we must be considering energies per particle and not per mol - which we did. We therefore do not have a mol fraction but a particle number fraction; but this is identical, anyway. All we have to do to get the activity is to reshuffle the **ln**:

$$\frac{\partial G}{\partial n_V} = \mu_V = G_F + kT \cdot \ln \frac{n}{N} = G_F + kT \cdot \ln a_V$$

- Now this is exactly the formula for an ideal gas or solid if we identify the formation enthalpy G_F of a vacancy with its standard chemical potential μ_0**(vacancy)** - and we did that already, too.
  - Replacing the concentration n/N of the vacancies with the activity of the vacancies is fine - but fortunately, for vacancy concentrations in elemental crystals, there is no difference between concentration and activity, because vacancy concentrations are always small (below 10^{-4}) - the vacancies are far apart and therefore do not interact very much - *they do behave like an ideal gas*!
- The situation, however, may be completely different for point defects in *large concentrations*, e.g. impurity atoms or vacancies and interstitials in ionic crystals.
  - The latter case is special because the *concentration of intrinsic point defects may depend on the stoichiometry and on impurities*: If there is e.g. a trace of Ca^{++} in a **NaCl** crystal, there must be a corresponding concentration of **Na** - vacancies to maintain charge neutrality and this concentration can not only be much larger than the maximum concentration in thermal equilibrium for "perfect" crystals, it will also be constant, i.e. independent of the temperature!
  - How to use the chemical potentials and activities in this context is described in a series of modules in the "backbone II" section of chapter 2. Here we will only give one example - equilibrium between phases.

Chemical Potential and Phase Equilibrium

- Consider some substance at constant pressure and temperature, but with *two possible phases*.
  - An everyday example is water in contact with ice, or any binary substance with a given composition (e.g. **Pb** and **Sn** - solder) at some point at its phase diagram where two phases coexist (consult the module "phase diagrams"), for that matter.
  - How many particles will be contained in phase **1** and how many in phase **2**? Given N particles altogether, we will have N_1 particles in phase **1** and $N_2 = N - N_1$ in phase **2**. How large is N_1?
- Lets look at the free enthalpy of the substance, or better yet, at its change with the particle numbers. In full generality, we have two equations:

$$1.\quad dG(p, T, N_1, N_1) = \frac{\partial G}{\partial T}\cdot dT + \frac{\partial G}{\partial p}\cdot dp + \frac{\partial G}{\partial N_1}\cdot dN_1 + \frac{\partial G}{\partial N_2}\cdot dN_2$$

$$2.\quad N_1 + N_2 = N = \text{const}$$

  - Since we look at a situation with *constant pressure and temperature*, we have that $dT = 0 = dp$.
  - For equilibrium, we demand $dG = 0$. From equ. (2) we get

$$dN_1 = -dN_2$$

  - Substituting that in equ. (1) yields

$$\frac{\partial G}{\partial N_1}\cdot dN_1 - \frac{\partial G}{\partial N_2}\cdot dN_1 = dG = 0$$

$$\frac{\partial G}{\partial N_1}\cdot dN_1 = \frac{\partial G}{\partial N_2}\cdot dN_1$$

$$\frac{\partial G}{\partial N_1} = \frac{\partial G}{\partial N_2}$$

$$\mu(N_1) = \mu(N_2)$$

  - ***q.e.d***.

- What happens if $\mu(N_1) > \mu(N_2)$; i.e. if we have non-equilibrium conditions with $\mu(N_1)$, *the chemical potential of the particles in phase **1***being larger than in phase **2**?
  - We now must change the particle numbers in the phases until equilibrium is achieved.
  - So do we have to increase N_1 (at the same time decreasing N_2) or should it go the other way around?
  - Well, whatever we do, *it must decrease **G***, so dG must be negative if we change the particle numbers the right way. For dG we had (a few lines above)

$$dG = \frac{\partial G}{\partial N_1} \cdot dN_1 - \frac{\partial G}{\partial N_2} \cdot dN_1$$

$$dG = \mu(N_1) \cdot dN_1 - \mu(N_2) \cdot dN_1$$

  - For positive dN_1, we will have $dG > 0$ since $\mu(N_1) > \mu(N_2)$. This necessarily leads to the general conclusion:
  - dN_1 **must be < 0** if the system is to move towards equilibrium.
- In words this means: The phase with the larger chemical potential will have to to shrink and the phase with the smaller chemical potential will grow until equilibrium is achieved and $\mu(N_1)=\mu(N_2)$.
  - *This is a very general truth*. Electrons, e.g., move from the phase with the higher chemical potential (than called Fermi energy) to the phase with the lower one.
  - We can also turn it around: Vacancies in supersaturation will tend to move to vacancy agglomerates and increase their size. It follows that the chemical potential of supersaturated single vacancies must be larger than that of vacancies in an agglomerate.
- Following up this line of thought leads straight to the law of mass action, which will be dealt with in another module.

Mass Action Law

Advanced

General Remarks

- This module is registered in the "advanced" part, because it uses the concept of the chemical potential also developed in the advanced part.
  - We use a rather general derivation, but do not go too deep into the details.
- The mass action law is usually taught in high school chemistry, so we know what we want to find: We look at some chemical reaction, e.g.

$$2H_2 + O_2 \Leftrightarrow 2H_2O$$

- The mass action law, as we know it, than asserts that the concentrations of the particles (= molecules in this case) in equilibrium can be written as

$$\frac{[H_2]^2 \cdot [O_2]}{[H_2O]^2} = K(T, p)$$

  - With **K =** reaction constant.
  - Or in words: The product of the concentration of the reaction partners with all concentrations always taken *to the power of their stoichiometric factors*, equals a constant **K** which has a numerical value that depends on the temperature and pressure. The constant ***K*** is called **reaction constant**.
- This statement, however, includes already a *generalization* and a *convention*:
  - There can be any number of particles reacting or resulting from the reaction, and we always bring the *results of the reaction*, (in the example the H_2O), to the *right side* of the equation and assign a *negative value to its stoichiometric factors* - the reaction products thus end up in the *denominator* of the concentration products. We mostly use *integers* for the stoichiometric factors, but that is not de rigeur.
  - An alternative way of writing the reaction equations that shows the "minus" sign more clearly, is

$$2H_2 + O_2 - 2H_2O = 0$$

  - The mass action law is deceptively simple, it is however not so trivial to derive it from thermodynamics *including a value for the reaction constant*, and it is often quite tricky to use for real cases!
- We will now give a standard derivation; an alternative way is given in another module.

Standard Derivation Using the Chemical Potential

- First we define arbitrary reactions of any kind by the equation

$$\nu_1 \cdot A_1 + \nu_2 \cdot A_2 + ... + \nu_f \cdot A_f = \nu_g \cdot A_g + \nu_{g+1} \cdot A_{g+1} + ... + \nu_i \cdot A_i$$

  - The A_x denote the particles (or reactions partners involved) - atoms, ions, molecules, vacancies, electrons, holes, .. - we want to be very general at this point. The corresponding stoichiometric factors are the ν_x, and they are usually (but not always) integers. Bringing the *products of the reaction* to the left side of the equation *which gives their stoichiometric factors a negative sign*, leads to the simple version

$$\sum_{1}^{i} \nu_i \cdot A_i = 0$$

- Chemical reactions as written down in standard notation always inherently assume that we have exactly the right amount of the chemicals (or, as we prefer to call it, particles) that are needed.
  - The reaction above, for our example, thus takes *two* mols of H_2 (= A_1) for *every* mol of O_2 (= A_2); or in our lingo, *two* H_2 particles (= molecules in this case) for *one* O_2 particle., yielding *two* H_2O (= A_3) particles.

- We have $\nu_1 = 2, \nu_2 = 1, \nu_3 = -2$.

Real life is different. You mix *some* number of H_2 particles with *some* number of O_2 particles, and after the reaction you have *some* number of all three particles involved (with one number probably being very low, or ideally zero, if the most scarce particle was completely used up in the reaction).

- In deriving the mass action law, we have to allow for this by allowing arbitrary starting concentrations c_i^0 of the particles involved including, if we wish, some concentration of the reaction products even before a reaction took place - nobody keeps us from filling some water into the container with H_2 and O_2 before we start the reaction.

We want to get a statement about the *concentration of the particles in equilibrium* for an *arbitrary mix of concentrations at the start of the reaction in non-equilibrium*; for ease of writing we denote the equilibrium concentration of the component ***i*** with c_i; the concentration at the start than is c_i^0, and an arbitrary concentration is C_i.

- The various c_i may be the number of mols, the absolute number of particles, or the concentration relative to some fixed value - it doesn't matter as long as the same definition is used throughout.
- As pointed out above, it is important to realize, that the c_i^0 can have any initial values whatsoever - you always can throw into a closed container whatever you want - but the dC_i; the *changes in the concentrations*, are tied to each other via the reaction equation.
- If you produce one mol of H_2O from any initial quantity of H_2 and O_2; you will have reduced the H_2 concentration by **1 mol** and the O_2 concentration by **0,5** mol - the dC_i thus are *not independent*.

The whole mixture of stuff - at whatever composition, i.e. for the whole range of the C_i - will have some free enthalpy $G(C_i, p, T)$.

- The important question is: For which concentration values of the various particles, do we have equilibrium and thus the minimum of G?
- In other words: For what conditions is $dG = 0$?
- Lets write it down. With $G = G(C_i, p, T)$ we have for dG

$$dG = \frac{\partial G}{\partial C_1} \cdot dC_1 + \frac{\partial G}{\partial C_2} \cdot dC_1 + \ldots + \frac{\partial G}{\partial C_i} \cdot dC_i + \frac{\partial G}{\partial T} \cdot dT + \frac{\partial G}{\partial p} \cdot dp$$

- The $(\partial G / \partial C_i)$ by definition are the chemical potentials μ_i of the particle sort x in the mixture, and the two last terms are simply $= 0$ if we look at it at constant pressure and temperature. For equilibrium.this leaves us with

$$dG = \sum_{1}^{i} \mu_i \cdot dC_i = 0$$

Now comes a decisive step. We know that our dC_i are tied somehow, but how?

- To see this, we "wiggle" the system a little and react some particles, changing the concentrations a little bit. As a measure of this change we introduce a "**reaction coordinate**" $d\xi$; a somewhat artificial, but useful quantity (without a unit).
- The changes in the concentrations of the various particles of our system then must be proportional to $d\xi$ *and the proportionality constants are the stoichiometric indices* ν_i. Think about it! However you wiggle - if the concentration of O_2 changes some, the concentration of H_2 will change twice as much.
- In other words, or better yet, in math, we have

$$dC_i = \nu_i \cdot d\xi$$

Substituting that into the equation for dG from above, we obtain

$$dG = \sum_{1}^{i} \mu_i \cdot \nu_i \cdot d\xi = d\xi \cdot \sum_{1}^{i} \mu_i \cdot \nu_i = 0$$

- Since $d\xi$ is some arbitrary number, the sum term must be zero by itself and we have as **equilibrium condition**

$$\sum_{1}^{i} \mu_i \cdot \nu_i = 0$$

- This looks (hopefully) familiar. It is the equilibrium condition we had before for particles not reacting with each other when we looked at the meaning of the chemical potential.

Now all we have to do is to take the "master equation" for the chemical potential so beloved by the more chemically minded, and plug it into the equilibrium condition for our reactions.

- In order to stay within our particle scheme, we use **k** instead of **R** and the activity A_i of the component **i** instead of its concentration C_i. Feel free to read "*activity*" as "*somewhat corrected concentration*" if you are unfamiliar or uncomfortable with activities. We have

$$\mu_i = \mu_i^0 + kT \cdot \ln A_i$$

- And, since we are treating equilibrium, the activity A_i now is the equilibrium activity a_i (= concentration c_i if everything would be "ideal") instead of the arbitrary concentration C_i because we are treating equilibrium now by definition.
- Inserting this formula in the equilibrium condition from above (and omitting the index "**i**" at the sum symbol for ease of writing) yields

$$\Sigma (\nu_i \cdot \mu_i^0) + kT \cdot \Sigma (\nu_i \cdot \ln a_i) = 0$$

Going through the mathematical motions now is easy.

- Expressing the sum of **ln**'s as the **ln** of the products of the arguments, and rearranging a bit gives

$$\ln \prod_{(a_i)} {}^{\nu_i} = -\frac{1}{kT} \cdot \Sigma \mu_i^0 \cdot \nu_i = -\frac{1}{kT} \cdot \Delta G^0$$

- Because $\Sigma \nu_i \cdot \mu_i^0$ is just the sum over all standard reaction enthalpies involved, which we call ΔG^0.

The product on the right hand side is just a fancy way to write down one part of the mass action law, it would give exactly what we formulated for the case of $2H_2 + O_2 \Leftrightarrow H_2 + O$ from above. Putting everything in the exponent finally yields the mass action law:

$$\prod_{(a_i)} {}^{\nu_i} = \exp -\frac{G^0}{kT} = K^{-1} = (\text{Reaction Constant})^{-1}$$

- It doesn't matter much, but it is standard to write K^{-1}. In other words, put the products of the reaction in the nominator to get **K**.

There seems to be a bit of magic involved: We started with arbitrary amounts of components, let them react an arbitrary amount (we even defined a new quantity, the reaction coordinate ξ) - and none of this shows up in the final formula! There are certainly some questions.

- What's left are only equilibrium concentrations (or activities) - what happened to the starting concentrations?
- Can't we derive the mass action law then without introducing quantities that seem not to be needed?

Some short answers:

- At some point, we essentially switched to changes (= derivatives) of prime quantities - and everything not changing is now gone. It is still there, however, if we do real calculations because then we need more information - the mass action law, after all, is just *one* equation for *several* unknown concentrations.
- There probably is a more direct way to get the mass action law that does not involve the somehow superfluous reaction coordinate. However - I do not know it and I'm in good company. Several text books I consulted do not know a better way either. Still, try the link for some alternatives.

Lets go back to our original question and mix arbitrary amounts of whatever and than let the buggers react. What will we get, throwing in the reaction equation and possibly some reaction enthalpies?

- The mass action law now gives us *one* relation between the equilibrium concentration, but not the absolute amounts. There are, after all, just as many unknowns for the equilibrium concentrations as you have components, and you need more than one equation to nail everything down.
- Additionally, the way we have spelled out the mass action law here also has a number of pitfalls; if you want to really use it, you must know a bit more, in particular about conventions that must be strictly adhered to.

All that is essentially beyond the scope of this "Defect" lecture, but for the hell of it, a few more modules intertwining mass action law and chemical potentials were made; they are accessible via the following links.

- Pitfalls and extensions of the mass action law
- Some standard (chemical) examples of applying mass action law
- Alternative derivations of the mass action law
- Some defects in ionic crystal related applications of the mass action law

Appendix: Some Necessary Math

This appendix contains one of the necessary mathematical transformations used above to refresh your memory

$$\Sigma \ln (c_i)^{\nu_i} = \ln(c_1)^{\nu_1} + \ln(c_2)^{\nu_2} + ... = \ln\{(c_1)^{\nu_1} \cdot (c_2)^{\nu_2} \cdots\} = \ln \Pi (c_i)^{\nu_i}$$

Pitfalls and Extensions of the Mass Action Law

Advanced

What is Reacting?

- Lets look at *ammonia synthesis*, a major chemical breakthrough at the beginning of the **20 th** century, as a pretty simple chemical reaction *between gases* (Remember: In the chemical formalism invoking the mass action law, point defects behave like (ideal) gases).
  - The reaction equation that naturally comes to mind is

$$N_2 + 3H_2 \Leftrightarrow 2NH_3$$

  - and the mass action law tells us that

$$\frac{[N_2] \cdot [H_2]^3}{[NH_3]^2} = K_1$$

  - With K_1 = reaction constant for this process. We used the square brackets **[..]** as the notation for concentrations, but lets keep in mind that the mass action law in full generality is *formulated for activities or fugacities*!
- However, we also could look at the *dissociation of ammonia* - equilibrium entails that some ammonia is formed, some decays; the "⇔" sign symbolizes that the reaction can go both ways. So lets write

$$2NH_3 \Leftrightarrow N_2 + 3H_2$$

  - The mass action law than gives

$$\frac{[NH_3]^2}{[N_2] \cdot [H_2]^3} = K_2 = \frac{1}{K_1}$$

- To make things worse, we could write the two equations also like

$$1/2N_2 + 3/2H_2 \Leftrightarrow NH_3$$

$$\frac{[N_2]^{1/2} \cdot [H_2]^{3/2}}{[NH_3]} = K_3 = (K_1)^{1/2}$$

  - and nobody keeps us from using the reaction as a source for hydrogen via

$$2/3NH_3 - 1/3N_2 \Leftrightarrow H_2$$

$$\frac{[NH_3]^{2/3} \cdot [N_2]^{1/3}}{[H_2]} = K_4 = ?$$

- And so on. Now *what does it mean ?* What exactly does the mass action law tell us? There are two distinct points in the examples which are important to realize:
  - **1.** Only the mass action law *together* with the reaction equation *and* the convention of what we have in the nominator and denominator of the sum of products makes any sense. A reaction constant given as some number (or function of ***p*** and ***T***) by itself is meaningless.

- **2.** The standard chemical potentials μ_i^0 that are contained in the reaction constant (via $\Sigma_i \mu_i^0$) where defined for reacting *one* standard unit, usually **1 mol**. The reaction constant in the mass action law thus is the reaction constant for producing **1** unit, i.e. one mol and thus applies, loosely speaking, to the component with the stoichiometry index **1**.
- That was N_2 in the first example. Try it. Rearranging the reaction equation to produce *one mol* of N_2 gives

$$2NH_3 - 3H_2 = N_2$$

$$\frac{[NH_3]^2 \cdot [H_2]^{-3}}{[N_2]} = K_1^{-1}$$

- Which is just what we had for the inverse reaction before.

So the right equation for figuring out what it takes to make *one mol* NH_3 is actually the one with the fractional stoichiometry indexes!

This looks worse than it is. All it takes is to remember the various conventions underlying the mass action law, something you will get used to very quickly in actual work. The next point is the tricky one!

Concentrations Relative to What?

Lets stick with the ammonia synthesis and give the concentrations symbolized by **[..]** a closer look. What we have is a *homogeneous* reaction, i.e. only gases are involved (a *heterogeneous* reaction thus involves that materials in more one kind of state are participating).

- We may then express the concentrations as partial pressures, (or, if we want to be totally precise, as fugacities). We thus have

$$[N_2] = p_{N2}$$

$$[H_2] = p_{H2}$$

$$[NH_3] = p_{NH3}$$

- And the total pressure is $p = \Sigma p_i$

But what is the actual total pressure?

- If we stick **1** mol N_2 and **3** mols H_2 in a vessel keeping the pressure at the beginning (before the reaction takes place) at its standard value, i.e. at atmospheric pressure, the pressure *must have changed* after the reaction, because we now might have only **2** mols of a gas in a volume that originally contained **4**!
- If you think about it, that happens whenever the number of mols on both sides of a reaction equation is not identical. Since the stoichiometry coefficients ν count the number of mols involved, we only have identical mol numbers before and after the reaction if $\Sigma \nu_i = 0$.

This is a tricky point and it is useful to illustrate it. Lets construct some examples. We take one reaction where the mol count changes, and one example where it does not. For the first example we take our familiar

$$N_2 + 3H_2 \Leftrightarrow 2NH_3$$

- We put **1** mol N_2 and **3** mols H_2, i.e. $N_0 =$ mols into a vessel keeping the pressure at its standard value (i.e. atmospheric pressure p_0). This means we need **4** "standard" volumes which we call V_0.
- Now let the reaction take place until equilibrium is reached. Lets assume that **90 %** of the starting gases react, this leaves us with **0,1** mol N_2, **0,3** mol of H_2, and **1,8** mols of NH_3. We now have **N = 2,2** mols in our container

The pressure p must have gone down; as long as the gases are ideal, we have

$$p_0 \cdot V_0 = N_0 \cdot RT$$

$$p \cdot V_0 = N \cdot RT$$

$$\Rightarrow \quad p \cdot \frac{N}{N_0} = 0{,}55\, p_0$$

- N or N_0 is the total number of mols contained in the reaction vessel at the pressure p or p_0, respectively.
- This equation is also valid for the partial pressure p_i of component i (with $\Sigma_i\, p_i = p$) and gives for the partial pressure and the number of mols N_i of component i, respectively

$$p_i = p_0 \cdot \frac{N_i}{N_0}$$

$$N_i = N_0 \cdot \frac{p_i}{p_0}$$

Now for the second example. Its actually not so easy to find a reaction between gases where the mol count does *not* change (think about it!), Lets take the formal reaction producing ozone, albeit a chemist might shudder:

$$2O_2 \Leftrightarrow O_3 + O$$

- Lets take comparable starting values: **2** mols O_2, **90 %** of which react, leaving **0,2** mol of O_2 and forming **0,8** mols of O_3 and **0,8** mols of **O** (think about it!) - *we always have two mols in the system*.

The mass action law followed from the chemical potentials and the decisive factor was **ln** c_i with c_i being a measure of the concentration of the component i. We had several ways of measuring concentrations, and it is quite illuminating to look closely at how they compare for our specific examples.

- In real life, for measuring concentrations, we could use for example:

  - The *absolute number of mols* $N_{i,mol}$ for component **i**. In general, the total number of mols in the reaction vessel, $\Sigma_i\, N_{i,mol}$, does *not* have to be constant as outlined above.
  - The *absolute particle number* $N_{i,\,p}$, which is the same as the absolute number of mols $N_{i,\,mol}$ if you multiply $N_{i,\,mol}$ with Avogadros constant (or Lohschmidt's number) $A = 6{,}02214\ mol^{-1}$; i.e. $N_{i,\,p} = A \cdot N_{i,\,mol}$. Note that the absolute number of particles (= molecules) does *not* have to stay constant, while the *absolute number of atoms*, of course, never changes.
  - The *partial pressure* p_i of component **i**, which is the pressure that we actually would find inside the reaction vessel if only the the component **i** would be present. The sum of all partial pressures p_i thus gives the *actual* pressure p inside the vessel; $\Sigma_i\, p_i = p$ and p does *not* have to be constant in a reaction. This looks like a violation of our basic principle that we look at the minimum of the free enthalpy at constant pressure and temperature to find the mass action law. However, the mass action law is valid for the equilibrium and the pressure at equilibrium - not for *how* you reach equilibrium!
  - The *activity* a_i (or the fugacity f_i) which for ideal gases is identical to $a_i = p_i/p = p_i/\Sigma_i\, p_i$. This is more or less also what we called the *concentration* c_i of component **i**.
  - The *mol fraction* X_i, which is the number of mols divided by the total number of mols present in the system: $X_i = N_{i,\,mol}/\Sigma_i\, N_{i,\,mol}$. This is the same thing as the concentration defined above because the partial pressure p_i of component i is proportional (for an ideal gas) to the number of mols in the vessel. We thus have $X_i = c_i$ ($= a_i = f_i$ as long as the gases are ideal).
  - The *"standard" partial pressure* p_i^0 defined relative to the standard pressure p^0. This is the pressure that we would find in our reaction vessel if we multiply all absolute partial pressure with a factor so that $p = p^0$. We thus have $p_i^0 = (p_i \cdot N_{i,mol}^0)/N_{i,\,mol}$ with $N_{i,\,mol}^0$ = number of mols of component i at the beginning of the reaction (and p = standard pressure) as outlined above.

- For ease of writing (especially in **HTML**), the various measures of concentrations will always be given by the square bracket **"[*i*]"** for component i .

We now construct a little table writing down the *starting* concentrations and the *equilibrium* concentrations in the same system of measuring concentrations. We then compute the reaction constant **K** for the respective concentrations, always by having the reaction products in the denominator (i.e taking $K = [NH_3]^2/[H_2]^3 \cdot [N_2]$ **or** $K = \{[O_3] \cdot [O]\}/[O_2]^2$, respectively).

Measure for c	Starting values	Equilibrium values	Reaction constant
N_{Mol} absolute number of Mols equivalent via $N_{i,p} = A \cdot N_{i,\,mol}$ to $N_{i,\,p}$ the absolute number of particles	$[H_2] = 3$ $[N_2] = 1$ $[NH_3] = 0$ $p = p^0$ $\Sigma_i\, N = 4$	$[H_2] = 0{,}3$ $[N_2] = 0{,}1$ $[NH_3] = 1{,}8$ $p = 0{,}55p^0$ $\Sigma_i\, N_i = 2{,}2$	$K_N = 1200$
	$[O_2] = 2$ $[O_3] = 0$ $[O] = 0$ $p = p^0$ $\Sigma_i\, N = 2$	$[O_2] = 0{,}2$ $[O_3] = 0{,}8$ $[O] = 0{,}8$ $p = p^0$ $\Sigma_i\, N_i = 2$	$K_N = 16$
Partial pressure p_i in units of p^0	$[H_2] = 3/4$ $[N_2] = 1/4$ $[H_3] = 0$	$[H_2] = 0{,}3/4 = 0{,}075$ $[N_2] = 0{,}1/4 = 0{,}025$ $[NH_3] = 1{,}8/4 = 0{,}450$	$K = 19\,200\ (p^0)^{-2}$
	$[O_2] = 2/2 = 1$ $[O_3] = 0$ $[O] = 0$	$[O_2] = 0{,}2/2 = 0{,}1$ $[O_3] = 0{,}4$ $[O] = 0{,}4$	$K_p = 16$
Activity a_i identical to the concentration c_i identical to the Mol fraction X_i	$[H_2] = 3/4$ $[N_2] = 1/4$ $[H_3] = 0$	$[H_2] = 0{,}3/2{,}2 = 0{,}136$ $[N_2] = 0{,}1/2{,}2 = 0{,}0454$ $[N_3] = 1{,}8/2{,}2 = 0{,}818$	$K_{act} = 5\,808$
	$[O_2] = 2/2 = 1$ $[O_3] = 0$ $[O] = 0$	$[O_2] = 0{,}2/2 = 0{,}1$ $[O_3] = 0{,}8/2 = 0{,}4$ $[O] = 0{,}82 = 0{,}4$	$K = 16$
"Standard" partial pressure p_i^0	$[H_2] = 3/4$ $[N_2] = 1/4$ $[NH_3] = 0$	$[H_2] = 0{,}136$ $[N_2] = 0{,}045$ $[NH_3] = 0{,}818$ $p_i^0 = 1.1818 p_i$	$K = 5\,914$
	$[O_2] = 2/2 = 1$ $[O_3] = 0$ $[O] = 0$	$[O_2] = 0{,}1$ $[O_3] = 0{,}4$ $[O] = 0{,}4$	$K = 16$

- Well, you get the point. The reaction constant may be wildly different for different ways of measuring the concentration of the components involved *if the mol count changes in the reaction* (which it mostly does).
  - Well, at least it appears that we do not have any trouble calculating **K** if the concentrations are given in whatever system. *But this is not how it works!* We do not want to compute **K** from measured concentrations, we want to use *known* reactions constants assembled from the standard reaction enthalpies or standard chemical potentials to calculate what we get.
  - So we must have rules telling us how to change the reaction constant if we go from from one system of measuring concentrations to another one.
  - Essentially, we need a translation from absolute quantities like particle numbers (or partial pressures) to relative quantities (= concentrations), which are always absolute quantities divided by some reference state like total number of particles or total pressure. The problem clearly comes from the changing reference state if the mol count changes in a reaction.
- Lets look at the the conversion from activities to particle numbers; this essentially covers all important cases.

Conversion of Reaction Constants

Well, lets go back to the final stage in the derivation of the mass action law and see what can be done. We had

$$\Pi\,(a_i)^i \;=\; \exp\left(-\frac{G_0}{kT}\right) \;=\; K \;=\; K_{act} \;=\; \text{Reaction constant for activities}$$

- The a_i are the activities, which we defined when discussing the chemical potential analogous to the fugacities for gases. Fugacities, in turn, were introduced to take care of non-ideal behavior of gases.
- However, as long as we look at gases and as long as they are ideal, the fugacity (or activity), the prime quantity in the chemical potential for gases was the *concentration* of gas **i** given by its partial pressure p_i divided by the actual pressure p, a relative quantity. For the purpose of this paragraph it is sufficient to consider

$$a_i \;=\; \frac{p_i}{p} \;=\; \frac{p_i}{\Sigma_i\, p_i}$$

Lets now switch to an absolute quantity. We take the number of mols of gas i. $N_{i,\,mol}$; now lets see how the mass action law changes.

- We can express p_i by

$$p_i \;=\; N_i \cdot \frac{p_0}{N^0}$$

- With p^0 = standard pressure, and N_0 = starting number of mols, and $p = \Sigma p_i = (\Sigma N_i) \cdot p_0/N^0$.
- With this we can reformulate the mass action law by substituting

$$\frac{p_i}{p} \;=\; \frac{N_i \cdot \dfrac{p_0}{N^0}}{\Sigma\, N_i \cdot \dfrac{p_0}{N^0}} \;=\; \frac{N_i}{\Sigma\, N_i}$$

- This gives (afer some fiddling around with the products and sums)

$$\ln\Pi\,(a_i)^{\nu_i} \;=\; \ln\,\Pi\left(\frac{p_i}{p}\right)^{\nu_i} \;=\; \ln\,\Pi\left(\frac{N_i}{\Sigma N_i}\right)^{\nu_i} \;=\; \ln\left(\left(\Sigma N_i\right)^{-\Sigma\nu_i} \cdot \left(\Pi\, N_i\right)^{\nu_i}\right) \;=\; \ln K_{act}$$

- If this looks a bit like magic, you are encouraged to go through the motions in fiddling around the products and the sums yourself. If you don't want to - after all we are supposed to be dealing with defects, not with elementary albeit tricky math - look it up.
- We want the mass action law for the particle numbers N_i, i.e. we want an expression of the form

$$\Pi\,(N_i)^{\nu_i} \;=\; K_N$$

So if we write down the mass action law now for particle number N_i we have

$$\prod (N_i)^{\nu_i} = \left(\Sigma N_i\right)^{\Sigma \nu_i} \cdot K_{act} = K_N$$

$$K_N = \left(\Sigma N_i\right)^{-\Sigma \nu_i} \cdot K_{act}$$

Lets try it. For our ammonia example we have

$$\Sigma N_i = 2{,}2$$

$$\Sigma \nu i = 1 + 3 - 2 = 2$$

$$\left(\Sigma N_i\right)^{\Sigma \nu_i} = 2{,}2^2 = 4{,}84$$

- Well, the two constants from the table above are $K_N = 1200$ and $K_{act} = 5\,808$; $K_{act}/K_N = 4{,}84$ as it should be? Great - but shouldn't it be the other way around?
- Indeed, we should have $K_N/K_{act} = 4{,}84$ according to the formula above - just the other way around. However, the way we formulated the mass action law above, we should have written K^{-1} to compare with the values in the table!

OK; this is unfair - but look at the title of this subchapter!

One last word before we turn irreversibly into chemists:

- With the equations that couple pressure and mol-numbers, we can express ΣN_i by $\Sigma N_i = p \cdot (N^0/p_0)$ which, inserted into the expression between mass action constants from above, gives

$$K_N = \left(p \cdot \frac{N^0}{p_0} \right)^{\Sigma \nu_i} = p \cdot K'$$

- In words: The reaction constant is proprotional to the pressure. If you do not just accept whatever pressure you will get after a reaction, but keep the system at a certain pressure, you can influence how much (or little) of the reaction products you will get.

Lets deal with the $\ln \prod (N_i/\Sigma N_i)^{\nu_i}$ term step by step:

- First it is important to realize that ΣN_i is a fixed number. Even so it has an index **i**, after the summation is done the index is gone and it does not get "afflicted" by the $\prod$ sign.
- We thus have .

$$\ln \prod \left(\frac{N_i}{\Sigma N_i} \right)^{\nu_i} = \frac{\left(N_1\right)^{\nu_1} \cdot \left(N_2\right)^{\nu_2} \cdot \ldots}{\left(\Sigma N_i\right)^{\nu_1} \cdot \left(\Sigma N_i\right)^{\nu_2} \cdot \ldots} = \frac{\prod \left(N_i\right)^{\nu_i}}{\prod \left(\Sigma N_i\right)^{\Sigma \nu_i}}$$

- Keeping in mind that $\ln (a/b^x) = \ln (a \cdot b^{-x}) = \ln b^{-x} + \ln a$, we obtain .

$$\ln \prod \left(\frac{N_i}{\Sigma N_i} \right)^{\nu_i} = \ln \left(\left(\Sigma N_i\right)^{-\Sigma \nu_i} \cdot \prod \left(N_i\right)^{\nu_i} \right)$$

Chemical Examples for Mass Action Law Applications

Advanced

- What can one do with the mass action law? A lot - but it is not always very obvious. Lets ask a few "dumb" questions and see how far we get.
- First we look at a really simple reaction, e.g

$$H_2 + Cl_2 \Leftrightarrow 2HCl$$

  - To keep it easy, we start with equal amounts of H_2 and Cl_2 .
- How much **HCl** do we get? Notice that we have the same number of mols on both sides of the reaction equation.
  - Well, in equilibrium (denoted by **[..]**) we have

$$\frac{[HCl]^2}{[H_2] \cdot [Cl_2]} = K$$

  - or, with $[H_2] = [Cl_2] = [equ]$

$$[HCl] = [equ] \cdot K^{1/2}$$

- One equation with two unknowns; not sufficient for calculating numbers.
  - But then we also have the condition that the number of the atoms involved stays constant, i.e.$[H_2] + 2[HCl] =$ constant = e.g. the number of H_2 mols before the reaction.

- Next, a little bit harder. Lets start with arbitrary concentrations of something and see what we can say about the *yield* of the reaction. For varieties sake lets look at

$$H_2 + CO_2 \Leftrightarrow H_2O + CO$$

  - Again a simple reaction with the same number of mols on both sides, so we do not have to worry about the precise form of the mass action law.
- We start with n^0_{H2} and n^0_{CO2} mols of the reacting gases and define as the *yield* ***y*** the number of mols of H_2O that the reaction will produce at equilibrium. This leaves us with

$$n_{H2O} = y$$

$$n_{CO} = y$$

$$n_{H2} = n^0_{H2} - y = \text{equilibrium concentration of } H_2$$

$$n_{CO2} = n^0_{CO2} - y = \text{equilibrium concentration of CO}$$

$$\Sigma n = n^0 = n^0_{H2} + n^0_{CO2}$$

  - The last equation holds because the mol count never changes in this example.
  - The mass action law now gives

$$\frac{y^2}{(n^0_{H2} - y) \cdot (n^0_{CO2} - y)} = K$$

$$y = \left(\frac{1}{2}\left(1 - K\right)\right) \cdot \left(\left(-n^0 \cdot K \pm \left((n^0 \cdot K)^2 + 4 \cdot (1 - K) \cdot n^0_{H2} \cdot (n^0 - n^0_{H2}) \cdot K\right)^{1/2}\right)\right)$$

- The starting concentration of CO_2, i.e n_{CO2}, is expressed as $n_{CO2} = n^0 - n^0_{H2}$.

Looks extremely messy, but this is just the standard solution for a second order equation. Whatever this solution means in detail, it tells us that the yield is a function of the starting concentrations of the ingredients.

- What kind of starting concentrations will give us *maximum* yield? To find out, we have to form $dy/dn^0_{H2} = 0$.
- Well, go through the math yourself; this is elementary stuff. The solution is

$$n^0_{H2} = \frac{n^0}{2}$$

$$n^0_{CO2} = \frac{n^0}{2} = n^0_{H2}$$

In other words: maximum yield is achieved if you mix just the right amounts of the starting stuff. This result is always true, even for more complicated reactions.

- At this point we stop, again because otherwise we might turn irreversibly into chemists.

Alternative Derivations of the Mass Action Law

Advanced

If despite all the efforts made in this Hyperscript you still don't like chemical potentials - here is the physicists way to deduce the mass action law *without* invoking chemical potentials and all that.

We start from the reaction equation with stoichiometric indices as before

$$\Sigma_i \nu_i \cdot A_i = 0$$

- If we denote with N_i the quantity of substance A_i in *mols*, the free enthalpy G of the mixture contains the sum of the free enthalpies g_i of the constituents and the mixing entropy S_m, we have

$$G = \Sigma_i (N_i \cdot g_i) - T S_m$$

S_m is calculated in the usual way by considering the number of possibilities for mixing the substances in question, as a result one obtains

$$S_m = - R \cdot \left(\Sigma_i N_i \cdot \ln \frac{N_i}{\Sigma_i N_i} \right)$$

- In this formulation the **ln** is negative because $(N_i/\Sigma_i N_i) << 1$, and we thus must assign a negative sign to the total entropy, cf. the link.
- The total free enthalpy now is

$$G = \Sigma_i \cdot (N_i \cdot g_i) + T \cdot R \left(\Sigma_i N_i \cdot \ln \frac{N_i}{\Sigma_i N_i} \right)$$

We are looking for the minimum of the free enthalpy. For that we consider what happens if we change the N_i by some ΔN_i.

- This changes G by some ΔG expressible as a total differential $\Delta G = \Sigma_i (\partial G / \partial N_i) \cdot \Delta N_i$. It is not much fun to go through the motions, but it is just a simple differentiation job without any problems. We obtain

$$\Delta G = \Sigma_i (\Delta N_i \cdot g_i) + T \cdot R \left(\Sigma_i \ln (N_i \cdot \Delta N_i) + \Sigma_i (\Delta N_i) - \Sigma_i \left(\Delta N_i \cdot \ln (\Sigma_i N_i) \right) - \Sigma_i \left(N_i \cdot \frac{\Sigma_i \Delta N_i}{\Sigma_i N_i} \right) \right)$$

- Horrible, but it can be simplified (the third and fifth term actually cancel each other) and expressed via the reaction coordinate ξ using

$$\Delta N_i = \nu_i \cdot \Delta \xi_i$$

- We had that before; we obtain

$$\Delta G = \Delta \xi \cdot \left(\Sigma_i g_i \cdot \nu_i + RT \cdot \Sigma_i \nu_i \cdot \ln N_i - RT \cdot \Sigma_i N_i \cdot \ln (\Sigma_i \nu_i) \right)$$

For equilibrium we demand $\Delta G = 0$ and this means that the expression in large brackets must be zero by itself:

$$0 = \left(\Sigma_i g_i \cdot \nu_i + RT \cdot \Sigma_i \nu_i \cdot \ln N_i - RT \cdot \Sigma_i N_i \cdot \ln (\Sigma_i \nu_i) \right)$$

- Division with RT, putting both sides in the exponent, noticing that a sum in an exponent can be written as a product, and dropping the index i at ν, g, and N because it is clear enough by now (and cannot be written properly in **HTML** anymore), yields

$$\prod_i N^{\nu} = \left(\Sigma_i N\right)^{\Sigma\nu} \cdot \exp - \frac{\Sigma_i g_i \cdot \nu_i}{RT}$$

- which is the mass action law in its most general form.

No chemical potentials μ, no standard chemical potential μ_0, no fugacities or activities. *Everything is clear*.

- The catch, of course, is the entropy formula. It is *only* valid for classical non-interacting particles. However, if this is not the case, it is clear what need to be modified - it may not be so clear, however, how.
- The other issue that takes perhaps a little thought, is the sum of the free enthalpies $\mathbf{g_i}$ of the constituents. Since we look at the equilibrium situation, it is the free enthalpy of one mol of the substances present with respect to the prevailing condition.

Schottky Defects

- Historically, point defects in crystals were first considered in ionic crystals, not in the much simpler metal crystals. The reason was that some known properties of ionic crystals (e.g. their conduction mechanism by ion migration at high temperatures) could be understood for the first time in terms of point defects, while no special properties of metals (in the twenties) were in desperate need of an explanation.
  - Since point defects in ionic crystals are charged, they only can come in pairs to maintain charge neutrality.
  - **Schottky defects** then are differently charged pairs of vacancies, i.e. missing **Na^+** and **Cl^-** ions in the **NaCl** crystal (the other principally possible pairing of point defects is described by Frenkel defects). Schottky defects are dealt with in chapter 2.1.3.
  - Since the number of atoms has to stay constant, no matter how many Schottky defects are present, the surplus atoms must be thought of as sitting on the surface - the crystal expands (measurably) when Schottky defects are formed!
- Researchers with a chemical or ceramics background tend to classify all point defects in the category "Schottky" or "Frenkel".
  - In this classification system, the simple (uncharged) vacancy in metals the would be a Schottky defect.
  - However, it is not always useful to force all possible point defect assemblies in the narrow corset of "Schottky" or "Frenkel". The general situation of arbitrary numbers of several different point defects will be dealt with in chapter 2.2

Frenkel Defects

- Historically, point defects in crystals were first considered in ionic crystals, not in the much simpler metal crystals. The reason was that some known properties of ionic crystals (e.g. there conduction by ion migration at high temperatures) could be understood for the first time in terms of point defects, while no special properties of metals (in the twenties) were in desperate need of an explanation.
  - Since point defects in ionic crystals are charged, the only can come in pairs to maintain charge neutrality.
  - **Frenkel defects** than are charged interstitial - vacancy pairs carrying automatically different charge, e.g. a vacancy on a Na^+ site and a Na^+ interstitial (the other principally possible pairing of point defects is described by Schottky defects). Frenkel defects are dealt with in detail in chapter 2.1.2.
  - In contrast to Schottky defects, there is no (or only a negligible) volume expansion of the crystal when Frenkel defects are formed.
- Researchers with a chemical or ceramics background tend to classify all point defects in the category "Schottky" or "Frenkel"
  - In this classification system, Frenkel defects do not appear in thermal equilibrium in simple (elemental) crystals. They may, however be produced in non-equilibrium, e.g. by energetic irradiation which transfers sufficient energy to crystal atoms to displace them into interstitial sites while at the same time creating a vacancy.
  - The defect situation in Si, however, where vacancies and interstitials coexist in thermal equilibrium in comparable, but not necessarily equal amounts, cannot be accounted for in the "Schottky" /"Frenkel" system.
- It is thus not always useful to force all possible point defect assemblies in the narrow corset of "Schottky" or "Frenkel". The general situation of arbitrary numbers of several different point defects will be dealt with in chapter 2.2

Internal Energy, Enthalpy, Entropy, Free Energy, and Free Enthalpy

A Thermodynamics Primer

General Remarks and State Functions

- Let's face it: Thermodymanics is not easy! *It is not possible to learn it by just reading through this module.*
  - However, if you fought your way through thermodynamics proper at least once, and thus are able to look at it from a distance without getting totally confused by the "details" (which you don't have to know anymore, but must be able to understand when they come up), it's not so difficult either.
  - It gets even easier by restricting ourselves to solids which means that most of the time we don't have to worry about the *pressure* ***p*** anymore – it is simply constant. We nevertheless specify it here for the sake of general validity.
- In this primer we will review the most important issues necessary for understanding defects, including defects in semiconductors. In order to stay simple, we must "cut corners". This means:
  - We will usually not show the functional relationships by showing the variables. We thus simply write ***G*** for the free enthalpy, and not $G(T, p, n_i)$, showing that ***G*** is a function of the temperature ***T***, the pressure ***p*** and the particle numbers n_i.
  - In the same spirit, we will omit the indexes showing what stays constant for partial derivations, i.e. we write for the chemical potential μ_i of the particle sort **i** the simple form

$$\frac{\partial G}{\partial n_i} = \mu_i$$

  - In full splendor it should be

$$\left.\frac{\partial G(T, p, n_j)}{\partial n_i}\right|_{p,\, T,\, n_{j \neq i}} = \mu_i(T, p, n_j)$$

  - We also are sloppy about standards. n_i may refer to particle numbers *or* concentrations; in the latter case **particles/cm^3** or **mol/cm^3** - you must know what is meant from the context. You are also supposed to know that if Boltzmann's constant **k** comes up in an equation, we are working with properties per particle, whereas the gas constant **R** signifies properties per mol.
- This, admittedly, is dangerous. But multi-indexed quantities are confusing (and not easily written in HTML, anyway)! Let's stay simple and refer to complications whenever they come up.
- If we restrict ourselves to crystals, it is rather easy to consider the concepts behind the all-important thermodynamic quantities **Internal Energy**, **Enthalpy**, **Entropy**, **Free Energy** and **Free Enthalpy**. We start with the *internal energy* ***U*** of a crystal.
  - Neglecting *external* energies (e.g. the gravitational potential) and *internal energies that never change* (e.g. the energy of the inner electrons), we are essentially left with the internal energy ***U*** being contained in the *vibrations of the crystal atoms* (or molecules), which express themselves in the *temperature* ***T*** *of the system* according to

$$U = \tfrac{1}{2} \cdot f \cdot kT$$

  - With ***U*** = average energy per atom, ***f*** = degree of freedoms for "investing" energy in an atom (**f = 6** for crystals; **3** for the kinetic energy in v_x, v_y, v_z, and **3** for the potential energy at **(*x*, *y*, *z*)**); **k =** Boltzmann's constant and ***T*** = (absolute) temperature
- The *(macro)state of the system* is thus given by the number of atoms ***N***, the pressure ***p*** and the temperature ***T***. Knowing these numbers is all there is to know about the system *on a macroscopic base*.
  - We can change the state of the system by adding or removing heat ***Q***, putting mechanical work ***W*** into the system or taking it out, and by changing the number of atoms (or more generally, particles) by some ΔN.
  - Since at this point we keep the number of atoms in our crystal constant, we only have to consider ***Q*** and ***W*** if we change the state. The following basic equations (a formulation of the **1st law of thermodynamics**) *(german link)* holds

$$\mathbf{d}U = \mathbf{d}Q - \mathbf{d}W$$

- With the changes written in differential form. Note that the regular **"d"** is not the sign for (partial) derivatives (that would be ∂) but for "delta". **d*U***, e.g., stands for **total change of *U***. Note that *sometimes* the **d** indicates a total differential (e.g. in the case of **d*U***), sometimes it does not (e.g. in the case of **dQ** or **d*W***).
- If we reserve the letter "**d**" for total differentials only, the equation above acutally should have been written as $\mathbf{d}U = \delta Q - \delta W$ but we will not be that rigid here.
- Any mechanical work must change the volume (something must move); for the normal conditions encountered with crystals where the pressure stays constant it can always be expressed as

$$\mathbf{d}W = p \cdot \mathbf{d}V$$

This term ***p*dV*** is cumbersome as long as only situations involving crystals under constant pressure are considered. We thus introduce a *new state function* called enthalpy ***H*** and define it as

$$H = U + pV$$

- If we again change the state of the system by adding or subtracting heat ***Q*** and mechanical work ***W***, we now obtain for the total *change in enthalpy* **d*H***

$$\mathbf{d}H = \mathbf{d}U + p\mathbf{d}V + V\mathbf{d}p$$

- With $V\mathbf{d}p = \mathbf{0}$, because the pressure ***p*** is constant, and $\mathbf{d}U = \mathbf{d}Q - p\mathbf{d}V$ we obtain

$$\mathbf{d}H = \mathbf{d}Q$$

- This is a simple relation always best suited for systems under constant pressure and also clarifying why we tend to think of enthalpy as **heat**.
- **d*H*** is a measure of the energy needed to form a substance in a given state, it is occasionally also called the **heat of formation** (always refering to the difference between two states).
- Of course, not much happens if the substance is just heated a bit but does not change its chemical nature – let's say we look at a mixture of $\mathbf{H_2}$ and $\mathbf{O_2}$ which we heat up a bit. All the fun comes from chemical reactions (or phase changes) – in our example it would be the formation of $\mathbf{H_2O}$ in a somewhat violent fashion.

It was thought that the sign of **d*H*** would indicate if a reaction should or should not occur. A negative sign would mean that the reaction would transfer energy to the surroundings and thus could easily happen, whereas a positive sign would tell us that energy would have to be pumped into the system – nothing would happen by itself.

- It's not that simple! While this point of view was true enough for relatively large **d*H*** (let's say **> 100 kcal/mol**), the criterion often does not work for smaller changes of **d*H***.
- The reason, of course, is that we neglected the change of the **entropy *S*** of the system, **d*S***, that occurs parallel to **d*H***.

Purely mechanical systems (consisting of non-interacting mass points) would be in equilibrium for the lowest possible internal energy, i.e. for a minimum in their potential energy and no movement – just lying still at the lowest possible point. *But thermodynamic systems* consisting of many interacting particles and some externally fixed condition (e.g. a constant temperature), are in equilibrium if the *best possible balance between a **small energy and a large entropy** is achieved*!

- We just take that as an article of faith (or law of nature) at this point.
- Often, both quantities are opposed to each other: High entropies mean high energies and vice verse. The entropy part becomes more important at high temperatures, and the thermodynamic potential which has to be minimized for systems under constant pressure, is the **free enthalpy *G*** (also called **Gibbs energy**). It is defined as

$$G = H - T \cdot S$$

- With ***S*** = **entropy** = $\mathbf{d}Q_{rev}/T$ in classical thermodynamics (the suffix "rev" refers to reversible processes).
- If you have a system with constant volume (and variable pressure), the best suited state function is the **free energy *F*** (also called **Helmholtz energy**). It is defined as

$$F = U - T \cdot S$$

- Before turning to the entropy, a word to the choice of state functions. We now already have four: ***U***, ***H***, ***G***, ***F*** – but for a given system, there is only *one* state. Two things are important in this context:
  - State functions, by definition, must describe the state of a system no matter how this state developed – they must, in other words, meet all the requirements for potentials and thus are **thermodynamic potentials**. We have not proved if this is the case for ***U***, ***H***, ***G***, ***F*** – turn to the potential module for some input to this question – but they really are potentials.
  - *Any* state function or thermodynamic potential can be used to describe *any* system (always for equilibrium, of course), but for a given system some are more convenient than others. The most convenient (and thus important) one for crystals (usually under constant pressure) is the free enthalpy.

Entropy - Statistical Consideration

- The key question is:

What is entropy?

  - There is a classical answer, but here we only use the statistical definition where *entropy is the measure of the "probability" **w** of a given macrostate*, or, essentially the same thing, *the number **P** of microstates possible for the given macrostate*.
  - Not too helpful: What is a **microstate** or a **macrostate**? Or the **probability of a macrostate**?
  - Well, *any* particular arrangement of atoms (or more generally, particles) where we look only on *average quantities* is a *macrostate*, while any *individual arrangement* defining the properties (e.g. location and momentary velocity) of all the particles for a given macrostate is a *microstate*.
  - In other words, and somewhat simplified: For a microstate it matters what individual particles do, for the macrostate it does not.
  - The difference between microstates and macrostates is best illustrated for for a gas in a closed container: We can define many possible macrostates, e.g.

    - **1.** All molecules are in the left half of the container.
    - **2. 70 %** of the molcules are in the left half of the container, **30 %** in the right half.
    - **3.** Equal (average) distribution of the molecules.

    and all these macrostates (plus many more) could have exactly the same internal energy ***U*** (or ***H***).
  - However, the *probability* of experimentally finding one or the other of those macrostates is very different. The probabilities of the macrostates **1.** and **2.** are certainly much much smaller than the probability of macrostate No. **3.**
  - For all the possible macrostates, the *state function* tells us which one will be realized (= is most probable) in thermal equilibrium.
- How do we calculate the probability of a macrostate? Let's see:
  - For every possible macrostate we can think of, there are many microstates to realize it. Its exactly like playing dice: Let's assume you have **3** dice. A macrostate would be some possible number you may throw; e.g. **9**. The corresponding microstates are the possible combinations of the individual dice. For throwing **9** we have

Dice	1. Poss.	2. Poss.	3. Poss	....
1	1	1	1	
2	1	2	3	
3	7	6	5	

  - - and so on. You get the picture.
- The *probability* for such a macrostate would be the number of microstates divided by the number of all possible combinations of the dice (which is a constant). We can see off-hand that the macrostates **"3"** and **"18"** are the most unlikely ones, having only one microstate at their disposal, while **9**, **10**, or **12** are more likely to occur.

- Now we know what *the number of possible ways to generate the same macrostate* means and why the "probability" ***w*** of a given macrostate is *"almost"* the same thing.

An example just as easy as playing dice, comes from our friend, the vacancy. We simply ask: How many ways ***P*** (= *microstates*) are there to arrange ***n*** vacancies (= *macrostate*) in a crystal of ***N*** atoms?

- When we figure that out, we can use the equilibrium condition to select the most likely macrostate and this gives us the number of vacancies in equilibrium.

The fundamental point now is that just knowing the internal energy ***U*** of a system with a constant volume and temperature is not good enough to tell us what the equilibrium configuration will be because *we* could think of many macrostates with the same ***U*** (and mother nature, to be sure, can come up with lots more).

- That's why just minimizing ***U*** (or ***H***) is not good enough, we have to minimize ***F = U – TS*** or ***G = H – TS*** to find the equilibrium configuration of the system, and for that we have to know the entropy, because we now can interprete these formulas:
- Of all the many *macrostates* possible for a given ***U*** (or ***H***) the one with the *largest entropy* at the given temperature will be the one that the system will adopt

Obviously, we need to be able to calculate the entropy of a certain macrostate and this is done by employing the statistical definition of the entropy ***S***, the famous **Boltzmann entropy equation** *(german link)*:

$$S = k \cdot \ln w$$

or

$$S = k \cdot \ln P$$

With ***w*** = probability of a macrostate and ***P*** = number of microstates for a macrostate.

If you feel that the ambiguity with respect to taking ***w*** or ***P*** is a bit puzzling - that's because it is! You should consult the link to see that at least it is nothing to worry about. Whatever you chose to work with, the results you will get in the end will be the same.

- Entropy ***S***, by the way, is *not a state function* (***TS*** would be one).

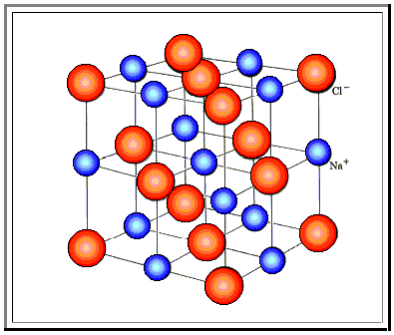
We used the statistical definition of entropy and the minimization of the free enthalpy in chapter 2.1; and in an exercise module it can be seen in detail how to apply it to derive the formula for the vacancy concentration

Ionic Crystals

- *Ionic* crystals have at least two atoms in their base which are ionized. Charge neutrality demands that the total charge in the base must be zero; so we always need ions with opposing charge.
  - The binding between the ions is mostly electrostatic and rather strong (binding energies around **1000 kJ/mol**); it has no directionality.
  - Ionic crystals thus can be described as an ensemble of hard spheres which try to occupy a minimum volume while minimizing electrostatic energy at the same time (i.e. having charge neutrality in small volumes, too).
  - There are no free electrons, ionic crystals are insulators.
- Ionic crystals come in simple and more complicated lattice types; the latter is true in particular for oxides which are often counted among ionic crystals. Some prominent lattice types follow

The NaCl Structure

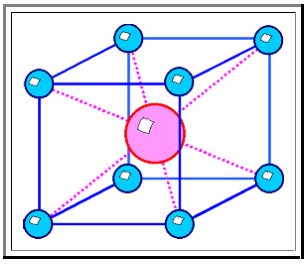
- The lattice is *face centered cubic* **(fcc)**, with *two* atoms in the base: one at **(0, 0, 0)**, the other one at **(½, 0, 0)**



  - Many salts and oxides have this structure, e.g. **KCl**, **AgBr**, **KBr**, **PbS**, ...
    or
    MgO, **FeO**, ...

The CsCl Structure

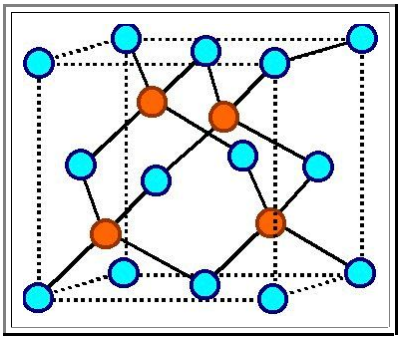
- The lattice is *cubic primitive* with *two* atoms in the base at **(0,0,0)** and **(½, ½, ½)**. It is a common error to mistake it for a bcc lattice.



  - Intermetallic compounds (not necessarily ionic crystals), but also common salts assume this structure; e.g. **CsCl**, **TlJ**, ...,
    or **AlNi**, **CuZn**,

The ZnS (or Diamond, or Sphalerite) Structure

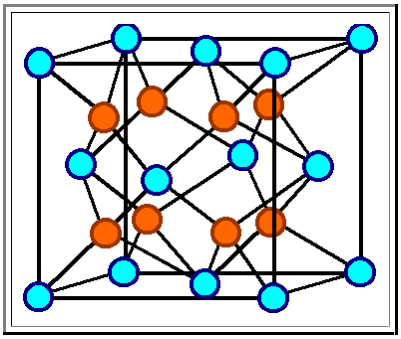
The "**zinc blende**" lattice is *face centered cubic* **(fcc)** with two atoms in the base at **(0,0,0)** and **(¼, ¼, ¼)**.



- It is not only an important lattice for other ionic crystals like **ZnS**, which gave it its name, but also the typical lattice of *covalently bonded group IV semiconductors* (**C** (diamond form), **Si**, **Ge**) or III-V compounds semiconductors (**GaAs, GaP, InSb, InP,** ..)
- The **ZnS** lattice is easily confused with the **ZrO_2** lattice below.

The CaF_2 or ZrO_2 Structure

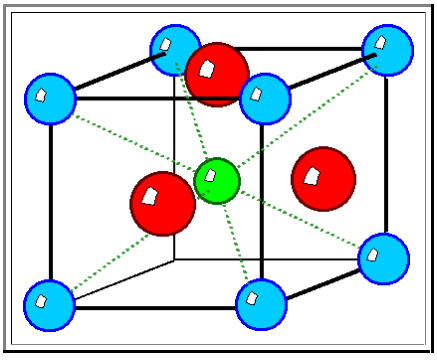
The lattice is *face centered cubic* **(fcc)** with *three* atoms in the base, one kind (the cations) at **(0,0,0)**, and the other two (anions of the same kind) at **(¼, ¼, ¼)**, and **(¼, ¾, ¼)**.



- It is often just called the "**fluorite structure**".

Perovskite Structure

The lattice is essentially *cubic primitive*, but may be distorted to some extent and then becomes *orthorhombic* or worse. It is also known as the **$BaTiO_3$** or **$CaTiO_3$** lattice and has *three* different atoms in the base. In the example it would be **Ba** at **(0,0,0)**, **O** at **(½, ½, ,0)** and **Ti** at **(½, ½, ½)**.

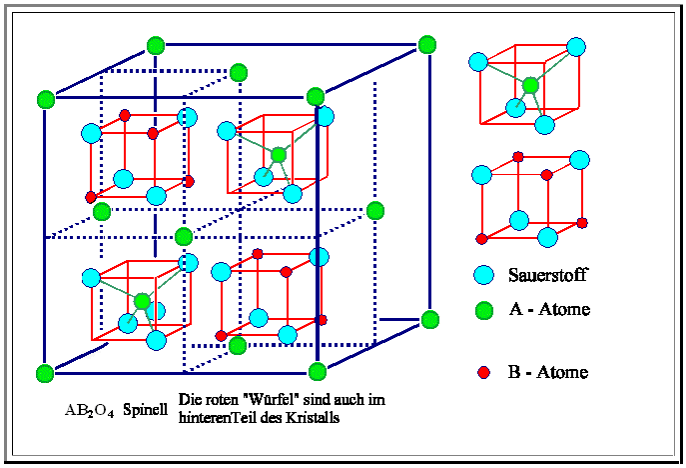


- A particular interesting perovskite (at high pressures) is **$MgSiO_3$**. It is assumed to form the bulk of the mantle of the earth, so it is the most abundant stuff on this planet, neglecting its Fe/Ni core. The mechanical properties (including the movement of dislocations) of this (and related) minerals are essential for geotectonics - forming the continents, making and quenching volcanoes, earthquakes - quite interesting stuff!

Spinel Structure

The spinel structure (sometimes called **garnet structure**) is named after the mineral *spinel* (**$MgAl_2O_4$**); the general composition is **AB_2O_4**. It is essentially *cubic*, with the **O** - ions forming a fcc lattice. The cations (usually metals) occupy **1/8** of the tetrahedral sites and **1/2** of the octahedral sites and there are **32 O**-ions in the unit cell.

- This sounds complicated, but it is not as bad as it could be; look at the drawing. We "simply" have two types of cubic building units inside a big **fcc O**-ion lattice, filling all **8** octants.



- The spinel structure is very flexible with respect to the cations it can incorporate; there are over **100** known compounds. In particular, the **A** and **B** cations can mix! In other words, the composition with respect to one unit cell can be

  - **$(A_8)\ (B_{16})O_{32}$**, or
  - **$A_8\ (B_8A_8)O_{32} = A(AB)O_4$** in regular chemical spelling, or
  - **$(A_{8/3}B_{16/3})\ (A_{16/3}B_{32/3})O_{32}$**

  and so on, with the atoms in the brackets occupying the respective site at random.

- A few examples (in regular chemical symbols)

  - Magnetite; **$Fe^{3+}(\ Fe^{2+}\ Fe^{3+})O_4$**
  - Spinel; **$Mg^{2+}(\ Al_2{}^{3+})O_4$**
  - Chromite; **$Fe^{3+}(Cr_2{}^{3+})O_4$**
  - Jacobsite; **$Fe^{3+}(\ Mn^{2+}\ Fe^{3+})O_4$**

- The spinel structure is also interesting because it may contain **vacancies as regular part of the crystal**. For example, if magnetite is slowly oxidized by lying around a couple of billion years, or when rocks cool, **Fe^{2+}** will turn into **Fe^{3+}** (oxidation, in chemical terms, means you take electrons away). If all **Fe^{2+}** is converted into **Fe^{3+}**, charge balance requires a net formula of **$Fe_{21,67}O_{32}$** per unit cell and this means that 2,33 sites must be vacant - we have what is called a **defect spinel**. In a way, the composition is now **$Fe_{21,67}Vac_{2,33}O_3$**; having lots of vacancies as an *integral part of the structure*.

Debye Length

- The Debye length is treated in some detail and in a simple approach in the Hyperscript "**Electronic Materials**"; the link brings you there.
  - A more involved treatment can be found in the Hyperscript "**Semiconductors**".

Vagaries in the Statistical Definition of the Entropy

- The statistical definition of the entropy appears in many forms; almost every textbook finds its own version - and all versions are equally correct. You will always find the definition

$$S = k \cdot \ln P$$

- But the meaning of ***P*** may be quite different on a first glance. Let's look at a few examples:

- "***P*** means the probability of a macrostate, where ***P*** in turn is proportional to the number of microstates accessible to the system contained within that macrostate".
  - The quote is from: *R.P. Baumann*; Modern Thermodynamics with Statistical Mechanics, p. 337.
  - Note that a probability is a number $\leq$ **1**; ***S*** thus would be always negative.
- ***P*** is the volume in **phase space** occupied by the system.
  - *Becker*, Theorie der Wärme, S. 117 (That's what I had as a student).
  - Note that this looks like a number with a dimension!
- Now some random finds without the detailed quote.
- ***P*** is the number of indistinguishable microstates belonging to one macrostate.
  - That is the definition we used in the script. Note that this is a pure, and mostly very large number.
- ***P*** is the probability for a macrostate, i.e. the number P_i of microstates belonging to a certain macrostate **i** divided by the sum over all possible P_i.
  - Note that ***P*** than is a pure number between **0** and **1**.
- What is correct? The numerical value of ***S*** obviously could be positive or negative and generally very different depending on which definition one uses.
  - The answer, of course, draws on the old fact that in classical physics (including thermodynamics) there is *no absolute scale* for energies (and entropy times temperature is a form of energy).
  - We thus can always use a ***P**** instead of ***P***, defined by $P^* = P/P_0$ with $P_0 =$ arbitrary constant factor (that does not depend on the variables of the system under consideration). All that happens is that you add a constant factor to the entropy or free energy of a system; i.e. you change the zero point of the energy scale.
  - If we replace ***P*** by ***P****, we obtain for the entropy .

$$S^* = k \cdot \ln P^* = k \cdot \ln \frac{P}{P_0} = k \cdot \ln P - k \cdot \ln P_0 = S - \text{const.}$$

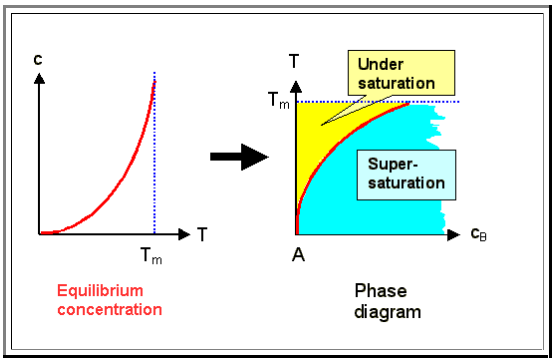
  - For the free enthalpy we then simply have $G^* = G - kT \cdot \ln P_0 = G - \text{const}$
  - Moreover, since in practice most applications contain the derivative of ***S*** with respect to some variable ***x*** of the system, constant factors will disappear, i.e..

$$\frac{\partial S^*}{\partial x} = \frac{\partial S}{\partial x}$$

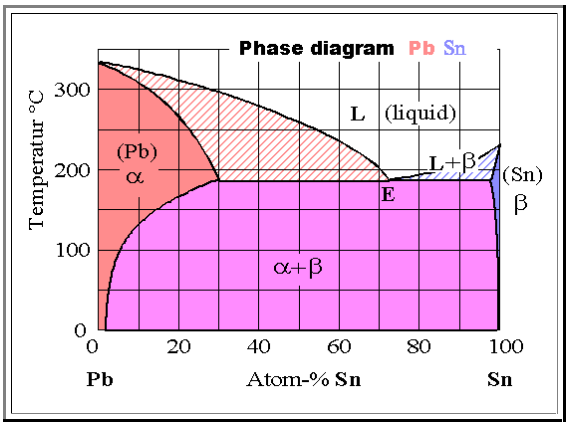
- In short, all definitions are equivalent and you don't have to worry about the additional constant factors that may appear. Feel free to use the definition that is most easily applied to the problem under consideration.
  - However, if you *like* to worry, or noticed that there was a little disclaimer above, read on in the advanced section.

Phase Diagrams

- Phase diagrams are the mainstay of materials science and technology. They may be seen as a *map* in temperature - composition space that shows the particular structure with the minimum free enthalpy at any point of the "map".
  - If we take the formula for the equilibrium concentration of self-interstitials to also describe the equilibrium concentration of extrinsic interstitial atoms if we replace the formation enthalpy by a corresponding property (lets call it *solubility enthalpy*), the resulting diagram of the equilibrium concentration over temperature can be interpreted as part of a phase diagram.
  - All we have to do is to switch the axes from the normal representation *concentration vs. temperature* to *temperature vs. concentration*:



- We now have a diagram for the composition of material **A** (the matrix) with material **B** (the extrinsic interstitial) for small concentrations of **B**.
  - The red line denotes the limit of solubility of **B** in **A**; it corresponds to the equilibrium concentration.
  - In the yellow or blue areas, the **B**-interstitials are *undersaturated* or *supersaturated*, respectively. We must expect that something new is going to happen in the supersaturated region, e.g. the precipitation of some $\mathbf{A_xB_y}$ compound, or a phase separation of **A** and **B**.
  - On the other end of the composition axis, things would be much the same, only that now **A** is the interstitial in **B**. The equilibrium line and the melting point would be different too, of course.
- The phase diagram **Pb** - **Sn** (familiar solder) provides a real example:



- Pure lead and lead with **Sn** interstitials has a **fcc** lattice; we call this the α-phase; pure **Sn** and **Sn** with **Pb** interstitials is tetragonal, we call this the β-phase.
  - In the supersaturated region something new has happened ndeed, we have an *eutectic phase separation* and a mixture of α and β.
  - In the high temperature regime, we have something new, too: Mixtures of liquid **(L)** and solid phases
- Be that as it may (and it can be much more complicated), the essential points are
  - The system always goes for the minimum free enthalpy, and this minimum could be calculated in principle following the same, albeit much more involved line of reasoning we employed for equilibrium concentrations of point defects.
  - The (experimentally determined) phase diagrams are maps of the particular minimum free enthalpy configuration out of many possible arrangements for a given composition and temperature.
  - Changing the temperature or the composition of a system thus takes us from one area in the phase diagram to another; the boundaries we have to cross give us an idea of what has to happen kinetically.

Arrhenius-Plot

- Many thermally excited reactions are described by

$$y = y_0 \exp - \frac{E_a}{kT}$$

  - With E_a = activation energy (or enthalpy) of the process, and kT with its usual meaning.
  - This equation governs not only the equilibrium concentration of point defects, but also, for example, the emission of electrons from a hot wire or the growth of bacterial cultures.
- An Arrhenius plot of this equation is simply a plot of **log y** (or **ln** y) over $1/T$ (or $1/kT$). This produces a straight line:

$$\ln y = \ln y_0 - \frac{E_a}{k_B} \cdot \frac{1}{T}$$

  - The (extrapolated) cut with the **ln y**-axis gives directly the value of the pre-exponential factor y_0, and the slope of the straight line gives the activation energy.
- An Arrhenius plot is extremely useful if data are determined experimentally. It shows at a glance if the scatter of the data points is small or large, if we have an Arrhenius relation at all (i.e. a straight line), and if we have enough data points to get unambigous values for the activation energy and the pre-exponential factor.

- In the following Java module, you can play a bit with the representations of the exponential law.
  - Shown is the function

$$c_V = c_0 \cdot \exp - \frac{H_F}{k_B}$$

  - in a direct plot and in an Arrhenius plot. You can change the values of the parameters and see what happens.

Potentials

- Mechanically, a potential $A(\underline{r})$ was the difference $A(\underline{r}_2) - A(\underline{r}_1)$ and equal to the work needed to go from $\underline{r}_1$ to $\underline{r}_2$.
  - The decisive point was that $A(\underline{r})$ did *not* depend on the particular way chosen to go from $\underline{r}_1$ to $\underline{r}_2$ as long as the acting forces $F(\underline{r})$ were given as the derivative of the potential $A(\underline{r})$ with respect to the coordinates. We have

$$F(\underline{r}) = -\,\mathrm{grad}\,[A(\underline{r})] = -\,\nabla[A(\underline{r})] = -\left(\frac{\partial A}{\partial \underline{x}}, \frac{\partial A}{\partial \underline{y}}, \frac{\partial A}{\partial \underline{z}}\right)$$

- These are the well known relations for mechanic (and electrostatic) potentials. If one knows the potential and the momentum of a massive (and charged) particle, one knows everything one can know and needs to know.
  - The history, i.e. how the particle came to its present position $\underline{r}$ with the potential $A(\underline{r})$ and momentum, is totally irrelevant.
  - Potential *and* momentum together then define the **state** of the particle.
- We may first generalize the idea of a potential by allowing **generalized coordinates**, i.e. any variables (and not just space coordinates) that describe the state of a system.
  - This allows to treat thermodynamic systems consisting of *many particles*, where individual coordinates loose significance and average values describing the system take precedence.
- Lets consider the free energy as a first example. It is a thermodynamic potential *and* at the same time a **state function**, i.e. it describes *completely* the state of systems with the generalized coordinates temperature T, volume V and particle numbers N_i. The mechanical potential by itself, in contrast, is *not* a state function (we would need the momentum, too, to describe the state of a mechanical system).
  - A change of state thus necessarily demands a change in at least one of the three generalized coordinates from the above example.
  - In formulas we write

$$F = \begin{matrix}\textbf{thermodynamic}\\ \textbf{potential}\end{matrix} = F(V, T, N_1, \ldots N_i)$$

  - The N_i denote the number of particles of kind **i**.
- *Is it that easy? Can we elevate all kinds of functions to the state of thermodynamic potentials and state functions?*
- The answer, of course, is *No!* The statement that the function $P(x_i)$ is a potential with respect to the generalized coordinates x_i is *only* true if a number of conditions are met. *For the case of equilibrium* (which requires that nothing changes and therefore that all $\partial P / \partial x_i = 0$) those requirements are:
  - The values of the generalized coordinates describe the state of the system *completely* (this means you have the right number *and* the right kinds of coordinates).
  - The value of $P(x_i)$ for a set of values of the generalized coordinates is *independent from the path chosen* to arrive at the particular state and thus from the history of the system. In formulas:

$$\Delta P = \begin{matrix}\textbf{change in } P\\ \textbf{between a state 1}\\ \textbf{and a state 2}\end{matrix} = \int_1^2 dP = \int_1^2 \frac{\partial P}{\partial x_1}\cdot dx_1 + \int_1^2 \frac{\partial P}{\partial x_2}\cdot dx_2 + \ldots := P_2 - P_1$$

  - i.e. the difference depends *only* on the starting and end state.
- This can only be true if dP is a **total differential** of P, i.e.

$$dP = \frac{\partial P}{\partial x_1}\cdot dx_1 + \frac{\partial P}{\partial x_2}\cdot dx_2 + \ldots.$$

  - i.e. the changes in the generalized coordinates describe unambiguously the total change in P.

- The numerical values of the partial derivatives of the thermodynamic potentials describe the "effort" it takes to change the potential by fiddling with the particular coordinate considered. This can be understood as a **generalized force**. If the derivative happens to be taken with respect to a *space coordinate* of the system, the *generalized* force is a *real* mechanical force - it describes the change of energy with space as it should.
  - If the derivative happens to be taken with respect to a *particle number*, however, the resulting quantity is not called, e.g., "*particle number changing force*" (which would have been a perfectly good name), but **chemical potential**, which might be a bit confusing at times.
- How do we know if a given function is a thermodynamic potential and a state function? For a given function (in differential form), it is not necessarily obvious if it is a total differential. You have to resort to *physical* or *mathematical* reasoning to find out. Lets first look at an example of physical reasoning:
- The first law of thermodynamics was defined as follows:

$$\mathbf{d}U = \mathbf{d}Q - \mathbf{d}W$$

- Are these three differential quantities total differentials and thus state function and thermodynamic potentials, or are they not?
  - *Physical* reasoning tells us that **d*U*** *must* be a total differential because **U** must be a thermodynamic state function - otherwise we can construct a **perpetuum mobile**! Lets see why:
  - If we start from some value U_1, characterized by as many variables x_i (= coordinates) as you like (e.g. pressure p_1, volume V_1, temperature T_1) and than move to a second value U_2 by adding, for example, a heat quantity Q; we have described a path **1** to move from U_1 to U_2.
  - If we return to the same values of the generalized coordinates by a different path; e.g. by now extracting some mechanical work, but have a value U_1' that is not identical to U_1, we are now in a position to construct a cycle between the states **1** and **2** as characterized by the generalized coordinates that allows us to extract work in every cycle - we have a *perpetuum mobile*.
  - So **d*U*** must be a total differential - and this requires that **d*Q*** and **d*W*** are *not* total differentials! Because it is entirely possible to go from one state of ***U*** to another one by adding different amounts of ***Q*** and ***W*** - the path described by these circumstances must not matter!
  - $\Delta Q = Q_2 - Q_1$ or $\Delta W = W_2 - W_1$ thus are not independent of the path, therefore they are *not* state functions and their differentials cannot be total differentials. This is essentially tied to the fact that the system entropy may change in these cases.
- Now lets turn to *mathematical* reasoning. Obviously, if a function $P(x_i)$ of several variables x_i is given, it is always possible to calculate the total differential $\mathbf{d}P(x_i)$. *But the reverse is not necessarily true*:
  - If ***P*** is given in differential form, it always can be written as

$$\mathbf{d}P(x,y) = g(x,y) \cdot \mathbf{d}x + \mathbf{f}(x,y) \cdot \mathbf{d}y$$

  - We used only two variables ***x*** and ***y*** for simplicities sake. The functions ***g(x,y)*** and **f(*x*,*y*)** are arbitrary functions - what ever you like is allowed.
- *If* **d*P*** is a total differential, we have *necessarily*

$$g(x,y) = \frac{\partial P}{\partial x}$$

$$\mathbf{f}(x,y) = \frac{\partial P}{\partial y}$$

  - On the other hand for *all* functions ***P(x,y)*** the following equalities must obtain

$$\frac{\partial}{\partial y}\left(\frac{\partial P}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial P}{\partial y}\right)$$

  - This requires that

$$\frac{\partial g(x,y)}{\partial y} = \frac{\partial f(x,y)}{\partial x}$$

Only if the above relation is fulfilled, is **dP(*x,y*) = g(*x,y*)dx + f(*x,y*)dy** a total differential. So we have a simple checking procedure for a given differential function (a mathematical rule) to find out if it is indeed a total differential.

Boltzmann's Constant and Gas Constant

- We have repeatedly stressed the fact that whenever you encounter Boltzmann's constant **k** we deal with the particle unit "atom" or "molecule", while whenever we encounter the gas constant **R**, we deal with the the unit "mol". Here we will quickly survey the connection.
  - First we have the general law for ideal gases with volume V, pressure p and temperature T

$$p \cdot V = \text{const} \cdot T$$

  - This was first an empirical law that later became fully understood by statistical thermodynamics.
- The next step is to realize that if you increase the volume while keeping everything else constant, the "const" in the law must increase in the same proportion. This leads to the much more universal formulation that is generally used:

$$p \cdot V = n \cdot R \cdot T$$

  - With n = quantity of the gas, and **R** = gas constant with a value depending on how you measure **n**.
- This would still leave room for **R** being different for different kind of gases. **Avogadro** enters, proposing that identical volumina of gases under identical pressure and temperature contain identical numbers of particles.
- This permits to define **R** for all (ideal) gases and to measure it. We find

$$R = \frac{p \cdot V}{n_{mol} \cdot T} = 8.32441\ J \cdot K^{-1} \cdot mol^{-1}$$

  - *if* me measure the quantity n of the gas in *mols*. º
  - One **mol** of a substance, *per definition*, contains just as many particles, objects, or building blocks of that substance (i.e. atoms, molecules, electrons, vacancies, ...), as there are carbon atoms in **1 g** of ^{12}C which gives

$$1\ mol = 6.022 \cdot 10^{23}$$

  - **Avogadros constant** then automatically is

$$N_A = 6.022 \cdot 10^{23}\ mol^{-1}$$

  - i.e. we have $6.022 \cdot 10^{23}$ particles *per mol* of a substance.
- If we set $n = 1$ we have for the mol-volume V_m, i.e. for the volume that **1** mol of a gas occupies

$$V_m = \frac{n \cdot R \cdot T}{p} = 22.414\ \frac{l}{mol}$$

  - This is valid for for "old" standard conditions (p = **1013 mbar = 101 325 Pa**, and T = **0°C**).
  - For the "new" standard conditions (p = **100 000 Pa,** T = **298.15 K**) we have V_m = **24.789 l/mol**
  - *Why the international standards and units of measurements must change all the time is beyond me, but that's the way it is. I have suffered through **4** changes in the units for pressure by now, not to mention the big pain caused by the fact that the Americans normally don't care and still stick to **psi**.*
- If we now measure substance quantities not per mol, but per particle, we must divide **R** by Avogadros constant N_A and obtain

$$p \cdot V = n_{part} \cdot \frac{R}{N_A} \cdot T = n_{part} \cdot kT$$

- and n_{part} is now the *number of particles* in V.
- For R/N_A **=: k = Boltzmann's constant** we obtain

$$k = \frac{8.32441}{6.022 \cdot 10^{23}} \; J \cdot K^{-1} = 8.616 \cdot 10^{-5} eV \cdot K^{-1}$$

Fine, we can see that as a *definition* of Boltzmann's constant **k**. But now we have two questions:

- **1.** Why is the **k** from the gas law the same number as in the famous entropy equation $S = k \cdot \ln P$?
- *Not obvious* - and not exactly easy to prove. Essentially, you have to unleash the full power of statistical thermodynamics to show that both **k's** are identical. So either grab your thermodynamic textbook, or believe your professor at this point.
- **2**. Is there a way to calculate the numerical value of **k** from some more fundamental constants? Well, as far as I know, it *cannot be done*. So **k** is a basic *constant of nature*, in the same league as other fundamental constants of nature, like the speed of light, the gravitational constant, or the elementary charge.

Finally something to make things really complicated:

- Changing from *mols* to *particle numbers* or *densities*, changes the precise formulation of the mass action law. Consult the link for details.

A short reminder of basic Thermodynamics with links to other Hyperscripts and to more detailed considerations.

Link to the newest version of this page without insets

Internal Energy, Enthalpy, Entropy, Free Energy, and Free Enthalpy

A Thermodynamics Primer

Basics

General Remarks and State Functions

- Lets face it: Thermodymanics is not easy! *It is not possible to learn it by just reading through this module.*
  - However, if you fought your way through thermodynamics proper at least once, and thus are able to look at it from a distance without getting totally confused by the "details" (which you don't have to know anymore, but must be able to understand when they come up), it's not so difficult either.
  - It gets even easier by restricting ourselves to solids which means that most of the time we don't have to worry about the *pressure* ***p*** anymore - it is simply constant.
- In this primer we will review the most important issues necessary for understanding defects, including defects in semiconductors. In order to stay simple, we must "cut corners". This means:
  - We will usually not show the functional relationships by showing the variables. We thus simply write ***G*** for the free enthalpy, and not $G(T, S, n_i)$, showing that ***G*** is a function of the temperature ***T***, the entropy ***S*** and the particle numbers n_i.
  - In the same spirit, we will omit the indexes showing what stays constant for partial derivations, i.e. we write for the chemical potential μ_i of the particle sort **i** the simple form

$$\frac{\partial G}{\partial n_i} = \mu_i$$

  - In full splendor it should be

$$\left.\frac{\partial G(T, S, n_i)}{\partial n_i}\right|_{S,\, T,\, n_{i \neq j}} = \mu_i(T, S, n_{i \neq j})$$

  - We also are sloppy about standards. n_i may refer to particle numbers *or* concentrations; in the latter case **particles/cm^3** or **mol/cm^3** - you must know what is meant from the context. You are also supposed to know that if Boltzmanns constant **k** comes up in an equation, we are working with properties per particle, whereas the gas constant **R** signifies properties per mol.
- This, admittedly, is dangerous. But multi-indexed quantities are confusing (and not easily written in HTML, anyway)! Lets stay simple and refer to complications whenever they come up.
- If we restrict ourselves to crystals, it is rather easy to consider the concepts behind the all-important thermodynamic quantities **Internal Energy**, **Enthalpy**, **Entropy**, **Free Energy** and **Free Enthalpy**. We start with the *internal energy* ***U*** of a crystal.
  - Neglecting *external* energies (e.g. the gravitational potential) and *internal energies that never change* (e.g. the energy of the inner electrons), we are essentially left with the internal energy ***U*** being contained in the *vibrations of the crystal atoms* (or molecules), which express themselves in the *temperature **T** of the system* according to

$$U = \tfrac{1}{2} \cdot f \cdot kT$$

  - With ***U*** = average energy per atom, ***f*** = degree of freedoms for "investing" energy in an atom (**$f = 6$** for crystals; **3** for the kinetic energy in v_x, v_y, v_z, and **3** for the potential energy at (x, y, z)); **k =** Boltzmanns constant and ***T*** = (absolute) temperature
- The *(macro)state of the system* is thus given by the number of atoms ***N***, the pressure ***p*** and the temperature ***T***. Knowing these numbers is all there is to know about the system *on a macroscopic base*.

- We can change the state of the system by adding or removing heat ***Q***, putting mechanical work ***W*** into the system or taking it out, and by changing the number of atoms (or more generally, particles) by some ΔN.
- Since at this point we keep the number of atoms in our crystal constant, we only have to consider ***Q*** and ***W*** if we change the state. The following basic equations (a formulation of the **1st law of thermodynamics**) *(german link)* holds

$$dU = dQ - dW$$

- With the changes written in differential form. Note that the regular **"*d* "** is not the sign for (partial) derivatives (that would be ∂) but for "delta". **d*U***, e.g., stands for **total change of *U***. Note that *sometimes* the **d** indicates a total differential (e.g. in the case of **d*U***), sometimes it does not (e.g. in the case of **dQ** or **d*W***).
- Any mechanical work must change the volume (something must move); for the normal conditions encountered with crystals where the pressure stays constant it can always be expressed as

$$dW = p \cdot dV$$

This term ***p*d*V*** is cumbersome as long as only situations involving crystals under constant pressure are considered. We thus introduce a *new state function* called enthalpy ***H*** and define it as

$$H = U + pV$$

- If we again change the state of the system by adding or subtracting heat ***Q*** and mechanical work ***W***, we now obtain for the total *change in enthalpy* **d*H***

$$dH = dU + pdV + Vdp$$

- With ***V*d*p* = 0**, because the pressure ***p*** is constant, and **d*U* = d*Q* – *p*d*V*** we obtain

$$dH = dQ$$

- This is a simple relation always best suited for systems under constant pressure and also clarifying why we tend to think of enthalpy as **heat**.
- **d*H*** is a measure of of the energy needed to form a substance in a given state, it is occasionally also called the **heat of formation** (always refering to the difference between two states).
- Of course, not much happens if the substance is just heated a bit but does not change its chemical nature - lets say we look at a mixture of H_2 and O_2 which we heat up a bit. All the fun comes from chemical reactions (or phase changes) - in our example it would be the formation of H_2O in a somewhat violent fashion.

It was thought that the sign of **d*H*** would indicate if a reaction should or should not occur. A negative sign would mean that the reaction would transfer energy to the surroundings and thus could easily happen, whereas a positive sign would tell us that energy would have to be pumped into the system - nothing would happen by itself.

- It's not that simple! While this point of view was true enough for relatively large **d*H*** (lets say **> 100 kcal/mol**), the criterion often does not work for smaller changes of **d*H***.
- The reason, of course, is that we neglected the change of the **entropy *S*** of the system, **d*S***, that occurs parallel to **d*H***.

Purely mechanical systems (consisting of non-interacting mass points) would be in equilibrium for the lowest possible internal energy, i.e. for a minimum in their potential energy and no movement - just lying still at the lowest possible point. *But thermodynamic systems* consisting of many interacting particles and some externally fixed condition (e.g. a constant temperature), are in equilibrium if the *best possible balance between a **small energy and a large entropy** is achieved*!

- We just take that as an article of faith (or law of nature) at this point.
- Often, both quantities are opposed to each other: High entropies mean high energies and vice verse. The entropy part becomes more important at high temperatures, and the thermodynamic potential which has to be minimized for systems under constant pressure, is the **free enthalpy *G*** (also called **Gibbs energy**). It is defined as

$$G = H - T \cdot S$$

- With ***S*** = **entropy = dQ_{rev}/*T*** in classical thermodynamics (the suffix "rev" refers to reversible processes).

- If you have a system with constant volume (and variable pressure), the best suited state function is the **free energy *F*** (also called **Helmholtz energy**). It is defined as

$$F = U - T \cdot S$$

- Before turning to the entropy, a word to the choice of state functions. We now already have four: ***U***, ***H***, ***G***, ***F*** - but for a given system, there is only *one* state. Two things are important in this context:
  - State functions, by definition, must describe the state of a system no matter how this state developed - they must, in other words, meet all the requirements for potentials and thus are **thermodynamic potentials**. We have not proved if this is the case for ***U***, ***H***, ***G***, ***F*** - turn to the potential module for some input to this question - but they really are potentials.
  - *Any* state function or thermodynamic potential can be used to describe *any* system (always for equilibrium, of course), but for a given system some are more convenient than others. The most convenient (and thus important) one for crystals (usually under constant pressure) is the free enthalpy.

Entropy - Statistical Consideration

- The key question is:

What is entropy?

  - There is a classical answer, but here we only use the statistical definition where *entropy is the measure of the "probability" **w** of a given macrostate*, or, essentially the same thing, *the number **P** of microstates possible for the given macrostate*.
  - Not too helpful: What is a **microstate** or a **macrostate**? Or the **probability of a macrostate**?
  - Well, *any* particular arrangement of atoms (or more generally, particles) where we look only on *average quantities* is a *macrostate*, while any *individual arrangement* defining the properties (e.g. location and momentary velocity) of all the particles for a given macrostate is a *microstate*.
  - In other words, and somewhat simplified: For a microstate it matters what individual particles do, for the macrostate it does not.
  - The difference between microstates and macrostates is best illustrated for for a gas in a closed container: We can define many possible macrostates, e.g.

    - **1.** All molecules are in the left half of the container.
    - **2. 70 %** of the molcules are in the left half of the container, **30 %** in the right half.
    - **3.** Equal (average) distribution of the molecules.

    and all these macrostates (plus many more) could have exactly the same internal energy ***U*** (or ***H***).
  - However, the *probability* of experimentally finding one or the other of those macrostates is very different. The probabilities of the macrostates **1.** and **2.** are certainly much much smaller than the probability of macrostate No. **3.**
  - For all the possible macrostates, the *state function* tells us which one will be realized (= is most probable) in thermal equilibrium.
- How do we calculate the probability of a macrostate? Lets see:
  - For every possible macrostate we can think of, there are many microstates to realize it. Its exactly like playing dice: Lets assume you have **3** dice. A macrostate would be some possible number you may throw; e.g. **9**. The corresponding microstates are the possible combinations of the individual dice. For throwing **9** we have

Dice	1. Poss.	2. Poss.	3. Poss	....
1	1	1	1	
2	1	2	3	
3	7	6	5	

  - - and so on. You get the picture.

- The *probability* for such a macrostate would be the number of microstates divided by the number of all possible combinations of the dice (which is a constant). We can see off-hand that the macrostates **"3"** and **"18"** are the most unlikely ones, having only one microstate at their disposal, while **9**, **10**, or **12** are more likely to occur.
  - Now we know what *the number of possible ways to generate the same macrostate* means and why the "probability" ***w*** of a given macrostate is *"almost"* the same thing.
- An example just as easy as playing dice, comes from our friend, the vacancy. We simply ask: How many ways ***P*** (= *microstates*) are there to arrange ***n*** vacancies (= *macrostate*) in a crystal of ***N*** atoms?
  - When we figure that out, we can use the equilibrium condition to select the most likely macrostate and this gives us the number of vacancies in equilibrium.
- The fundamental point now is that just knowing the internal energy ***U*** of a system with a constant volume and temperature is not good enough to tell us what the equilibrium configuration will be because *we* could think of many macrostates with the same ***U*** (and mother nature, to be sure, can come up with lots more).
  - That's why just minimizing ***U*** (or ***H***) is not good enough, we have to minimize ***F = U – TS*** or ***G = H – TS*** to find the equilibrium configuration of the system, and for that we have to know the entropy, because we now can interprete these formulas:
  - Of all the many *macrostates* possible for a given ***U*** (or ***H***) the one with the *largest entropy* at the given temperature will be the one that the system will adopt
- Obviously, we need to be able to calculate the entropy of a certain macrostate and this is done by employing the statistical definition of the entropy ***S***, the famous **Boltzmann** **entropy equation** *(german link)*:

$$S = k \cdot \ln w$$

or

$$S = k \cdot \ln P$$

  With ***w*** = probability of a macrostate and ***P*** = number of microstates for a macrostate.
- If you feel that the ambiguity with respect to taking ***w*** or ***P*** is a bit puzzling - that's because it is! You should consult the link to see that at least it is nothing to worry about. Whatever you chose to work with, the results you will get in the end will be the same.
  - Entropy ***S***, by the way, is *not a state function* (***TS*** would be one).
- We used the statistical definition of entropy and the minimization of the free enthalpy in chapter 2.1; and in an exercise module it can be seen in detail how to apply it to derive the formula for the vacancy concentration

Back to the "Guided Tour" home page

Combinatorics

- Here we just look at the different ways to generate **combinations** or **variations** of "things" (= *elements*) belonging to a certain set of things.
  - The *set of "things"* could be the numbers **{0, 1, 2, ..., 9}**; the letters **{a, b, c, ..., f}** of the alphabet; the atoms of a crystal; the people on this earth, in Europe, or in your hometown - you get the drift. We generally assume that the complete set has ***N*** such elements.
  - We then define *subsets* **{*k*}** that contain ***k*** elements from the set **{0...9}**; eight letters, a certain number ***n*** of atoms, and ask what kind of combinations or variations are possible between ***N*** and ***k***.
- Note that we do *not* ask what you can do with ***k*** elements after you made a choice.
  - To make that clear: If we have, for example, ***N* = {0, 1, 2, ..., 9}**, and ***k* = 3**, we may ask: How many possibilites are there to pick three members of ***N***? We do *not* ask: How many different numbers can I form with the subset **{2,4,5}**?
  - That seems to be a good question, so why don't we allow it? Because the set **{*k*} = {2,4,5}** has no relation anymore with **{*N*}**. How many numbers you can form with the integers **2,4,5** is completely independent of **{*N*}**; so it is not an eligible question if want to look at relations between **{*N*}** and **{*k*}**.
- This is a bit abstract, so let's look at examples:
  - For the first example we may ask:
    - How many three digit numbers (or subsets) can we form with the elements of the set **{0,1,2,....,9}** allowing *everything* (e.g. that the number starts with "**0**", e.g. **043**, and that we may have identical elements, e.g. **{*k*} = {3,3,3}** is allowed)
    - How many numbers with five digits can we form, but allowing only *distinguishable* elements? That means that, e.g., neither ***k* = {3,3,3,3,3}** is allowed, nor the number **12343**.
    - How many "numbers" with ***k*** digits can we form, allowing only *distinguishable* elements and counting all different arrangements of the same elements as identical (i.e. **123**; **231**; **312**, ...are seen as one "number" or arrangement.
- It is obvious: Even for the most simple examples, there is no end of questions you can ask concerning possible arrangements of your elements.
  - Some answers to possible questions are rather obvious, some certainly are not. For some, you might feel that you would find the answer given enough time; some you might feel are hopeless - just for you, or possibly for everybody?
  - Moreover, for some answers you have a *feeling* or some rough idea of what the result could be. It's just clear that all problems involving three digits have less than **1.000** possibilities, and that with more restrictions the number of possibilities will decrease. For other problems, however, you may not have the faintest idea of what the result might be. That is a big problem and makes combinatorics often very abstract.
- How to be systematic about this? That is an easy question: Study *combinatorics* - a mathematical discipline - for quite some time and you will find out.
  - In particular you will find out that there is a small number of *standard cases* that include many of the typical questions we posed above, and that there are standard formulas for the answers. Let's summarize these standard cases in what follows.
- Quite generally, we look at a situation where we have ***N*** elements and ask for the number of arrangements we can produce with ***k*** of those elements.
  - Some Examples:
    - The elements are the natural numbers **{0, 1, 2, ..., 9}**; i.e. ***N* = 10**. With ***k* = 3** we now ask how many *numbers* we can form with **3** of those elements.
    - The elements are two different things (e.g. ♣ and ♥, yes and no, place occupied, place free, ...) How many different strings (or other arrangements) consisting of ***k* = 6** elements can you form (e.g. ♣♥♣♣♥; ♥♥♥♣♥♣, and so on)?
    - The elements are ***N*** coins all lined up and with face up. How many *different strings* can you form if you flip ***k*** coins over?
- The questions we ask, however, are not yet specific enough to elicite a definite answer. We have to construct **2 × 2 = 4** general cases or groups of questions.
  - *First* we have to distinguish between two basic possibilities of selecting elements for the combinatorial task:

    1. We only allow *different* elements. We pick, e.g. **2** or **9** of the **10** given elements **0, 1, 2, ...9**; or generally ***k*** *different* elements. Obviously ***k*** ≤ **10** applies. For ***k* = 3**, we may thus pick **{1,2,3}**, or **{0, 5,7}**, but *not* **{1,1,2}** or **{3,3,5}**. However, it just means that you can pick a given element only once. If we look at the set **{N} = {1,1,1,2,3,4}** and have **k = 4**, we may select the sets **{1,1,1,3}**, or **{1,1,2,4}**, because **{N}**

contains three "**1's**", but not, e.g., **{1,1,1,1}**. Of course, it is a bit confusing that this case includes subsets where the elements *look* identical, even so they are not, according to the definition we used.

2. We allow *identical* elements. Again we pick ***k*** elements, but we may pick any element as often as we like, at most, of course, ***k*** times. If we work with **{N} = {1,2,3, ... ,9}** and three elements, we now might use **{1,1,1}**, **{1,1,2}**, **{1,2,2}**, **{1,2,3}**, while only **{1,2,3}** would have been allowed in the case of *different* elements from above

Second, we have to distinguish between possibilities of **arranging** the elements. An *arrangement* in this sense, simply speaking, can be anything that allows to visualize the combinations we make with the elements selected - e.g. a string as shown above. We than have two basic possibilities:

1. *Different* arrangements of the same elements count as *different* combinations/variations. **(1,3,2)** thus is a string different from **(3,1,2)** if we work with different elements from the **{0,1,2,3,...,9}** set. Likewise, **(1,1,3)** is a string different from **(1,3,1)** if identical elements are allowed.
2. *Different* arrangements of the same elements do *not* count as different combinations/variations. **(1,2,3)**, **(3,1,2)**, **(2,1,3)**, **(2,3,1)**, and so on, then would all count as *one* case or string. Note that it does not matter, if the arrangements are *really* indistinguishable or not, but only if what they *encode* is indistinguishable. For example, the string **123**, interpreted as the *number* hundred-twenty-three, is certainly distinguishable form **312**, but both strings would be indistinguishable arrangements if. e.g., interpreted as the the sequence of arranging electrons (**132** = take an electron from the first atom, then one from the third and finally one from the second and put them "in a box").

Sticking to *natural numbers* as elements of the set **{N}** for examples, we now can produce the following table for the four basic cases:

Case Distinction			
We must select *different* elements		We may select *identical* elements.	
Different *arrangements* of the same elements count. ("Distinguishable arrangements")	Different *arrangements* of the same elements do *not* count ("Indistinguishable arrangements")	Different *arrangements* of the same elements count.	Different *arrangements* of the same elements do *not* count.
We ask for the number of possible **Variations** $V^D(k, N)$	We ask for the number of possible **Combinations** $C^D(k, N)$	We ask for the number of possible **Variations** $V^I(k, N)$	We ask for the number of possible **Combinations** $C^I(k, N)$
$C^D(k, N) = \frac{N!}{(N-k)!} = \binom{N}{k} \cdot k!$	$C^I(k, N) = \frac{N!}{(N-k)! \cdot k!} = \binom{N}{k}$	$V^D(k, N) = N^k$	$V^I(k, N) = \frac{(N+k-1)!}{(N-1)! \cdot k!} = \binom{N+k-1}{k}$
Examples			
N **= {1,3,4,5}** ***k*** **= 3** **All 3-digit numbers with different elements** **134, 143, 135, 153, 145, ...** $C^D(k, N) = 4!/1! = 24$	***N*** **= {1,3,4,5}** ***k*** **= 3** **3-digit numbers with different elements and only one combination** **134, 135, 145, 345** $C^D(k, N) = 24/k! = 24/6 = 4$	***N*** **= {0,3, ... ,9)** ***k*** **= 3** **All 3-digit numbers** **000, 001, ... , 455, ... , 999** $V^D(k, N) = 10^3 = 1000$	***N*** **= {1,3,4,}** ***k*** **= 2** **All 2-digit numbers with only one combination** **11, 12, 22, 13, 23, 33** $C^D(k, N) = 4!/2! \cdot 2! = 6$

Since the fraction marked in red comes up all the time in combinatorics, it has been given its own symbol and name.

We define the **binomial coefficient** of ***N*** and ***k*** as

$$\binom{N}{k} = \text{Binomial coefficient} = \frac{N!}{(N-k)! \cdot k!}$$

Yes - it is a bit mind boggling. But it is not quite as bad as it appears.

- The third column gives an obvious result. How many three digit numbers can you produce if you have **0 - 9** and every possible combination is allowed (i.e,. **001 = 1** etc.) and counted. Yes - all numbers from **000, 001, 002, ..., 998, 999** - makes exactly **1000** combinations, or $C^D(k, N) = 10^3$ as the formula asserts.

- Always ask yourself: Am I considering a *variation* (*all* possible arrangement counts) or a *combination* ("indistinguishable" [1)] arrangements don't count separately)?
  - Look at it from the *practical* point of view, not from the *formal* one, and you will get into the right direction without too much trouble.
  - The rest you have to take on faith, or you really must apply yourself to combinatorics.
- All more complicated questions not yet contained in the cases above - e.g. we do not allow the element "**0**" as the first digit, we allow one element to be picked k_1 times, a second one k_2 times and so on, may be constructed by various combinations of the **4** cases (and note that I don't say "*easily* constructed").

Arrangement of Vacancies

- OK. For the example given the cases may be halfway transparent. But how about the *arrangement of vacancies in a crystal*? What are the elements of this *combinatorial* problem, and what is k?
  - The elements obviously are the N atoms of the crystal. The subset k equally obviously selects $k = n_V$ **= number of vancanies** of these elements.
  - This is exactly the "confusing" case mentioned above: All elements in $\{N\}$ look the same; nevertheless it makes a difference if I allow identical or different elements for $\{k\}$. We can make the situation a bit more transparent if we *number* the atoms in our thoughts.
- Now what exactly is the question to ask? There are often many ways in stating the same problem, but one way might be better than others in order to see the structure of the problem.
  - We could ask for example:

    - How many ways do we have to arrange n_V vacancies in a crystal with N atoms? That's the question, of course, but it just does not go directly with the math demonstrated above.
    - How many digital numbers can we produce with $N - n_V$ "**1's**" and n_V zeros? Here we simply count the vacancies as zero's. Good question, but still not too clear with respect to the cases above.
    - We have N *numbered* atoms. How many possibilities do we have to select n_V *different* elements? Moreover, we don't care about the arrangement of the atoms taken out, all "numbers" we could produce with the numbered atoms we have taken out counts as *one* arrangement.
  - It is clear now that we have to take the "*different elements*" and "*different arrangements of the same elements do not count*" case - which indeed gives us the correct formula that we derived from scratch in exercise 2.1-4

---

1) There is a certain paradoxon here: In order to explain *in words* that certain arrangements are *indistinguisghable*, we have to list them separately, i.e. we distinguish them. But that is not a real problem, just a problem with words.

Stirlings Formula

- Stirlings formula is an indispensable tool for all combinatorial and statistical problems because it allows to deal with **factorials**, i.e. expressions based on the definition **1 · 2 · 3 · 4 · 5 · .... · N := N!**
- It exists in several modifications; all of which are approximations with different degrees of precision. It is relatively easy to deduce its more simple version. We have

$$\ln x! = \ln 1 + \ln 2 + \ln 3 + ... + \ln x = \sum_{1}^{x} \ln y$$

  - With **y =** positive integer running from **1 to x**
  - For large **y** we may replace the sum by an integration in a good approximation and obtain

$$\sum_{1}^{x} \ln y \approx \int_{1}^{x} (\ln y) \cdot dy$$

  - With $\int (\ln y) \cdot dy = y \cdot \ln y - y$, we obtain

$$\ln x! \approx x \cdot \ln x - x + 1$$

- This is the simple version of Stirlings formula. it can be even more simplified for large ***x*** because then $x + 1 << x \cdot \ln x$; and the most simple version, perfectly sufficient for many cases, results:

$$\ln x! \approx x \cdot \ln x$$

- **However!!** We not only produced a simple approximation for ***x*!**, but turned a *discrete* function having values for integers only, into a *continuous* function, giving numbers for something like **3,141!** - which may or may not make sense.
  - This may have dire consequences. Using the Strirling formula you may, e.g., move from *absolute probabilities* (always a number between **0** and **1**) to *probability densities* (any positive number) without being aware of it.
- Finally, an even better approximation exists (the prove of which would take some **20** pages) and which is already rather good for small values of ***x***, say $x > 10$:

$$x! \approx (2\pi)^{1/2} \cdot x^{(x + ½)} \cdot e^{-x}$$

Multiple Choice Test zu

2.1.1 Simple Vacancies and Interstitials

Start Multiple Choice

Multiple Choice Test zu

2.1.2 Frenkel Defects

Start Multiple Choice

Multiple Choice Test zu

2.1.3 Schottky Defects

Start Multiple Choice

Multiple Choice Test zu

2.2.1 Impurity Atoms and Point Defects

Start Multiple Choice

Multiple Choice Test zu

2.2.2 Local and Global Equilibrium

Start Multiple Choice

Exercise 2.1-1

Find the Mistake

Below, two pages from the book "Defects and Defect Processes (**Hayes and Stoneham**) are shown.

- Can you find the mistake?
- Why is the result correct anyway?

Illustration

Link to the Solution

3.1.1 Schottky and Frenkel Defects

Consider first a crystal of N atomic sites that, for simplicity, we may imagine to be a rare-gas crystal. We suppose there are no complications due to electronic or orientational degeneracy. Let g be the work necessary to create a single vacancy at constant temperature and pressure. We shall assume the vacancy concentration to be so low that defect–defect interactions can be ignored and hence that g is independent of the vacancy content. The free energy of the defective crystal contains three principal terms:

(i) A term G_0 corresponding to the perfect crystal.
(ii) A term $G_1 = ng$ coming from the work done in creating n vacancies.
(iii) A term $G_2 = -TS_{\text{config}}$ from the configurational entropy.

The concentration of vacancies at equilibrium is determined by the extremal condition

$$\frac{\partial G}{\partial n} = \frac{\partial}{\partial n}(G_0 + G_1 + G_2) = 0 \qquad (3.2)$$

In our case G_0 is independent of n by definition, and $(\partial G_1/\partial n)$ is simply g. Equation (3.2) reduces to

$$g - T\frac{\partial}{\partial n}(S_{\text{config}}) = 0 \qquad (3.3)$$

The configurational entropy term is obtained using a standard combinatorial method. The vacant lattice points can be arranged in P_L ways,

$$P_L = N!/[(N-n)!] \qquad (3.4)$$

The configurational entropy is given by $k_B \ln P_L$, so that we have

$$S_{\text{config}} = k_B \ln\{N!/[(N-n)!n!]\} \qquad (3.5)$$

We need $T\partial S_{\text{config}}/\partial n$, and this is found using Stirling's theorem that $\ln(x!)$ tends to $x \ln x$ at large x. If N, $(N-n)$ and n are all large, we find

$$T\frac{\partial S_{\text{config}}}{\partial n} = k_B T \ln\left(\frac{N-n}{n}\right) \tag{3.6}$$

If the defect concentration $f \equiv n/N$ is small, the right-hand side of Equation (3.6) reduces to $-k_B T \ln f$. Collecting together the terms in $\partial G/\partial n$ we find

$$g + k_B T \ln f = 0, \tag{3.7}$$

giving the fraction f of vacant sites as

$$f = \exp(-g/k_B T) \tag{3.8}$$

The fractional vacancy concentration depends exponentially on the Schottky formation energy g, and increases exponentially with increasing temperature.

Even the simple example discussed above raises some general points. The first concerns the reference energy state. Although for the simple Schottky defect this is the perfect crystal, there may be some choice in other cases. As an example we consider the energy of solution of an interstitial impurity, for example, oxygen, in a crystal. The reference state will be different when one considers the equilibrium of the crystal with a gas phase X_2 and when one considers the equilibrium of the same crystal with a liquid compound MX. The two cases can be related, one to the other, but there is opportunity for confusion in analyzing published data on energies of solution.

A second point that can cause confusion arises in connection with Schottky disorder. The number of defects present in equilibrium will depend to a large extent on the energy required to create them. If one creates a vacancy by moving a bulk atom to a surface site, the energy will depend on the precise surface site, and hence the equilibrium would at first sight seem to depend on details such as the crystal habit. The resolution of this problem can be given in several forms. One way is to define Schottky disorder for an infinite crystal with no free surfaces or sinks. The vacancy is introduced by rearranging N atoms on $N+1$ sites, with a volume change if the process is to occur at constant pressure. It is the formation energy so defined that determines the equilibrium concentration of defects, in principle.

In many simple estimates of formation energies the experimental cohesive energy or the sublimation energy is required. It is here that problems could arise if the experiment measured some special property,

Exercise 2.1-2

Do the Math for the Formation Entropy

Illustration

Do the Math needed for going from the first to the second formula:

- **First formula:**

$$F = -kT \cdot \ln Z = kT \cdot \sum_i \left(\frac{h\omega_i}{4\pi \cdot kT} + \ln\left(1 - \exp - \frac{h\omega_i}{2\pi \cdot kT}\right)\right)$$

- **Second formula:**

$$S = k \cdot \sum_i \left(-\ln\left(1 - \exp \frac{h\omega_i}{2\pi \cdot kT}\right) + \frac{\frac{h\omega_i}{2\pi \cdot kT}}{\exp\left(\frac{h\omega_i}{2\pi \cdot kT} - 1\right)}\right)$$

Now, using the approximations referred to, derive the **Final formula**

$$S_F = k \cdot \sum_i \ln \frac{\omega_i}{\omega'_i}$$

Discuss the quality of the approximations

- (Hint: Use some real numbers or order of magnitudes for e.g. Debye temperatures, vacancy concentrations etc.)

Link to the Solution

Exercise 2.1-3

Calculate Formation Entropies

Illustration

Calculate the formation entropy for a simple cubic crystal.

- Assume that there are two kinds of springs holding the atoms together: Springs between next neighbors, and springs between second next neighbors.
- Look at the resonance frequency as a function of the spring constant ***D*** (make some assumptions about the spring constant D_2 of the second-next-neighbor spring in terms of the spring constant D_1 of the next neighbors).

Calculate the formation entropy first by only considering the D_1 springs; than consider the D_2 springs, too.

Link to the Solution

Exercise 2.1-4

Derive the Formula for the Vacancy Equilibrium Concentration

Illustration

Start from the formulas given, use the approximations described, and show that

$$c_V = \text{concentration of vacancies} = \frac{S_F}{k} \cdot \exp -\frac{H_F}{kT}$$

Would more advanced versions of Stirlings formula get better results?

Link to the Solution

Exercise 2.1-5

Do the Math for Mixed Point Defects

Illustration

Solve the system of the three equations given below and show that the result given is correct

$$c_V(C) \cdot c_i(C) = \frac{N'}{N} \cdot \exp - \frac{H_{FP}}{kT}$$

$$c_V(A) \cdot c_V(C) = \frac{N'}{N} \cdot \exp - \frac{H_S}{kT}$$

$$c_V(C) = c_V(A) + c_i(C)$$

What are the conditions for the limiting cases of pure Frenkel or Schottky defects?

Link to the Solution

Exercise 2.1-6

Enthalpy difference for the limiting cases of Schottky or Frenkel Defects

Illustration

The equations for the concentration of the three point defects contain in parts the differnce of the formation enthalpies of Schottky or Frenkel defects.

- Calculate the ratio of the concentration of Schottky to Frenkel defects as a function of this enthalpy difference.
- Discuss the result. Show in in particular, how large the difference must be if **90%** or **99%**, resp., of the defects are to be of *one* kind.

Link to the Solution

Exercise 2.1-7

Quick Questions to

2.1. Intrinsic Point Defects and Equilibrium

2.1.1 Simple Vacancies and Interstitials

Here are some quick questions:

- The *answers* are sometimes (and possibly only indirectly) contained in the links.

- Write down the free enthalpy of a crystal with ***N*** atoms for
  - **1** vacancy
  - ***n*** vacancies

  in terms of the relevant quantities "***G***", "***H***", "***S***" for a single vacancy.
- The binomial coefficient as defined below gives the number of possibilities that one has in selecting ***n*** *different elements* from a given set of ***N*** elements with the condition that *different arrangements* of the same elements do *not* count ("Indistinguishable arrangements").
  Example: Set = **{1,2,3,4,5,6,7,8,9}**; i.e. **N = 9**, Selected elements = **{1,5,8}**, i.e. ***n* = 3** = equivalent to **{5,1, 8}**; **{1,8,5}**, ...
  How do you have to phrase the question for the vacancy arrangement so that the answer is immediately obvious?

$$\binom{N}{k} = \text{Binomial coefficient} = \frac{N!}{(N-k)! \cdot k!}$$

- Write the entropy of mixing with the binomial coefficient from above and then express it with the help of the Stirling equation.
- What is the chemical potential μ, and what must be valid for ***n*** vacancies *in equilibrium*?
- In the famous Boltzmann entropy equation $S = k_B \cdot \ln P$ the number ***P*** could be a probability for a state; i.e. a number between **0** and **1**, or the number of arrangements, i.e. a huge number. Explain why this doesn't matter.
- How is the formation entropy defined formally? Why and how is it connected to a single vacancy? What does it describe or measure in practical terms? How large is it (order of magnitude)?
- Give some numbers for formation enthalpies of vacancies in common crystals.
  Give some numbers for formation enthalpies of self-interstitials in common crystals.
  Draw some conclusion.
- What kind of difference $\Delta H = H_i - H_V$ of the formation enthalpy of vacancy and self interstitial produces a concentration ratio of $n_V/n_i > 10$? Do a quick and dirty estimation!
- Write down the basic equation for the concentration of single vacancies. Produce a graph of this equation with some numbers on the axes:
  - in a direct representation.
  - in an Arrhenius representation.
- Describe how you can tell from the n_V curves for *two* crystals in *one* Arrhenius plot which one has the larger H_V and S_V.
- What is the relation between the Boltzmann distribution (or Boltzmann factor) and the vacancy concentration? How does on have to pose the question for the vacancy concentration to obtain the result directly from this?
- Write down the free enthalpy n_{2V} and the resulting concentration c_{2V} for *divacancies* in a crystal with ***N*** atoms.
- Generalize for n_{xV} and c_{xV}, i.e. for *multiple vacancy clusters*. Use two different approaches for this.
- Discuss the relative concentration of c_{xV}/c_{1V} in *equilibrium.*
- Describe the use of the mass action law for obtaining vacancy concentrations.

Exercise 2.1-8

Quick Questions to

2.1 Intrinsic Point Defects and Equilibrium

2.1.2 Frenkel Defects; 2.1.3 Schottky Defects; 2.1.4 Mixed Point Defects

Here are some quick questions:

- The *answers* are sometimes (and possibly only indirectly) contained in the links.

2.1.2 Frenkel Defects

- Why do we need "Frenkel" and "Schottky" defects besides vacancies, self-interstitials and their agglomerates?
- Draw a schematic picture of a crystal with "Frenkel" and "Schottky" defects. What kind of "conservation laws" do you have to consider, and why does that lead to a fundamental difference between the two defect kinds?
- What kind of charge would "the" vacancy carry in a **NaCl** crystal? (Consider only the realistic case).
- Give some crystals where Frenkel disorder prevails.
- Give an (approximate) equation for the concentration of Frenkel defects an discuss the important terms

---

2.1.3 Schottky Defects and **2.1.4. Mixed Defects**

- What, quite generally, is the *Debye length*?
- How does the *Debye length* come into consideration when discussion Schottky defects (and Frenkel defects)?
- If the formation enthalpies of two defect kinds differs by roughly ...???...eV, one defect type will be dominating and the other one can be neglected.
- Why do we usually consider *either* Schottky *or* Frenkel defects in ionic crystals but not mixed defects? For the answer check this exercise.
- Can you predict for a given ionic crystal which kind of defect type (Schottky or Frenkel) will be prevalent?
- Discuss the details in the following set of equations:

$$Na_{Na} + V_i \Leftrightarrow Na_i + V_{Na}$$

$$\Downarrow$$

$$\frac{[Na_{Na}] \cdot [V_i]}{[Na_i] \cdot [V_{Na}]} = const = \exp\frac{G_{Reaction}}{kT}$$

$$\Downarrow$$

$$c_V(C) \cdot c_i(C) = \frac{N'}{N} \cdot \exp -\frac{H_{FP}}{kT}$$

$$\Downarrow$$

$$c_V(C) = c_i(C)$$

- Derive in an equivalent way the final relation for Schottky defects:

$$c_V(A) \cdot c_V(C) = \exp - \frac{H_S}{kT}$$

$$c_V(A) = c_V(C)$$

Exercise 2.2-1

Properties of Johnson Complexes

Illustration

- Discuss the equation for the concentration of vacancy - impurity atom complexes (Johnson complexes).
  - Consider an impurity atom concentration of **1 %** and **1 ppm**, a vacancy formation enthalpy of **1 eV** (neglect the formation entropy) and several binding energies (including extremes).
  - Discuss the concentration of Johnson complexes as a function of temperature and in relation to the concentration of the impurity atoms and the equilibrium concentration of vacancies.
- Use approximations, order-of- magnitude considerations and reasonable numbers whenever possible.

Link to the Solution

Exercise 2.2-2

Quick Questions to

2.2 Extrinsic Point Defects and Point Defect Agglomerates

2.2.1 Impurity Atoms and Point Defects; 2.2.2 Local and Global Equilibrium

- Here are some quick questions:
  - The *answers* are sometimes (and possibly only indirectly) contained in the links.

- Let's look at combined defects - double vacancies, impurity atom - vacancy complex (and so on):
  - Derive from the mass action law and write down the *essential* equations for the concentrations of
    - Divacancies (c_{2V})
    - Clusters with n vacancies (c_{nV}).
    - Impurity atom - vacancy complex (c_C; c_F = impurity atom conc.).

    Forget pre-exponential factors etc., if you don't remember, or make simple assumptions.
- Now let's do something important. *You really should do this, it will teach you a lot!*
  - Make *sketches* of various concentrations in an Arrhenius plot. Try to produce intellegent and neat sketches with parameters as follows:
    - The concentration of single vacancies at T_m is roughly $c_V \approx 10^{-4}$
    - Positive binding energies in all case; about **10 % - 20 %** of the vacancy formation enthalpy H_F**(V)**.
    - Always include the Arrhenius plot for the single vacany as reference.
  - *First*, produce one Arrhenius diagram showing single *and* double vacancies.
  - *Second*, produce an Arrhenius diagram for single vacancies, impurities, and impurity atom - vacancy complex
  - *Third*, produce one Arrhenius diagram showing single *and* double vacancies *but* assume that the single vacancy concentration *cannot* decline anymore at some lower temperature.
  - Discuss you curves (in particular the **2nd** and **3rd**), take into account how the temperature changes in "theory" and in "real life"
- Now a few really quick ones:
  - If all vacancies present at thernal equilibrium near the melting point at a concentration of $c_V \approx 10^{-4}$ end up in vacancy clusters with an average of **100** vacancies, what is the concentration of these clusters? What is their average cluster distance compared to the average vacancy distance (assume a typical lattice constant around **0.3. nm**)?
  - Given a equilbrium vacancy concentration of c_V, an (substitutional) impurity concentration c_F, and some binding enthalpy and entropy H_C and S_C, the concentration c_C of vacancy - (substitutional) impurity complexes should be proportional to:**.....**?
  - What would you expect for the case of no binding enthalpy and entropy?

Solution to Exercise 2.1-1"Find the Mistake"

Illustration

- The binomial coefficient in Hayes and Stoneham is written as

$$P_L = \frac{N!}{(N - n)!}$$

- The correct formula, of course, is

$$P_L = \frac{N!}{(N - n)! \cdot n!}$$

- This is obviously a typo, otherwise one set of brackets would not have been necessary.
  - The final result is correct anyway, because the next equation (3.5) contains the complete formula.

Solution to Exercise 2.1-2"Do the Math for the Formation Entropy"

Illustration

- We start with

$$F = kT\sum_i \frac{\hbar\omega_i}{kT} + \ln\left(1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\right)$$

- Next we must do the differentiation, i.e. form $\partial F/\partial T$:

$$S = -\frac{\partial}{\partial T}\left[kT\sum_i \frac{\hbar\omega_i}{kT} + \ln\left(1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\right)\right]$$

- One can go straight ahead, of course. But here comes a little helpful trick: Multiply skillfully by T/T and re-sort; you get

$$\begin{aligned}
&= -\frac{\partial}{\partial T}\left[k\sum_i \frac{\hbar\omega_i}{k} + T\cdot\ln\left(1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\right)\right] \\
&= -k\sum_i \frac{\partial}{\partial T}\left[T\cdot\ln\left(1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\right)\right] \\
&= -k\sum_i \left[\ln\left(1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\right) + T\cdot\frac{-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\frac{\hbar\omega_i}{kT^2}}{1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)}\right] \\
&= k\sum_i \left[-\ln\left(1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\right) + \frac{\exp\left(-\frac{\hbar\omega_i}{kT}\right)\frac{\hbar\omega_i}{kT}}{1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)}\right] \\
&= k\sum_i \left[-\ln\left(1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\right) + \frac{\exp\left(-\frac{\hbar\omega_i}{kT}\right)\frac{\hbar\omega_i}{kT}\cdot\exp\left(\frac{\hbar\omega_i}{kT}\right)}{1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\cdot\exp\left(\frac{\hbar\omega_i}{kT}\right)}\right] \\
&= k\sum_i \left[-\ln\left(1-\exp\left(-\frac{\hbar\omega_i}{kT}\right)\right) + \frac{\frac{\hbar\omega_i}{kT}}{\exp\left(\frac{\hbar\omega_i}{kT}\right)-1}\right]
\end{aligned}$$

- Now we need to resort to approximations
  - First we realize that whenever **h ·ω/2π << kT**, then

$$\exp\left(-\frac{\hbar\omega_i}{kT}\right) \approx 1 - \frac{\hbar\omega_i}{kT}$$

  - This takes care of the first term.
  - The second term needs a somewhat more sophisticated approach. Substituting ***x*** for **h · ω/2π · *kT***, we can use a simple expansion formula, stop after the second term and re-insert the result. This gives

$$\begin{aligned}
&\lim_{x\to 0}\frac{x}{\exp(x)-1} = \frac{x}{\left[1+x+\frac{x^2}{2}+\ldots\right]-1} = \frac{x}{x+\frac{x^2}{2}+\ldots} = \frac{1}{1+\frac{x}{2}+\ldots} = 1 \\
&\Rightarrow S \approx k\sum_i\left[-\ln\left[1-\left(\frac{\hbar\omega_i}{kT}\right)\right]+1\right] = k\sum_i\left[-\ln\left(\frac{\hbar\omega_i}{kT}\right)+1\right] \approx -k\sum_i \ln\left(\frac{\hbar\omega_i}{kT}\right)
\end{aligned}$$

- That's as far as one can go. Now use ω' for the circle frequencies of the crystal with a vacancy and form $S_F = S' - S$

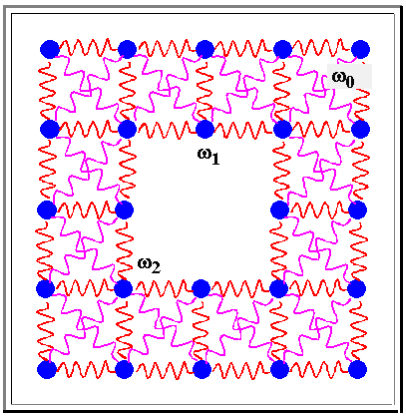
$$s_F = S'-S = -k\sum_i \ln\left(\frac{\hbar\omega_i'}{kT}\right) + k\sum_i \ln\left(\frac{\hbar\omega_i}{kT}\right) = k\sum_i \ln\left(\frac{\frac{\hbar\omega_i}{kT}}{\frac{\hbar\omega_i'}{kT}}\right) = k\sum_i \ln\left(\frac{\omega_i}{\omega_i'}\right)$$

  - *q.e.d.*

Solution to Exercise 2.1-3"Calculate the Formation Entropy"

Illustration

Lets look at a simple cubic lattice containing one vacancy and connect the atoms by springs symbolizing the bonds. It looks like this:



- We have two kinds of springs:
  - The red ones connect nearest neighbors and will heavily influence the vibration frequencies.
  - The violet ones, connecting diagonally. They will have some bearing on the vibration frequency, but since they must be weaker (the bond is weaker) than the red springs, their influence should be less pronounced.
- However, without the violet springs you could not make a stable crystal if you tried to built a model with balls and springs.
- Lets assign a spring constant ***D*** to the red springs and ***c · D*** to the violet springs, with ***c < 1***, and see what we get for the vibration frequency of an atom completely surrounded by other atoms and for the atoms around the vacancy.

Generally, the resonance (circular) frequency ω of a particle with mass ***m*** is given by

$$\omega^2 = \frac{D}{m}$$

In the most simple approximation, only accounting for the red springs, a regular atom would feel the force of *two* springs *per direction* and thus vibrate in any of the three dimensions with

.

$$\omega_0^2 = \frac{2D}{m}$$

The six atoms (for three dimensions) surrounding the vacancy and missing *one* red spring each (in one dimension), in contrast, would vibrate in one of the three dimensions with

$$\omega_1^2 = \frac{D}{m}$$

The entropy of formation S_F then becomes (note that we only have to sum over the "afflicted" dimensions):

$$S_F = k \cdot \sum_1^6 \ln \frac{\omega_0}{\omega_1} = k \cdot 6 \cdot \ln (2)^{1/2} = 3k \cdot \ln 2 = 3k \cdot 0{,}693$$

$$= 2{,}08\ k$$

Not bad for such a simple approximation. But now lets go one step further and add the violet springs.

- We have now for the frequency of the lattice atoms without a vacancy

$$\omega_0^2 = \frac{2D + 4c\,D}{m}$$

- and we simply include the factor $(1/2)^{-1/2}$, that would give us the component of the violet springs in the direction considered, into c; we thus have $c < 0{,}707$.

We now have to consider the **6** atoms with a missing red spring and **2** missing violet springs separately from the **12** atoms just missing one violet spring which are vibrating with ω_2, and consider the changed ω_0, too. Altogether we have

$$\omega_0^2 = \frac{2D + 4cD}{m}$$

$$\omega_1^2 = \frac{D + 2cD}{m}$$

$$\omega_2^2 = \frac{2D + 3cD}{m}$$

The entropy now is

$$S_F = k \cdot \left(\sum_1^6 \ln \frac{\omega_0}{\omega_1} + \sum_1^{12} \ln \frac{\omega_0}{\omega_2} \right)$$

- Crunching the numbers gives

$$S_F = 3k \cdot \ln \frac{2 + 4c}{1 + 2c} + 6 \cdot \ln \frac{2 + 4c}{2 + 3c} = 3k \cdot \ln (2) + 6k \cdot \ln \frac{2 + 4c}{2 + 3c}$$

- For $c = 0$ we must obtain our old result which indeed we do (check it), and for $c = 0{,}707$, the most extreme case possible, we find

$$S_F = 3k \cdot \ln (2) + 6k \cdot \ln (1{,}171) = 2{,}08\,k + 0{,}947k = 3{,}027\,k$$

In other words: For realistic c values, the correction is negligible and we can confidently claim that the formation entropy of a monovacancy in a cubic primitive lattice is around **2 k** in our ball and spring model approximation.

Solution to Basic Exercise 2.1-4"Derive the Formula for the Vacancy Equilibrium Concentration"

Illustration

First we need to determine the number of possibilites P_n to arrange n vacancies in a crystal of N atoms

This is most easily done by constructing a table and look at the cases $n = 1$, $n = 2$, etc. until it becomes obvious what the general law will be

n (= i)	p_n =	Comment
1	N	All N places are available
2	$\frac{N \cdot (N-1)}{2}$	N places for the first, only $N-1$ places for the second vacancy. Exchanging both vacancies does not change the situation - we have to divide by 2
3	$\frac{N \cdot (N-1) \cdot (N-2)}{2 \cdot 3}$	Exchanging vacancies does not change the microstate, we have to divide by the number of all possible exchanges = 6 = 2 · 3.
Make sure you understand the exchange argument: Here is the detailed reasoning: For vacancy No. 1 on place 1 , you have *two possibilities*: No. 2 on place 2, No. 3 on place 3 *or* No 2 on place 3 and No. 3 on place 2. You can do the same thing for No. 2 on place 2 (exchange No. 1 and No. 3) and for No. 3 on place 3., so you have 2 options 3 times = 6 indistinguishable arrangements.		
…	…	and so on
n	$\frac{N \cdot (N-1) \cdot (N-2) \cdot .. \cdot (N-(n-1))}{2 \cdot 3 \cdot .... \cdot n}$	The obvious law for n vacancies. {1· 2 · 3 · .... · n} of course is simply n!
n	$\frac{\{N \cdot (N-1) \cdot (N-2) \cdot .. \cdot (N-(n-1))\} \cdot \{(N-n)!\}}{n! \cdot \{(N-n)!\}}$	Extend the fraction by $(N-n)!$
n	$\frac{N!}{n! \cdot (N-n)!}$	Final result as used in subchapter 2.1 This is a standard expression in combinatorics and called the **binomial coefficient**.

The entropy of mixing thus is

$$S = k \cdot \ln \frac{N!}{n! \cdot (N-n)!} = k \cdot \left(\ln N! - \ln \{n! \cdot (N-n)!\} \right) = k \cdot \left(\ln N! - \ln n! - \ln (N-n)! \right)$$

We now can write down the free enthalpy for a crystal of N atoms containing n vacancies

$$G(n) = n \cdot G_F - kT \cdot [\ln N! - \ln n! - \ln (N-n)!]$$

Now we need to find the minimum of $G(n)$ by setting $dG(n)/dn = 0$ and for that we must differentiate ***factorials***. We will not do this directly (how would you do it?), but use suitable approximations as outlined in subchapter 2.1.

Mathematical approximation: Use the simplest version of the Stirling formula

$$\ln x! \approx x \cdot \ln x$$

- **Physical approximation**, assuming that there are far fewer vacancies than atoms:

$$n << N \quad \Rightarrow$$

$$\frac{n}{N - n} \approx \frac{n}{N} = c_V = \text{concentration of vacancies}$$

Now all that is left is some trivial math (with some pitfalls, however!). The links lead to an appendix explaining some of the possible problems.

- Essentially we need to consider $\mathbf{dS(n)/dn}$ using the Stirling formula

$$\frac{dS_n}{dn} = k \cdot \frac{d}{dn}\left(\ln N! - \ln n! - \ln (N-n)! \right) \approx k \cdot \frac{d}{dn}\left(N \cdot \ln N - n \cdot \ln n - (N-n) \cdot \ln (N-n) \right)$$

But we must not yet use the physical approximation, even so its tempting! With the formula for taking the derivative of products we obtain

$$\frac{dS_n}{dn} \approx k \cdot \left(\left((- \ln n - \frac{n}{n} \right) - \left(- \ln (N - n) + \frac{n - N}{N - n} \right) \cdot (-1) \right)$$

$$\frac{dS_n}{dn} \approx - k \cdot \left(\ln n + 1 - \ln (N-n) - 1 \right) = - k \cdot \left(\ln n - \ln (N-n) \right) = - k \cdot \ln \frac{n}{N - n}$$

- Now we can use the physical approximation and obtain

$$\frac{dS_n}{dn} \approx - k \cdot \ln c_V$$

Putting everything together gives

$$\frac{dG(n)}{dn} = 0 \quad G_F - T \cdot \frac{dS_n}{dn} = G_F + kT \cdot \ln c_V$$

- Reshuffling for c_V gives the final result

$$c_V = \exp - \frac{G_F}{kT}$$

- *q.e.d.*

- What happens if we use better approximations of the Stirling formula; e.g. $\ln x! \approx x\ln x - x$? Lets see:
  - We start with the equation from above and write it out with the better formula. With the extra terms in red, we obtain

$$\frac{dS_n}{dn} = k \cdot \frac{d}{dn}\left((N \cdot \ln N - N) - (n \cdot \ln n - n) - [(N - n) \cdot \ln (N - n) - (N - n)] \right)$$

  - After sorting out the signs, we have

$$\frac{dS_n}{dn} = k \cdot \frac{d}{dn}\left(N \cdot \ln N - N - n \cdot \ln n + n - [(N - n) \cdot \ln(N - n)] + N - n \right)$$

- Everything in red cancels and we are back to our old equation

Appendix: Mathematical tricks and Pitfalls

- Here are a few hints and problems in dealing with faculties and approximations.
- Having $n << N$, i.e. $n/(N - n) \approx n/N = c_V$ = concentration of vacancies does *not* allow us to approximate $d/dn\{(N - n) \cdot \ln (N - n)\}$ by simply doing $d/dn\{N \cdot \ln N\} = 0$.
  - This is so because d/dn gives the *change* of $N - n$ with n and that not only *might* be large even if $n << N$, but *will* be large because N is essentially constant and the only change comes from n.
- The derivative of $u(x) \cdot v(x)$ is: $d/dx(u \cdot v) = du/dx \cdot v(x) + dv/dx \cdot u(x)$.
  - The derivative of $\ln x$ is: $d/dx(\ln x) = 1/x$
- Easy mistake: Don't forget the *inner derivative*, it produces an important *minus* sign:

$$\frac{d}{dn}\left(\ln (N - n) \right) = \frac{1}{N - n} \cdot \frac{d(N - n)}{dn} = \frac{1}{N - n} \cdot (-1)$$

Solution to Exercise 2.1-5 "Do the Math for Mixed Point Defects"

Illustration

- *For obvious reasons some of the symbols deviate a little from the symbols used in the text; e.g. we have h_{FP} instead of H_{FP}.*

- We start with the system of equations that came from the mass action law

$$c_V(C) \cdot c_i(C) = \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)$$

$$c_V(A) \cdot c_V(C) = \exp\left(-\frac{h_S}{kT}\right)$$

$$c_V(C) = c_V(A) + c_i(C)$$

- We start with the calculation of $c_V(C)$:
  - Inserting the first and the second equation into the third equation yields:

$$c_V(C) = \frac{\exp\left(-\frac{h_S}{kT}\right)}{c_V(C)} + \frac{\frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)}{c_V(C)}$$

$$c_V^2(C) = \exp\left(-\frac{h_S}{kT}\right) + \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)$$

$$c_V(C) = \sqrt{\exp\left(-\frac{h_S}{kT}\right) + \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)}$$

$$c_V(C) = \sqrt{\exp\left(-\frac{h_S}{kT}\right) + \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right) \cdot \exp\left(-\frac{h_S}{kT}\right) \cdot \exp\left(+\frac{h_S}{kT}\right)}$$

$$c_V(C) = \exp\left(-\frac{h_S}{2kT}\right) \cdot \sqrt{1 + \frac{N'}{N} \cdot \exp\left(\frac{h_S - h_{FP}}{kT}\right)} \cdot$$

- That was the first equation for $c_V(C)$. Next we calculate $c_i(C)$.
  - Start with the third equation and eliminate $c_V(A)$ using the second. We have the final result after a series of mathematical manipulations:

$$c_i(C) = c_V(C) - c_V(A)$$

$$c_i(C) = c_V(C) - \frac{\exp\left(-\frac{h_S}{kT}\right)}{c_V(C)}$$

$$c_i(C) = \frac{c_V^2(C) - \exp\left(-\frac{h_S}{kT}\right)}{c_V(C)}$$

$$c_i(C) = \frac{1}{c_V(C)} \cdot \left\{\left[\exp\left(-\frac{h_S}{kT}\right) + \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)\right] - \exp\left(-\frac{h_S}{kT}\right)\right\}$$

$$c_i(C) = \frac{1}{c_V(C)} \cdot \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)$$

$$c_i(C) = \frac{\frac{N'}{N} \cdot \exp\left(\frac{h_S}{2kT}\right) \cdot \exp\left(-\frac{h_{FP}}{kT}\right)}{\sqrt{1 + \frac{N'}{N} \cdot \exp\left(\frac{h_S - h_{FP}}{kT}\right)}}$$

- That was the third equation. Next we calculate $c_V(A)$.
  - Start with the third equation and eliminate $c_i(C)$ using the first, we obtain

$$c_V(A) = c_V(C) - c_i(C)$$

$$c_V(A) = c_V(C) - \frac{N'}{N} \cdot \frac{\exp\left(-\frac{h_{FP}}{kT}\right)}{c_V(C)}$$

$$c_V(A) = \frac{c_V^2(C) - \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)}{c_V(C)}$$

$$c_V(A) = \frac{1}{c_V(C)} \cdot \left\{\left[\exp\left(-\frac{h_S}{kT}\right) + \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)\right] - \frac{N'}{N} \cdot \exp\left(-\frac{h_{FP}}{kT}\right)\right\}$$

$$c_V(A) = \frac{1}{c_V(C)} \cdot \exp\left(-\frac{h_S}{kT}\right)$$

$$c_V(A) = \frac{\exp\left(-\frac{h_S}{2kT}\right)}{\sqrt{1 + \frac{N'}{N} \cdot \exp\left(\frac{h_S - h_{FP}}{kT}\right)}}$$

That's it. Nothing to it. ;-)

- Well, not exactly. I myself certainly cannot solve problems like this without making some dumb mistakes in breaking down the math. Almost everybody does.
- However, I usually notice that I made a stupid mistake because the result just can't be true. And I can, if I really employ myself, get the right result eventually - because I did some exercises like this before. *And that is why you should do it, too!*.

As a last comment we may note that solving equations coming from the mass action law can become rather tedious very quickly - compare the example in the link, which is about as simple as it could be.

Now we look at the limiting cases of pure Schottky or pure Frenkel disorder.

- For pure Frenkel disorder we must have $h_{FP} << h_S$, and $c_V(A) = 0$.
- For pure Schottky disorder we must have $h_{FP} >> h_S$, and $c_i(C) = 0$.

For the first case - pure Frenkel disorder - just look at the expression

$$\left(1 + \frac{N}{N'} \cdot \exp\frac{h_S - h_{FP}}{kT}\right)^{1/2}$$

- For $h_S >> h_{FP}$, the exponential in this case is positive which means

$$\frac{N}{N'} \cdot \exp\frac{h_S - h_{FP}}{kT} >> 1$$

- So you may neglect the **1** in the above expression and replace the whole square root by

$$\frac{N}{N'} \cdot \exp\frac{h_S - h_{FP}}{2kT}$$

- This gives for $c_i(C)$

$$c_i(C) = \frac{N}{N'} \cdot \frac{\left(\exp\frac{h_S - 2h_{FP}}{kT}\right)^{1/2}}{\left(\exp\frac{h_S - h_{FP}}{kT}\right)^{1/2}} = \frac{N}{N'} \cdot \exp -\frac{h_{FP}}{kT}$$

- This is the result as as it should be.

With this we immediately obtain

$$c_V(C) = \frac{N}{N'} \cdot \exp - \frac{h_{FP}}{2kT}$$

$$c_V(A) = 0$$

- This is so because

$$\frac{N}{N'} \cdot \exp \frac{h_S - h_{FP}}{kT} >> 1$$

Contrariwise, if $h_S << h_{FP}$, $1 + N/N' \cdot \exp[(h_S - h_{FP})/kT] \approx 1$ obtains.

- Because $h_S - 2h_{FP}$ is a large negative number we get

$$c_i(C) = \frac{N}{N'} \cdot \exp \frac{h_S - 2h_{FP}}{2kT} \approx 0$$

- The expressions for $c_V(C)$ and $c_V(A)$ immediately reduce to the proper equation

$$c_V(C) = c_V(A) = \exp - \frac{h_S}{2kT}$$

Solution to Exercise 2.1-6 "Enthalpy difference for the limiting cases of Schottky or Frenkel Defects"

Calculate the ratio of the concentration of Schottky to Frenkel defect as a funtion of the enthalpy difference

The equations for the concentrations of the point defects in the "mixed" case are

Illustration

$$c_V(C) = c_S = \exp -\frac{H_S}{2kT} \cdot \left(1 + \frac{N'}{N} \cdot \exp \frac{H_S - H_{FP}}{kT} \right)^{1/2} = \exp -\frac{H_S}{2kT} \cdot K$$

$$c_V(A) = \exp -\frac{H_S}{2kT} \cdot \left(1 + \frac{N'}{N} \cdot \exp \frac{H_S - H_{FP}}{kT} \right)^{-1/2} = \exp -\frac{H_S}{2kT} \cdot K^{-1}$$

$$c_i(C) = c_{FP} = \frac{N'}{N} \cdot \exp \frac{H_S}{2kT} \cdot \exp -\frac{H_{FP}}{kT} \cdot \left(1 + \frac{N'}{N} \cdot \exp \frac{H_S - H_{FP}}{kT} \right)^{-1/2} = \frac{N'}{N} \cdot \exp \frac{H_S}{2kT} \cdot \exp -\frac{H_{FP}}{kT} \cdot K^{-1}$$

- Note that $c_V(C)$ or $c_i(C)$ is, by definition, identical to the concentration c_S or c_{FP} of Schottky or Frenkel defects, respectively. If you have problems with this, refer to the link.
- We abbreviated the root of the expression in square brackets by **K** for writing efficiency.

The ratio c_S / c_{FP} is easy to obtain. The **K's** cancel, we are left with

$$\frac{c_S}{c_{FP}} = \frac{N}{N'} \cdot \exp -\frac{(H_S - H_{FP})}{kT} = \frac{N}{N'} \cdot \exp -\frac{\Delta H}{kT}$$

- That is - of course - what we should have expected. The concentrations of Schottky and Frenkel defects are independent of each other and their relation could have been derived straight from the basic equations defining their equilibrium concentrations.

Show in in particular, how large the difference must be if **90%** or **99%** of the defects are to be of one kind.

We want to evaluate the equation for c_S/c_{FP} **= 0,011** or **= 0,001** (prevalence of Frenkel defects) and c_S/c_{FP} **= 90** or **= 99** (prevalence of Schottky defects).

- For the difference ΔH of the formation enthalpies as defined above we obtain

$$\Delta H = -kT \cdot \left(\ln \frac{N'}{N} + \ln \frac{c_S}{c_{FP}} \right)$$

- We have to define a value for N/N' ; we simply take this relation to be **1** or **0,1** as limiting cases.

Values are easily obtained, we arrange them in a little table

$\frac{c_S}{c_{FP}}$		99	90	10	0,1	0,011	0,010
ΔH [eV]	$\frac{N}{N'} = 1$	–0,115	–0,112	–0,058	0,058	0,112	0,115
	$\frac{N'}{N} = 10$	–0,172	–0,169	–0,115	–0,0004	0,054	0,057

Discuss the result.

We have two interesting results:

- If the formation enthalpies of the two defect kinds differ by just about **1/10** of an **eV**, we are fully justified to consider that only one defect kind is present.
- The pre-exponential factor ***N'/ N***, which describes the differences in the basic geometry for interstitials relative to vacancies, accounts at most for about **1/20** of an **eV** if expressed in enthalpy differences.

Solution to Exercise 2.2-1 "Properties for Johnson Complexes"

Illustration

- Discuss the equation for the concentration of vacancy - impurity atom complexes (Johnson complexes).
  - Consider an impurity atom concentration of **1 %** and **1 ppm**, a vacancy formation enthalpy of **1 eV** (neglect the formation entropy) and several binding energies (including extremes).
  - Discuss the concentration of Johnson complexes as a function of temperature and in relation to the concentration of the impurity atoms and the equilibrium concentration of vacancies.
- Use approximations, order-of- magnitude considerations and reasonable numbers whenever possible.

- The basic equation for the concentration of Johnson complexes is

$$c_C = \frac{z \cdot c_F \cdot c_V}{1 - z \cdot c_F} \cdot \exp \frac{\Delta S_C}{k} \cdot \exp \frac{H_C}{kT}$$

  - We first need to chose a coordination number, we take **z = 12** for **fcc** and **hcp** crystals. All other coordination numbers are smaller; we thus have the maximal effect of ***z***.
  - The given concentration of impurity atoms of **1 %** and **1 ppm** correspond to $c_F = 10^{-2}$ and $c_F = 10^{-6}$, respectively.
- First we note that the factor $1 - z \cdot c_F$ equals **0,88** or **0,999..**; i.e. we can forget it - at least for the low concentration.
  - Next we calculate the ratios c_C / c_F and c_C / c_V in order to get a feeling how the Johnson complex concentration relates to the (fixed) concentration of impurity atoms and the (temperature dependent) equilibrium concentration of vacancies. We have

$$\frac{c_C}{c_F} = (12 \ldots 13{,}6) \cdot c_V \cdot \exp \frac{\Delta S_C}{k} \cdot \exp \frac{H_C}{kT} = (12 \ldots 13{,}6) \cdot \exp - \frac{(H_F^V - G_B)}{kT}$$

$$\frac{c_C}{c_V} = (12 \ldots 13{,}6) \cdot c_F \cdot \exp \frac{\Delta S_C}{k} \cdot \exp \frac{H_C}{kT} = (12 \ldots 13{,}6) \cdot c_F \cdot \exp \frac{G_B}{kT}$$

  - The numbers in the bracket span the range of the given c_F concentrations.
- Our first result thus is simple: The ratios asked for are directly proportional to the concentration of vacancies or foreign atoms, respectively. The proportionality factor is about **2** times the Boltzmann factor of the free enthalpy of complex formation. So let's look at the role of the binding energy.
  - Let's look at binding energies (more precisely: binding free enthalpies G_B) of **– ∞ eV** (i.e. extreme repulsion between a vacancy and the foreign atom), **0 eV** (no interaction), **½** H_F^V (strong interaction), and H_F^V (extreme interaction). This gives us

G_B	– ∞	0	½ H_F^V	H_F^V
$\frac{c_C}{c_F}$	0	$\approx 12 c_V$	$\approx 12 \cdot (c_V)^{½}$	≈ 12
$\frac{c_C}{c_V}$	0	$\approx 12 c_F$	$\approx 12 \cdot c_F \cdot (c_V)^{-½}$	$\approx \frac{12 \cdot c_F}{c_V}$

- What does it mean?
  - *First*, for extreme repulsion, we simply do *not* form Johnson complexes as we would expect.

- *Second*, for zero interaction, we form Johnson complexes just *at random* - a vacancy just does not care if it sits next to an impurity atom or not. The concentration thus is directly given by the product of the concentrations of the partners (the factor **12** just accounts for the **12** different ways to form a Johnson complex with one vacancy).
- *Third*, for appreciable but not extreme binding energies the quotient c_C / c_F is always **< 1**, because $(c_V)^{1/2} << 1$; it decreases rapidly with temperature. *This means that in equilibrium only a small part of the foreign atoms will form Johnson complexes*.
- *Fourth*, for appreciable but not extreme binding energies the quotient c_C / c_V can be **>1** or **<1**, depending on $12c_F$ being larger or smaller than $(c_V)^{1/2}$. Below some temperature the vacancy concentration will always be so low that the ratio is **>1**, we then have *more* Johnson complexes than free vacancies. But that does not mean we have *many* - just more then the extremely few vacancies.
- *Fifth*, for extreme binding energies we have a problem. The relations given just must be wrong - we cannot for example, have **12** times as many Johnson complexes as we have foreign atoms. What went wrong?

Well, our starting formula is only valid under the assumption that $c_C << c_F$. This assumption is obviously violated for binding energies too large; we then must *not* use the simple formula.

- If we take the correct formula, we simply find that c_V times the exponential vanishes (i.e. c_C / c_V does not make sense anymore), and $c_C / c_F \approx z / (1 + z) \approx 1$ under all conditions, as we would expect.

Formation Enthalpies and Entropies for Vacancies and Self-Interstitials

Illustration

The following table contains some numbers found in the literature. for simple metals and **Si**. For more data activate the link

Crystal	H^F**(V) [eV]**	H^F**(i) [eV]**
Ag	1,1	No good numbers except $H^F(i) > H^F(V)$
Al	0,76	
Au	0,98	
Cu	1,0	
Si	? 2,0 - 4,5 not yet clear	? 2,0 - 4,5 not yet clear

Illustration

Detailed Derivation of Schottky Defect Equilibrium

Here is the detailed solution of the Poisson equation for Schottky defects:

Poisson equation of the problem

$$\boxed{\Delta V(\vec{r}) = -\frac{4\pi e N}{\epsilon\epsilon_0}\cdot\left\{\exp\left[\frac{-(h^- + eV(\vec{r}))}{kT}\right] - \exp\left[\frac{-(h^+ - eV(\vec{r}))}{kT}\right]\right\}} \quad (1)$$

$$\Delta V(\vec{r}) = -\frac{4\pi e N}{\epsilon\epsilon_0}\cdot\left\{\exp\left[\frac{-h^- - eV(\vec{r}) + \frac{h^+}{2} - \frac{h^+}{2}}{kT}\right] - \exp\left[\frac{-h^+ + eV(\vec{r}) + \frac{h^-}{2} - \frac{h^-}{2}}{kT}\right]\right\} \quad (2)$$

$$\Delta V(\vec{r}) = -\frac{4\pi e N}{\epsilon\epsilon_0}\cdot\left\{\exp\left[\frac{-eV(\vec{r}) + \frac{h^+}{2} - \frac{h^-}{2} - \frac{h^+}{2} - \frac{h^-}{2}}{kT}\right] - \exp\left[\frac{eV(\vec{r}) - \frac{h^+}{2} + \frac{h^-}{2} - \frac{h^+}{2} - \frac{h^-}{2}}{kT}\right]\right\} \quad (3)$$

$$\Delta V(\vec{r}) = -\frac{4\pi e N}{\epsilon\epsilon_0}\cdot\left\{\exp\left[\frac{-eV(\vec{r}) + \frac{h^+}{2} - \frac{h^-}{2}}{kT}\right]\cdot\exp\left[-\frac{\frac{1}{2}(h^+ + h^-)}{kT}\right] - \exp\left[\frac{eV(\vec{r}) - \frac{h^+}{2} + \frac{h^-}{2}}{kT}\right]\cdot\exp\left[-\frac{\frac{1}{2}(h^+ + h^-)}{kT}\right]\right\} \quad (4)$$

With $\boxed{v(\vec{r}) = \frac{eV(\vec{r}) - \frac{h^+}{2} + \frac{h^-}{2}}{kT}}$, (5)

follows

$$\Delta V(\vec{r}) = -\frac{4\pi e N}{\epsilon\epsilon_0}\cdot 2\cdot\exp\left[-\frac{\frac{1}{2}(h^+ + h^-)}{kT}\right]\cdot\frac{1}{2}\left[\exp(-v(\vec{r})) - \exp(+v(\vec{r}))\right] \quad (6)$$

$$\boxed{\Delta V(\vec{r}) = +\frac{8\pi e N}{\epsilon\epsilon_0}\cdot\exp\left[-\frac{\frac{1}{2}(h^+ + h^-)}{kT}\right]\cdot\sinh(v(\vec{r}))} \quad (7)$$

Multiplication with e/kT and using eq. 5 gives

$$\frac{e}{kT}\cdot\Delta V(\vec{r}) = \Delta v(\vec{r}) = \frac{8\pi e^2 N}{\epsilon\epsilon_0 kT}\cdot\exp\left[-\frac{\frac{1}{2}(h^+ + h^-)}{kT}\right]\cdot\sinh(v(\vec{r})) \quad (8)$$

With $\boxed{\chi^2 = \frac{8\pi e^2 N}{\epsilon\epsilon_0 kT}\cdot\exp\left[-\frac{\frac{1}{2}(h^+ + h^-)}{kT}\right]}$ (9)

we obtain finally $\boxed{\Delta v(\vec{r}) = \chi^2\sinh(v(\vec{r}))}$ (10)

1

Formation Enthalpies and Entropies for Frenkel and Schottky Defects

Illustration

The following table contains some numbers found in the literature. It is not complete, eventually it might get "fuller".

Schottky Disorder

Crystal	H_F [eV]	S_F [k]	H_M Cation vacancy [eV]	H_M Anion vacancy [eV]
LiF	2.5 [1)] 2.34 [2), 3)]	9.6 [1)]	0.7 [1)]	0.7 [1)]
LiCl	2.12 [2), 3)]			
LiBr	1.8 [3)]			
LiI	1,3 [3)]			
NaCl	2.3 [1), 2), 3)]	6 [1)]	0.7 [1)]	1.0 [1)]
KCl	2.3 [1)] 2.26 [3)]	6.5 [1)]	0.7 [1)]	1.0 [1)]
KBr	2.4 [1)]	8.6 [1)]	0.6 [1)]	0.9 [1)]
CsI	1.9 [1)]	-	0.6 [1)]	0.3 [1)]
MgO	6.6 [2)]			
CaO	6.1			

Frenkel Disorder

	H_F [eV]	S_F [k]	H_m Anion interstitial [eV]	H_M Anion vacancy [eV]
AgCl	1.6 [2), 3)]			
AgBr	1,20 [3)]			
β - AgI	0,7 [3)]			
CaF_2	2.7 [1)] 2.8 [2), 3)]	-	≈ 1.0 [1)]	0.6 [1)]
SrF_2	2.3 [1)] 0.7 [3)]	-	0.8 [1)]	0.9 [1)]
BaF_2	1.9 [1)]	-	0.7 [1)]	0.6 [1)]
PbF_2	1.1 [1)]	-	-	-

$SrCl_2$	1.7[1)]	-	-	-
ZrO_2	4.1[2)]			
UO_2	3.4[2)]			

- **1)** From "Hayes and Stoneham"; Defects an Defect Processes in Nonmetallic Solids
- **2)** From University of Hull, Lectures
- **3)** From Uni Lethbridge; California.

Point Defects in Ionic Crystals

Illustration

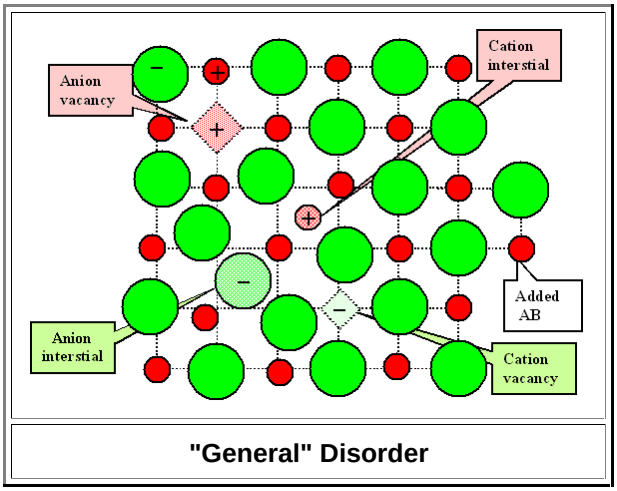
Most General (and Unrealistic) Case

It is important to be clear about the possibilities of producing defects in ionic crystals. It is also important to be clear about names:

- **Anions** move to a *positively* charged electrode also called *anode*, they are therefore *negatively* charged particles. Examples: The **Cl^-** ions in NaCl.
- **Cations** move to a negatively charged electrode also called *cathode*. Example: The **Na^+** ions in NaCl,

Now there is some room for confusion: If we take out the negatively charged cation **Na^+**, we have produced a *cation vacancy* that has a positive effective charge and thus behaves like an *anion*!

Here is all that can happen in a simple **NaCl** crystal:



"General" Disorder

Even for the most simple ionic crystals of the type **A^+B^-** like **LiCl** or **NaCl**, we can, *in principle*, produce arbitrary concentrations of two kinds of vacancies and two kinds of interstitials as shown on the left.
However, as we already learned in dealing with Schottky defects, global **charge neutrality** must be maintained. *Arbitrary* concentrations are thus not really allowed, we *must* demand that the the sum of the positively charged defects equals the sum of the negatively charged defects. In other words: we have to obey the *charge conservation law*.

If we also keep the number of atoms constant, we must add an **A** or **B** atom to the surface of the crystal for every pure vacancy we produce. In other (fancy) words, we have to obey the *mass conservation law.*

The picture on the left would not have needed the **AB** molecule, because we have two interstitials, too. But since it is supposed to illustrate the *general* case, with arbitrary numbers of defects, it needs to include **A** and **B** atoms on the surface.

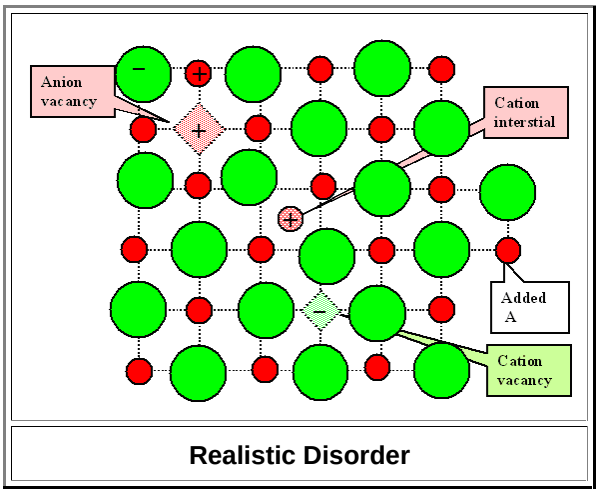
As always, you must bear in mind that pictures as shown here are *schematic* - in more realistic pictures the ions would touch! However, in more realistic pictures it also would be harder to show what is intended.

This will always be true if the anion is larger as the cation which is the case for many, but not all ionic crystals. We thus can safely assume that the concentration of *one* kind of interstitial, here the **anion interstitial**, is always far smaller than that of the other three defect kinds and we will simply neglect it from now on.

However, for crystals with a big and heavy cation (e.g. **Ca^+**) and a light anion (e.g. **F^-**), the cation might just be as big as the anion, and occur as interstitial (e.g. in the so-called "Anti-Frenkel defects").

Most General (and Realistic) Case

If we forget about the *anion interstitial*, we are left with three possible point defects.



Realistic Disorder

- The three now possible defect types are shown on the left. This is the general case of the **mixed defects** treated in the backbone
- Note that charge equilibrium demands that you *always* have more cation vacancies than anion vacancies or cation interstitials:

$$c_V(C) = c_V(A) + c_i(C)$$

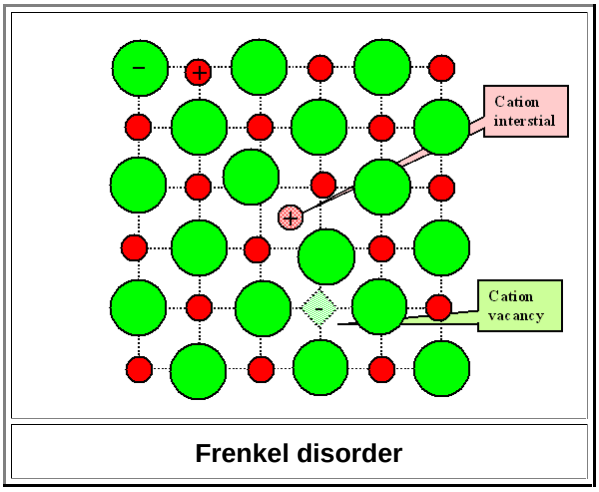
- This necessitates that some **AB** molecules must be added to the surface of the crystal if we keep the atom count in order, too (same concentration as the anion vacancy, to be precise).

The realistic mixed case thus contains Schottky and Frenkel defects in parallel.

- Note that the picture above does *not* show the equilibrium case, because we do not have charge neutrality - for that it would need another cation vacancy.
- Every cation vacancy finds an anion vacancy as a fictive partner, forming a formal Schottky defect, and every cation interstitial finds a cation vacancy, too, for a formal Frenkel pair, and the concentration of the anion vacancies is just so that it meets both demands for partners.
- We thus can identify $c_V(A)$ with the concentration of Schottky defects and $c_i(C)$ with the concentration of Frenkel defects.

Special Cases: Schottky and Frenkel Disorder

Schottky and Frenkel disorder may now simply be seen as *extreme* cases of the mixed disorder.

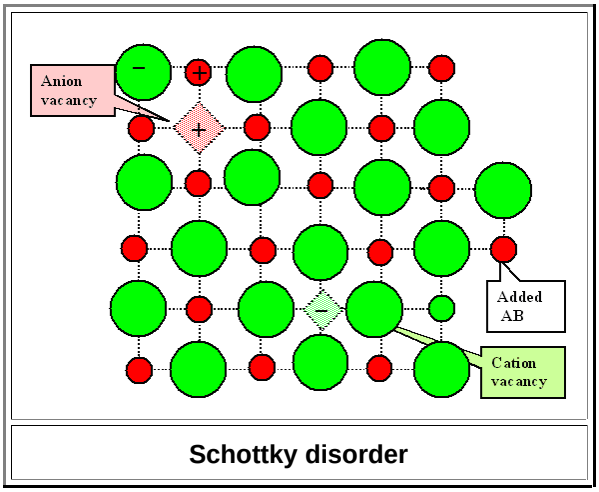


Frenkel disorder

- **Frenkel disorder** is predominant if the formation enthalpy of Frenkel pairs is smaller than that of Schottky pairs, i.e.

  $H_{FP} < H_S$

  We then only - or at least predominantly - have Frenkel pairs, i.e. equal numbers of cation vacancies and cation interstitials. We do not have an **AB** molecule for Frenkel disorder.



Schottky disorder

Schottky disorder is predominant if the formation enthalpy of Schottky pairs is smaller than that of Frenkel pairs, i.e.

$H_S < H_{FP}$.

We only - or at least predominantly - have vacancy pairs, i.e. equal numbers of cation vacancies and anion vacancies. And - in contrast to Frenkel disorder - we always need to form a lattice molecule, our **AB** molecule, to preserve atom numbers.

Just how much smaller the relevant formation enthalpy has to be for factual predominance of one defect type remains to be seen - in an exercise.

Microelectronics

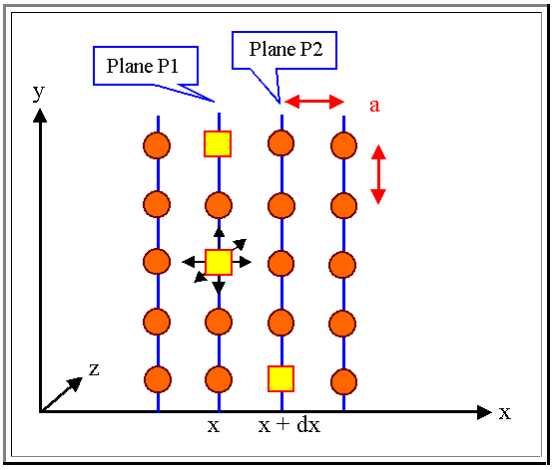
- *Microelectronics* is an inexhaustible topic in itself; it fills many books and we will encounter it in its own lectures. A taste of it is offered in the Hyperscript "Electronic Materials". At the very heart of microelectronics is the **Si wafer** and the structures integrated into it or on top of the the **Si**.
  - How is it done? By **defect engineering** - and there is the connection to defects! However, here this expression is used in a totally different context from that of an process engineer in a "wafer fab" i.e. a factory that makes **chips** (= integrated circuits) by processing wafers.
  - A process engineer considers "defect engineering" to be everything related to what produces "defective" chips, e.g. embedded particles, short-circuits etc.
- We, however, mean defects in the sense of this lecture, i.e. point defects, dislocations, etc. "*Defect engineering*" then comprises:
  - Growth of a **Si** single crystal with *no* grain boundaries or dislocations whatsoever, and very small and preferably very few point defect agglomerates and impurity precipitates.
  - Keeping that crystal clean and dislocation free - despite the fact that during the many high temperature processes necessary to make a chip, a lot of impurity atoms would like to diffuse into the **Si**. Temperature gradients, in addition, introduce mechanical stress which tends to relax by generation and movement of dislocations.
  - Get the right amount of dopant atoms in the right positions. This always involves defects - either generated by ion-implantation of the dopant, or the ones necessary for the diffusion. Still, they must be gone again in the end.
  - Have the right interface reactions, e.g. for forming an oxide. This involves point defects, too - oxidation, e.g., injects **Si** interstitials. Avoid, at all costs, to have those interstitials agglomerate into stacking faults!
- In summary: Chip making is indeed an exercise in defect engineering - as well as in equipment engineering, electrical engineering and so on.

Optoelectronics

- Optoelectronics includes all semiconductor devices which emit light through recombination of electrons and holes. Prime materials are **GaAs**, **GaAlAs**, **GaP**, **InSb** and generally all **III - V** semiconductors, but also **GaN** or **SiC**. More about optoelectronics can be found in an other Hyperscript.
  - Again, in making optoelectronic devices, *defect engineering* is needed. Diffusion plays a major role; the precise atomic mechanisms are not too well understood at present.
  - Moreover, defects in interfaces (= phase boundaries between different optoelectronic materials) play a major role; they essentially limit or prohibit applications in many cases.
- In contrast to **Si** microelectronics, defects may also play a role in the *finished device* while it is in operation. Dislocations, not wholly unavoidable in most **III - V** materials, may start to climb and degrade the function.
  - Early Lasers diodes, e.g., stopped working after few hours of operation because defects evolved that served as recombination centers impeding radiant recombination.

Diffusion Coefficient and Atomic Mechanisms

- We are looking for an equation that links the diffusion current ***j*** of Ficks **1st** law with the individual atomic jumps of a particle like an interstitial atom or a vacancy.
  - For simplicities sake we only consider vacancies in a primitive cubic lattice. The extension to interstitials is rather trivial.
  - We only consider a one-dimensional geometry.
  - Extensions to three dimensions, real crystals and exotic atomic mechanisms, albeit not necessarily easy, do not give new insights and will not be covered.
- Lets look at two lattice planes of a simple cubic crystal which are perpendicular to the ***x***-direction considered and which contain the diffusing particles - here vacancies.



- We are only interested in the flux of vacancies in the ***x***-direction, the **diffusion current *j*** of the *vacancies*. The flux or diffusion current of *atoms* that move via a vacancy mechanism, would have the same magnitude in the opposite direction.
  - We do not assume equilibrium, but a space-dependent vacancy concentration $c_V(x, y, z)$. Being one-dimensional, we only assume a concentration gradient in the ***x***-direction, $c_V(x, y, z) = c_V(x)$.
  - On any lattice plane perpendicular to ***x*** we have a certain number of vacancies per unit area (the area density in $\mathbf{cm^{-2}}$), which is computable by ***c(x)***. We distinguish this particular concentration with the index of the plane; i.e. P_1 is the number of vacancies on **1 cm²** area on plane **No. 1**, etc.
  - We then have

$$P_1 = a \cdot c_V(x)$$

$$P_2 = a \cdot c_V(x + dx)$$

  - With **d*x*** **= *a* =** *lattice constant*, because smaller increments make no physical sense, we obtain

$$P_2 = a \cdot c_V(x + a)$$

- Next we consider the jump rates in ***x***-direction, i.e. that part of all vacancy jumps out of the plane that are in **+*x***-direction. We define

r_{1-2} = **jump rate in *x* – direction from P_1 to P_2**

r_{2-1} = **jump rate in – *x* – direction from P_2 to P_1**

  - We obtain for our geometry:

$$r_{1-2}(T) = r_{2-1}(T) = \frac{1}{6} \cdot r(T)$$

- This means that **1/6** of the total number of possible jumps of a vacancy is in the ***+x*** or ***– x*** direction, the other possibilities are in the ***y-*** or ***z***-direction.

The jump rate itself is given by the usual Boltzmann formula

$$r = \nu_0 \cdot \exp - \frac{H^M}{kT}$$

- With ν_0 = vibration frequency of the particle, H^M = enthalpy of migration.

We obtain for the number of vacancies per **cm²** and second, which jump from P_1 to P_2, i.e. for the component of the diffusion current j_{1-2} flowing to the right (and this is *not* yet the diffusion current from Ficks law!):

$$j_{1-2} = P_1 \cdot r_{1-2}$$

- This is the current of vacancies flowing **out** in ***x***-direction from P_1. This current will be compensated to some extent by the current component j_{2-1} which flows **into** P_1. This current component is given by

$$j_{2-1} = P_2 \cdot r_{2-1}$$

- With the equation from above we obtain for the two components of the current

$$j_{1-2} = \frac{r}{6} \cdot a \cdot c(x)$$

$$j_{2-1} = \frac{r}{6} \cdot a \cdot c(x + dx)$$

The net j_x current in ***x***-direction, which *is* the current in Ficks laws, is exactly the difference between the two partial currents, we obtain

$$j_x = j_{1-2} - j_{2-1}$$

$$= -\frac{a \cdot r}{6} \cdot \{c(x + dx) - c(x)\}$$

- If we now multiply by **d*x*/d*x* = a/d*x*** we obtain directly Ficks first law for one dimension:

$$j_x = -\frac{a^2 \cdot r}{6} \cdot \frac{c(x + dx) - c(x)}{dx} = -\frac{a^2 \cdot r}{6} \cdot \frac{dc(x)}{dx}$$

- All we have to do is to indentify $(a^2 \cdot r)/6$ with the diffusion coefficient ***D*** of Fick's first law; we then have it in full splendor:

$$j_x = -D \cdot \frac{dc(x)}{dx}$$

- Ficks first law thus can be deduced in an unambiguous and physically sensible way for primitive cubic crystals in one dimension. (Mathematicians may have problems with the equality $dx = a$; but never mind).
  - We also obtain an equation for the *phenomenological* diffusion coefficient D in terms of the *atomic parameters* lattice constants and jump rate (for the simple cubic lattice).
- Considering arbitrary crystals now is easy.
  - The only parameters different in different crystal systems are the factor **1/6** and the jump distance, which does not have to be only a, but , in general, for jump type **i** will be Δx_i. With **i** we enumerate all geometrically different variants of jumps and take into account that the x- component may depend on **i**.
  - The diffusion coefficient then is given by

$$D = g \cdot a^2 \cdot r$$

  - And g is a constant which is specific for the lattice under consideration, it is the so-called **geometry factor** of the lattice for diffusion.
- If we reconsider how we obtained the factor **1/6** for the cubic primitive lattice used above, it is clear that in a general case the geometry factor is defined by the equation

$$g = \tfrac{1}{2} \cdot \Sigma_i \left(\frac{\Delta x_i}{a} \right)^2$$

  - The factor **1/2** takes into account that only **1/2** of all possible jumps must be counted, because the other half would be the jumps back. $\Delta x_i/a$ simply expresses the component of the jump in x-direction in units of a.
  - For simple lattices g is easily calculated; for the **fcc** and **bcc** lattice we have $g = 1$.
- Taking into account three dimension is easy, too:
  - In isotropic lattices (which, besides the cubic lattices, covers all poly-crystals) no direction is special, the above equations are equally valid for the y- and z-direction. We obtain then a vector equation for Ficks first law

$$j(r) = -D_0 \cdot \exp -\frac{E_M}{kt} \cdot \nabla c(x,y,z)$$

- In anisotropic crystals things are messy. Every direction has to be considered separately, the so far *scalar* quantity D evolves into a second-rank *tensor*. Fortunately, we do not have to consider this here.

Exercise 3.1-1

Calculate the Geometry Factor

Illustration

Calculate the geometry factor for diffusion using the formula given for

- The **fcc** lattice
- The **bcc** lattice
- The diamond lattice

Link to the Solution

Exercise 3.2-1

Crystal Identity

Illustration

Every now and then an atom in a crystal makes a jump to a neighboring place via its self-diffusion mechanism. After some time, we must expect that - on average - every atom has left its original place. Somehow we now have a different crystal. What then constitutes the "Crystal Identity"?

- Calculate how long it takes (on average and with simple approximations) until *every* atom in a crystal has made *one* jump as a function of the temperature and the relevant point defect parameters.
- Use the formula that gives the jump frequency of the atoms which are able to make a jump. Consider what must happen so *all* atoms can jump eventually.
- Discuss the results for some crystals of your choice. You may use the data provided in the script.

Link to the Solution

That this question is a bit philosophical becomes apparent if you substitute "brain" for crystal. It is known (from tracer techniques not unlike the ones used for studying diffusion) that most of the atoms that form your brain *now* will have disappeared after some weeks or month and *other* atoms of the same kind took their place (that's partially why you must eat!). Yet *your* identity seems completely unchanged.

Exercise 3.3-1

Quick Questions to

3. Point Defects and Diffusion

Here are some quick questions:

- The *answers* are sometimes (and possibly only indirectly) contained in the links.

3.1.1 Diffusion and Point Defects

- Give examples of products / processes / technologies that depend in a major way on point defect diffusion.
- Write down and discuss Fick's 1st law
- Write down and discuss Fick's 2nd law
- How do Ficks Laws connect to atomic diffusion? Give the two important equations dermining the diffusion coefficient***D*** i) by atomic jumping, and ii) via the diffusion length ***L***.
- What is the geometry factor ***g*** concerning jumping point defects in lattices? Can you give numbers for some common lattices?

3.2 Diffusion mechanisms

- Describe at least **2** possibilites for the diffusion of a substitutional and interstial impurity atom, respectively.
- Which diffusion mechanisms are the most important ones?
- For a substitutional impurity atom that diffuses via a vacancy mechanism, the diffusion coefficient ***D*** will be proprotional to ....?
- How do atoms diffuse in amorphous materials - e.g. glasses and polymers?
- What is self diffusion? The self-diffusion coefficient $\boldsymbol{D_{self}}$ is given by....?
- Any given atom in a given crystal will sooner or later leave its original place because of self diffusion. How long - roughly - does it take at high temperatures for all atoms to have changed positions? How about the atoms in your brain?

Solution to Exercise 3.1-1: "Calculate Geometry Factors"

Illustration

- The geometry factor (always for a single vacancy) was defined as

$$g = \tfrac{1}{2} \cdot \Sigma_i \left(\frac{\Delta x_i}{a} \right)^2$$

  - With Δx_i = component of the jump in x-direction.
- Looking at the fcc lattice we realize that there are **12** possibilities for a jump because there are **12** next neighbors.
  - **8** of the possible jumps have a component in x (or $-x$) -direction, and $\Delta x_i = a/2$
  - We thus have

$$g_{fcc} = \tfrac{1}{2} \cdot 8 \cdot \left(\frac{1}{2} \right)^2 = 1$$

- Looking at the bcc lattice we realize that there are **8** possibilities for a jump because there are **8** next neighbors.
  - All **8** possible jumps have the component $\Delta x_i = a/2$ in x-direction, again we have

$$g_{bcc} = \tfrac{1}{2} \cdot 8 \cdot \left(\frac{1}{2} \right)^2 = 1$$

- Looking at the diamond lattice we realize, after a bit more thinking (or drawing, or looking at a ball and stick model), that there are **4** possible jumps.
  - All **4** jumps have the component $\Delta x_i = a/4$ in x-direction, and we obtain

$$g_{diamond} = \tfrac{1}{2} \cdot 4 \cdot \left(\frac{1}{4} \right)^2 = 1/8$$

Solution to Exercise 3.2-1"Crystal Identity"

Illustration

- The jump rate of a vacancy is identical to that of an atom next to the vacancy. It was given by

$$\nu = \nu_0 \cdot \exp - \frac{G_m}{kT} \approx \nu_0 \cdot \exp - \frac{H_m}{kT}$$

  - The time t_a needed so that all the atoms with a vacancy next to them will make *one* jump thus is

$$t_a = \frac{1}{\nu} = \frac{1}{\nu_0} \cdot \exp \frac{H_m}{kT}$$

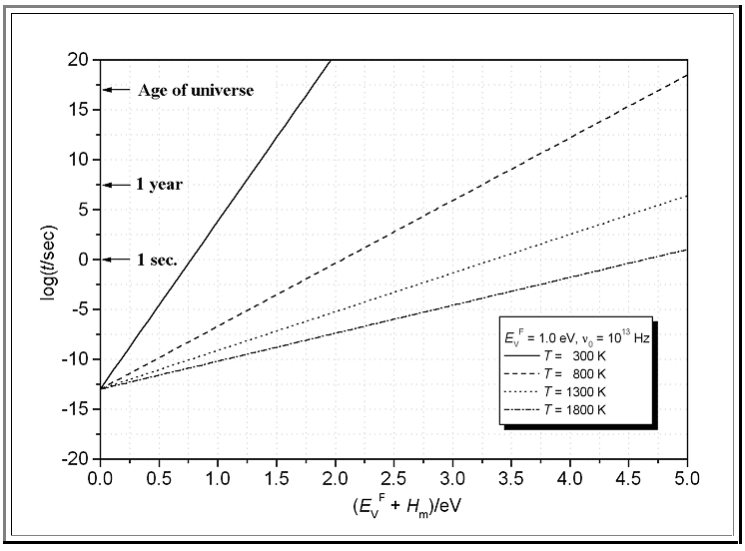
- After *that* time t_a, the fraction of all atoms that had a vacancy a a neighbor, has made *one* jump.
  - If you now wait another t_a, a *second* set of atoms can now make a jump. This second set may include atoms from the first set which simply jump back to their old position, but we ignore this effect for a rough estimate.
  - If all atoms of the crystal are supposed to make one jump, you have to wait for a time t_c that is a defined multiple of t_a. It is simply

$$t_c = m \cdot t_a = \frac{t_a}{c_V}$$

  - Because the multiplier m is of course the inverse of the vacancy concentration $c_V = \exp -(H_F)/kT)$
- t_c is the quantity we we are looking for, it is

$$t_c = \frac{1}{\nu_0} \cdot \exp \frac{H_m}{kT} \cdot \exp \frac{H_F}{kT} = \frac{1}{\nu_0} \cdot \exp \frac{H_m + H_F}{kT} = \frac{1}{\nu_0} \cdot \exp \frac{H_{SD}}{kT}$$

  - With H_{SD} = enthalpy of self diffusion.
- We may replace $1/\nu_0$ by $1/\nu_0 = g \cdot a^2 / D_{SD}$ and use the diffusion coefficient for self-diffusion to obtain values for specific materials, but lets just look at what we get in a very simple approximation with $\nu_0 = 10^{13}$ Hz



- Shown is t_c on a (rather far-reaching) **log** scale versus $H_m + H_F = H_{SD}$, i.e. the self-diffusion enthalpy H_{SD}, with the temperature as a parameter.
  - For $H_m + H_F = 0$, t_c is 10^{-13} s - as it should be.

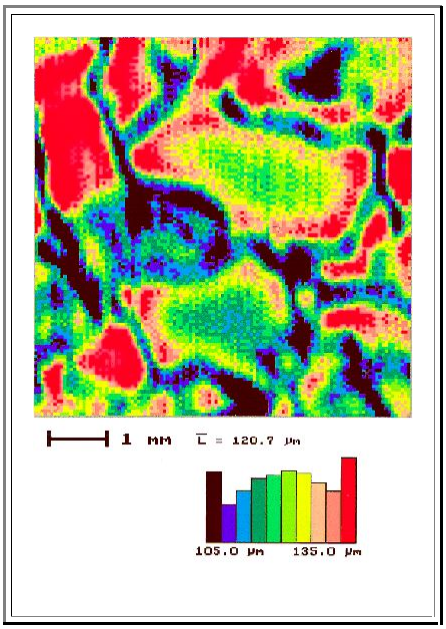
- For sensible values. e.g. $\mathbf{H_{SD} = 2\ eV}$, you must be very patient at room temperature, but at **800 °C**, your crystal has a different identity after **1 second**! Take **Si**, with $\mathbf{H_{SD}} \approx \mathbf{5\ eV}$ and a melting point of roughly **1700 K**, and again no atom will be where it was after a rather short time.

Using better values for ν_0 from the self-diffusion coefficient as stated above, just shifts the whole set of curves a "little bit" on the t - axis and thus t_c by the same (logarithmic) amount

Denuded Zones along Grain Boundaries

Illustration

Shown is a piece of polycrystalline Si.



- The colors code the **minority carrier diffusion length** which is a function of the concentration of certain impurities; in this case probably iron (in the **ppb** region). Light colors (yellow, red) denote large diffusion length and low impurity concentrations, dark colors the opposite. The scale gives precise numbers.
- The grain boundaries are essentially black, because carriers will recombine there; life time and thus diffusion length is small.
- Around the grain boundaries is a red/yellow zone, showing increased diffusion length as compared to the interior of a grain. This corresponds to a decreased impurity concentration, because the iron in the neighborhood of a grain boundary has diffused to the grain boundary where it is trapped. Knowing the thermal history of the sample allows an estimate of the diffusion coefficient of iron in Si.

Numbers for Point Defect Diffusion

Here a few numbers to point defect diffusion

- Numbers like this always should be taken with a grain of salt; they are often to a bit of doubt. It is not uncommon that newer measurements or new interpretations of old measurements give quite different results.

Illustration

Diffusing Atom	Host Crystal	Diffusion Mechanism	Migration Enthalpy (eV)	D_0 [cm^2/s]
C	Fe	interstitial	1,25	0,008
N	Fe	interstitial	0,78	0,007
H	Fe	interstitial	0,43	0,01
Ni	Fe	substitutional	2,86	0,5
Co	Fe	substitutional	2,34	0,2
Si	Fe	substitutional	2,08	0,4
Al	Cu	substitutional	1,69	0,07
S	GaAs	substitutional	4,0	4000
Zn	GaAs	substitutional	2,47	$1,5 \cdot 10^{-8}$
P	Si	substitutional		
As	Si	substitutionell		
B	Si	substitutional		
Si	Si			
O	Si			
Cu	Si			
Li	Si			

Numbers for Self-Diffusion

Illustration

- Here a few numbers for self diffusion
  - Numbers like this always should be taken with a grain of salt; they are often to a bit of doubt. It is not uncommon that newer measurements or new interpretations of old measurements give quite different results.
  - You may wonder a bit yourself, what self-diffusion in crystasl with two or more different atoms means, and how it relates to the prevalent defetc type, e.g. Schottky defects in **NaCl**.
- First some not-so-simple crystals:

Crystal	Diffusing Particle	Melting Point [°C]	Activation enthalpy H [eV] (= $H_{M,V} + H_{F,V}$)
H_2	H_2	- 259	0,016
Ar	Ar	- 189	0,18
H_2O	H_2O	100	0,58
NaCl	Cl	801	2,3
NaCl	Na	801	0,86
Ge	Ge	940	2,94
Si	Si	1412	5,11
GaAs	Ga	1238	5,54
GaAs	As	1238	9,96
Al	Al	660	1,47
Cu	Cu	1083	2,03
Ni	Ni	1455	2,86

- Now some metals; the values are form Neumann and Toelle (1986, 1990) as compiled by Kraftmakher
  - You will find *two* pre-exponential factors D_0 and *two* activation enthalpies H in tke left part of the table.
  - That is, because according to Neumann and Toelle, the self-diffusion data taken over a large region of temperatures do *not* form a straight line in an Arrhenius plot and therefore cannot be fitted with just *one* exponential.
  - So you fit with two exponentials, and it is anyone's guess what the second set (with the higher activation energy) actually describes. A common explanation is that you see the influence of double vacancies. While the formation energy is almost twice that of a single vacancy, the mgration energy canbe substantially lower - the sum thus may weel be relavant for self-diffusion.
  - But you could also argue that you see the influence of self-interstitials, or that this is all baloney; and that any curvature of the Arrhenius plots, if there is indeed some, is due to some temperature dependence of the formation/migration entrpopies and enthalpies (which could exist on theoretical reasons).
- However - the numbers you get are quite different for fits with one or two sets.
  - This serves as another example for how difficult it is to obtain unambiguos, air-tight data in the business!

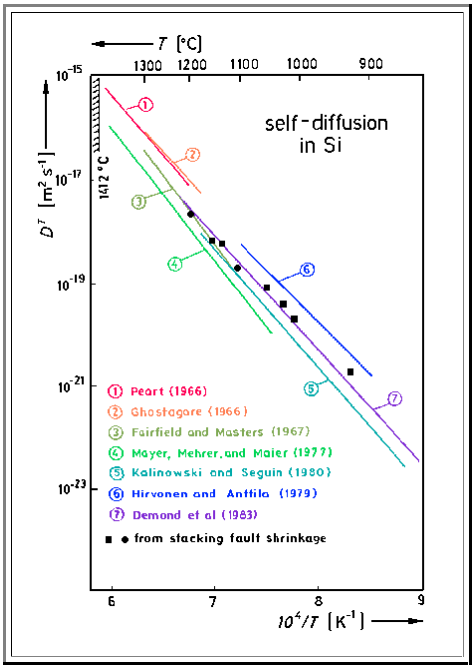
Crystal	Fitting with two sets				Fitting with one set	
	$D_0(1)$ cm^2s^{-1}	Activation enthalpy H_1 [eV]	$D_0(2)$ cm^2s^{-1}	Activation enthalpy H_1 [eV] $D_1(1)$ cm^2s^{-1}	D_0 cm^2s^{-1}	Activation enthalpy H_1 [eV]
Al						2,86
K	0.05	0.386	1	0.487		
Na	0.006	0.372	0.81	0.503		
Li	0.038	0.52	9.5	0.694		
Ag	0.055	1.77	15.1	2.35		
Au	0.025	1.70	0.83	2.20		
Cu	0.13	2.05	4.5	2.46		2,03
Ni	0.85	2.87	1350	4.15		1,47
Pt	0.034	2.64	88.6	4.05		
V	0.31	3.21	2420	4.70		
Nb	0.115	3.88	65	5.21		
Mo	0.13	4.54	140	5.70		
Ta	0.002	3.84	1.16	4.78		
W	0.13	5.62	200	7.33		

Illustration

Self-Diffusion in Si and some Metals

Arrhenius Representation of Self-Diffusion Data in Si

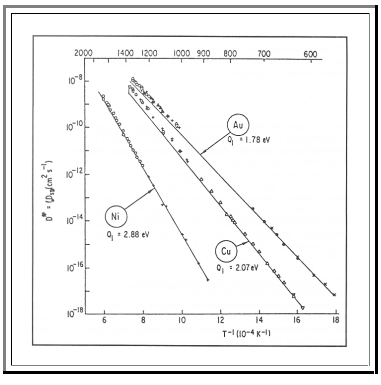
- Here are some tracer self-diffusion data from various researchers (compiled by *Frank et al.*, **1984**).



- Note that extrapolations to ***D**** differ by almost an order of magnitude.
- Be aware, that this discrepancy does not exist because some groups made "mistakes".
  - Measurements are what they are: measurements. Its the data you got and, if you made a honest effort, there is nothing "wrong" about it.
  - Why other groups obtain different data is an interesting question, often not easy to answer. While there must be reasons for it, its *not* because some groups are smart and others are dumb.

Arrhenius Representation of Self-Diffusion Data in Cu, Ni, and Au

- Here some data for impurity diffusion in **Si**



- There is far less spread of the data compared to Si self diffusion.

Self-Diffusion and some Related Quantities in Si

D_0 and Activation Energy E_{SD} for Self-Diffusion and for Various Impurities Including Intrinsic Point Defects in Si (fitted to $D = D^0 \exp(-E_{SD}/k_B T)$)

Illustration

The following table, which does not even come close to contain all relevant data, nicely illustrates how difficult it is to obtain reliable data for the point defect parameters. The situation has not yet (**1999**) changed, the values for formation and migration energies reported in the literature still vary from year to year.

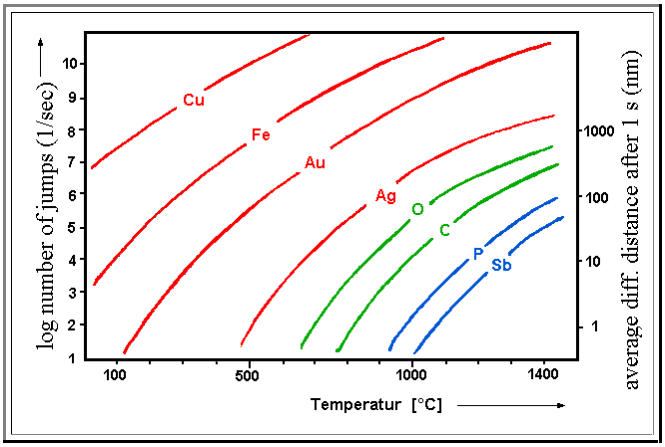
Diffusing Element	Measured Quantity (Type of Diffusion Mechanism)	D_0 [10^{-14} m^2 s^{-1}]	E_{SD} [eV]	Reference
Si	D^{Tracer}	1800	4.77	Peart, 1966
		1200	4.72	Ghostagore, 1966
		9000	5.13	Fairfield und Masters, 1967
		1460	5.02	Mayer et al., 1977
		8	4.1	Hirvonen und Antilla, 1974
		154	4.65	Kalinowski und Seguin, 1980
		20	4.4	Demond et al.; 1983
Si	$D_I C_I^{eq}$	914	4.84	Stolwijk et al.; 1984
		320	4.80	Stolwijk et al.; 1988
		2000	4.94	Hauber et al., 1989
		1400	5.01	Mantovani et al., 1986
Si	$D_V C_V^{eq}$	0.57	4.03	Tan und Gösele, 1985
		10^{-5}	0.4	Tan und Gösele, 1985
I	D_I	$3.75\ 10^{-9}$	0.13	Bronner und Plummer, 1985
		$8.6\ 10^{-5}$	4.0	Taniguchi et al.; 1983
		$2{,}42\ 10^{-5}$	0,937	Falster et al.98
V	D_V	0.1	2.0	Tan und Gösele, 1985
		$1{,}3\ 10^{-7}$	0,457	Falster et al.98
Ge	D^S	2500	4.97	Hettich et al.; 1979
Sn	D^S	32	4.25	Teh et al., 1968
C_S	D^S	1.9	3.1	Newman and Wakefield, 1961

C_i	D_i	4.4	0.88	Tipping and Newman, 1987
O	D_i	0.07	2.44	Mikkelsen, 1986

Impurity Diffusion in Si

Illustration

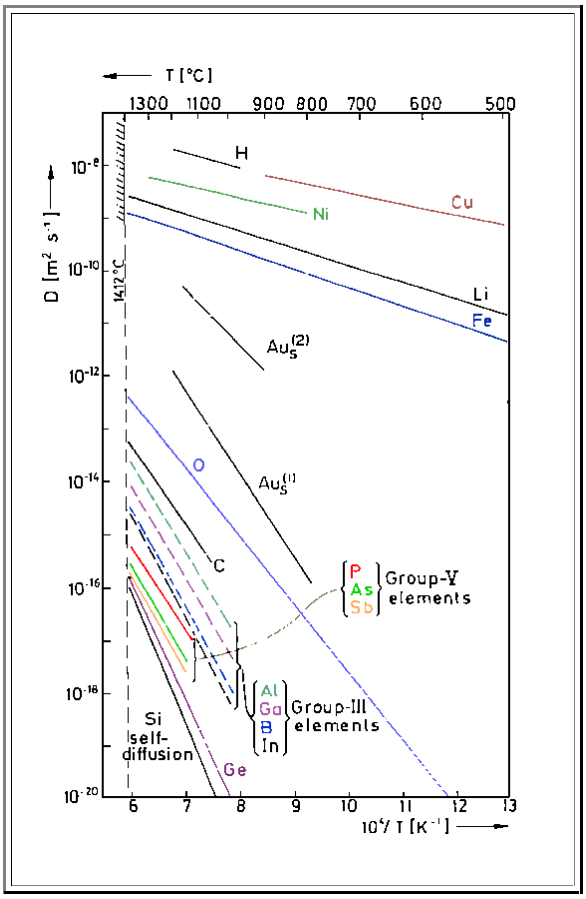
Shown is a somewhat unusual representation of impurity diffusion that gives a direct feeling for the number of jumps and distances covered as a function of the temperature. It is evident that waiting for long, but still reasonable times does not help anymore if the temperature is too low.



Impurity Diffusion in Si - Arrhenius Plot

Illustration

- Here are data of impurity diffusion in **Si**.
  - Quite trite - but these data are one of the basic cornerstones of the information society.



The Positron

- A positron is an elementary particle that behaves in all respects like an electron that has undergone certain symmetry operations that switched some signs - especially the sign of the elementary charge it carries.
  - Elementary particles with these reversed symmetries are called **anti-particles**, and *every* particle has an anti-particle as a partner in symmetry.
  - Even the photon has an anti-particle. However, since all photon properties for which the sign would be reversed upon the "anti"-operation are zero, the photon is its own anti-particle.
- Anti-particles can exist by themselves just as happily as "real" particles; they are, however, rare in *our* universe. There seems to be an excess of particles - all anti-particles have long since vanished. The prefix "anti", of course, just mirrors a human prejudice.
  - If a particle and an anti-particle meet, they annihilate each other in a burst of radiation; in the case of electrons ($\mathbf{e^-}$) and positrons ($\mathbf{e^+}$), two γ quanta with the combined energy of the two particles (according to $\boldsymbol{E=mc^2}$) are sent out (**511 keV** each if the particles were at rest).
- Do *not* confuse *positrons* with *holes* ($\mathbf{h^+}$).
  - Holes are merely missing electrons in energy levels that are otherwise completely filled with electrons; they do not exist by themselves outside of a crystal as positrons do!

Exercise 4.1-1

Lifetime of Positrons

Illustration

Show that the solution of the differential equations for the positron concentrations n_1 and n_2

$$\frac{dn_1}{dt} = -(\lambda_1 + \nu \cdot c_V) \cdot n_1$$

$$\frac{dn_2}{dt} = -\lambda_2 \cdot n_2 + \nu \cdot c_V \cdot n_1$$

- leads to the following formula for the average lifetime

$$\tau = \tau_1 \cdot \left(\frac{1 + \tau_2 \cdot \nu \cdot c_V}{1 + \tau_1 \cdot \nu \cdot c_V} \right)$$

Link to the Solution

Exercise 4.2-1

Diffusion During Cooling

Illustration

A (big) crystal cools down from its melting point T_m to room temperature T_r (about 0^o C) with $T = T_m \cdot \exp - (\lambda \cdot t)$. The point defects present have a diffusion coefficient given by $D = D_0 \cdot \exp - (E_m/kT)$.

How large is the average distance L that they cover during cooling down from some temperature T to T_r?

This is not an easy question. What you should do is:

- Use the Einstein relation for the diffusion length (and forget about lattice factors), but consider that the diffusion coefficient is a function of time, i.e.

$$L^2 = 6D \cdot t = \int_{t' = t_0}^{t' = \infty} D(t') \cdot dt'$$

- Proceed by first finding the values of λ for *initial* cooling rates at the melting point of **1 ^{o}C/s**, **10 ^{o}C/s**, **50 ^{o}C/s** and, for fun, **10^4 ^{o}C/s**.
- Using the following substitution will help with the integration

$$u(t) = \frac{E_m \cdot \exp \lambda \cdot t}{kT_m}$$

- The integral now runs from u_0 corresponding to t'_0 to whatever value of u corresponds to $t' = \infty$.
- You will obtain the following integral:

$$L^2 = 2D_0 \int_{?}^{?} \frac{1}{u} \cdot \exp - u \cdot du$$

- This integral cannot be solved analytically. In order to get a simple and good approximation, you may use the linear Taylor expansion for $1/u$ around u_0.
- Show that for realistic u_0 values you can replace $1/u$ by $1/u_0$ in a decent approximation and that you now can do the integral.

Now use typical values for melting temperatures, migration activation energies E_m, and D_0; e.g. from the backbone, two tables or diagrams given here. For missing values (e.g. D_0), make some reasonable assumptions.

- Plot L as a function of T for activation energies E = **1.0 eV**, E = **2.0 eV**, and E = **5 eV** with the four cooling rates given above as parameter.
- Play around a bit and draw some conclusions, e.g. with respect to

  - Average density of precipitates of point defects obtained in big crystals with few internal sinks.
  - Average size of these precipitates for some equilibrium concentration c_0 at T_m.
  - Possible errors made in quenching experiments.
  - Influence of sinks for point defects as a function of the average distance between sinks

Link to the Solution

Solution to Exercise 4.1-1 "Lifetime of Positrons"

Illustration

Show that the solution of the differential equations for the positron concentrations n_1 and n_2

$$\frac{dn_1}{dt} = -(\lambda_1 + \nu \cdot c_V) \cdot n_1$$

$$\frac{dn_2}{dt} = -\lambda_2 \cdot n_2 + \nu \cdot c_V \cdot n_1$$

The way to obtain the desired equation shall only be sketched.

- First, the coupled differential equation from above need to be solved for the initial condition

$$n_1(t = 0) + n_2(t = 0) = n_0$$

- n_0 is the number of thermalized positrons in the crystal at the beginning of the experiment: i.e. at $t = 0$

This is easy to do since the first differential equation does not contain n_2.

- The solution must be

$$n_1(t) = A \cdot \exp-(\lambda_1 + \nu \cdot c_V)t$$

Insertion in the second differential equation and using the initial condition yields

$$n(t) = n_0 \cdot \frac{\lambda_1 - \lambda_2}{\lambda_1 - \lambda_2 + \nu \cdot c_V} \exp-(\lambda_1 + \nu \cdot c_V)t + n_0 \cdot \frac{\nu \cdot c_V}{\lambda_1 - \lambda_2 + \nu \cdot c_V} \exp-\lambda_2 t$$

We have two components decaying with two lifetimes, τ_1 and τ_2, given by

$$\tau_1 = \frac{1}{\lambda_1 + \nu \cdot c_V}$$

$$\tau_2 = \frac{1}{\lambda_2}$$

Measurements usually only yield an *average* lifetime $<\tau>$

- The *average* lifetime is not simply the average of τ_1 and τ_2 because we need *weighted* averages, i.e.

$$<\tau> = \frac{1}{n_0} \int_0^\infty t \frac{dn(t)}{dt} dt$$

$$<\tau> = \tau_2 \cdot \frac{1 + \tau_2 \nu \cdot c_V}{1 + \tau_1 \nu \cdot c_V}$$

Doing the integral takes a few lines, but it is not too difficult - try it!

Solution to Exercise 4.2-1 "Diffusion During Cooling"

Illustration

For the diffusion length ***L*** we have the well known equations:

$$L = \sqrt{2\,Dt} \quad or \quad L^2 = 2\,Dt$$
$$D = D_0 \exp\left(-\frac{E}{kT}\right).$$

- ***E*** is the activation energy of the diffusing species an **k** is the Boltzmann constant. Because of $T = T_0 \cdot \exp -(\lambda \cdot t)$ we obtain for L^2

$$L^2 = 2\,D(t)t = 2\,D_0 \int_0^\infty \exp\left(-\frac{E}{kT(t)}\right) dt$$
$$= 2\,D_0 \int_0^\infty \exp\left(-\frac{E}{kT_0 \exp(-\lambda t)}\right) dt$$
$$= 2\,D_0 \int_0^\infty \exp\left(-\frac{E}{kT_0} \exp(\lambda t)\right) dt$$

Now we have a purely mathematical exercise which is not too difficult, but not too easy either. In order to solve the integral, we try the substitution

- $$u(t) = \frac{E}{kT_0} \exp(\lambda t), \quad dt = \frac{du}{\lambda u}$$

- The boundaries must be changed too, we obtain

  $t = 0$ changes to $u_0 = E/kT_0$
  $t = \infty$ changes to $u = \infty$.
  This gives us

$$L^2 = \frac{2\,D_0}{\lambda} \int_{u_0}^\infty \frac{1}{u} \exp(-u)\,du\ .$$

Now you must solve a simple looking integral. There are several ways of doing that

- 1. Find a good math book with lots of integrals and take the solution from there (the "Bronstein", however, won't do)
  2. Do a sensible approximation and solve it yourself in a simple way
  3. Go all the way and solve it completely - if you can.
- Here we go the second route.

We use a Taylor expansion for $1/u$ around u_0 because that's where u is felt most critically - for large values of u everything tends to be zero anyway. In full generality we have

$$\frac{1}{u} = \sum_{\nu=0}^{\infty} (-1)^\nu \frac{1}{u_0^{\nu+1}} (u - u_0)^\nu \approx \frac{1}{u_0} - \frac{u - u_0}{u_0^2}.$$

- If we keep it really simple, we could just use the first term, having $1/u \approx 1/u_0$; but we will go one step beyond this and take

$$\frac{1}{u} \approx \frac{1}{u_0} - \frac{u - u_0}{u_0^2}$$

- This gives us

$$L^2 = \frac{2 D_0}{\lambda u_0} \int_{u_0}^{\infty} \exp(-u) - \frac{u - u_0}{u_0} \exp(-u)\, du$$

$$= \frac{2 D_0}{\lambda u_0} \int_{u_0}^{\infty} 2 \exp(-u) - \frac{u}{u_0} \exp(-u)\, du$$

$$= \frac{2 D_0}{\lambda u_0} \left[-2 \exp(-u) + \frac{u + 1}{u_0} \exp(-u) \right]_{u_0}^{\infty}$$

$$= \frac{2 D_0}{\lambda u_0} \exp(-u_0) \left[1 - \frac{1}{u_0} \right]$$

$$= \frac{2 D_0 k T_0}{\lambda E} \exp\left(-\frac{E}{k T_0}\right) \left[1 - \frac{k T_0}{E} \right]$$

$$\Rightarrow L = \sqrt{\frac{2 D_0 k T_0}{\lambda E} \left[1 - \frac{k T_0}{E} \right]} \exp\left(-\frac{E}{2 k T_0}\right).$$

- The second term of the Taylor expansion brought in the factor **[1 – kT_0/E]** and since **kT_0 « E** in all normal cases, it is indeed not very important. If we neglect it, we may simply give the desired solution as

$$L = \left(\frac{2D_0 \cdot kT_0}{\lambda \cdot E} \right)^{1/2} \cdot \exp\left(-\frac{E}{2kT_0} \right)$$

- Now we can look at some typical cases and see what this formula means. However, first we have to find the right values for λ
  - For this we have to take the given values of the initial cooling rate, which we call λ', and see what λ values correspond to these cooling rates.
  - The initial cooling rate λ' is the derivative of the ***T*(*t*)** function at ***t* = t_0 = 0**, we thus have

$$\frac{d}{dt} (T_0 \cdot \exp -\lambda \cdot t \Big|_{t = 0} = \lambda' = -\lambda \cdot T_0 \cdot \exp -\lambda \cdot t \Big|_{t = 0} = -\lambda \cdot T_0$$

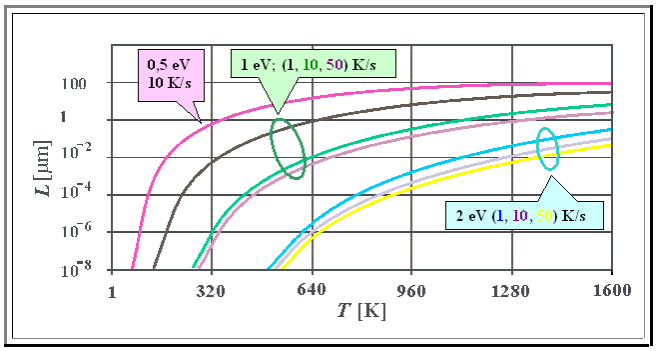
  - and obtain

$$\lambda = \frac{\lambda'}{T_0}$$

  - The "–" sign cancels, because our λ' must carry a minus sign, too, if it is to be a cooling and not a heating rate.
- Replacing λ by λ'/T_0 yields the final formula:

$$L = \left(\frac{2D_0 \cdot kT_0^2}{\lambda' \cdot E} \right)^{1/2} \cdot \exp\left(-\frac{E}{2kT_0} \right)$$

  - We have to evaluate this formula for cooling rates λ' given as (–) **1 ^{o}K/s**, **10 ^{o}K/s**, **50 ^{o}K/s**, **10^4 ^{o}K/s**, and activation energies of ***E* = 1.0 eV, 2.0 eV, 5 eV**. For D_0 we take **$D_0 = 10^{-5}$ cm^{2}s^{-1}**.
  - The result (including the **[1 – kT_0/E]** term is shown below



What can we learn from the formula and the curves?

1. The cooling rate is not all that important. Differences in the cooling rate of a factor of **50** produce only an order of magnitude effect or less since ***L*** is only proportional to $(1/\lambda)^{1/2}$.
2. The starting temperature T_0 is slightly more important than the activation energy ***E***; both have the same weight in the exponential, but T_0 appears directly in the pre-exponential while ***E*** enters only as square root.
3. The pre-exponential factor D_0 of the diffusion coefficient is exactly as important as λ' and ***E*** in the pre-exponential factor of the equation for ***L***

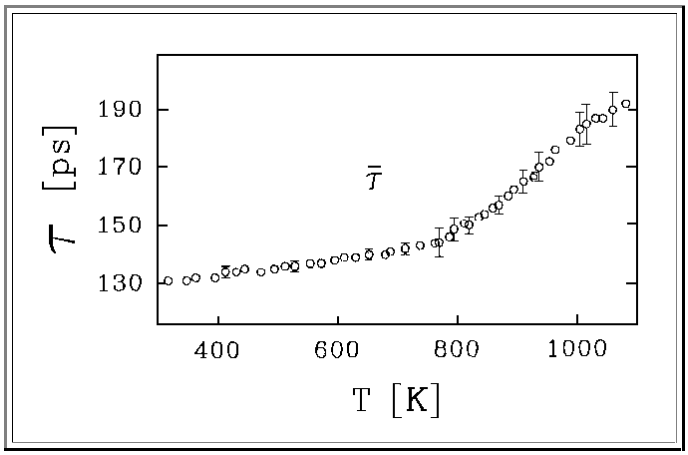
What can we do with the numbers? Quite simple:

1. ***L*** gives you the average of the largest distance between some point defect agglomerates, e.g. precipitates, because point defects farther away than ***L*** from some nuclei cannot reach it and must form their own agglomerate.
2. The average number of point defects in an agglomerate divided by L^3 gives a lower limit for the point defect concentration, because at least as many point defects as we find in an agglomerate must have been in the volume L^3.

Example for Positron Life Time Measurement in Ag

Illustration

- Here is an example for a positron life time measurement.

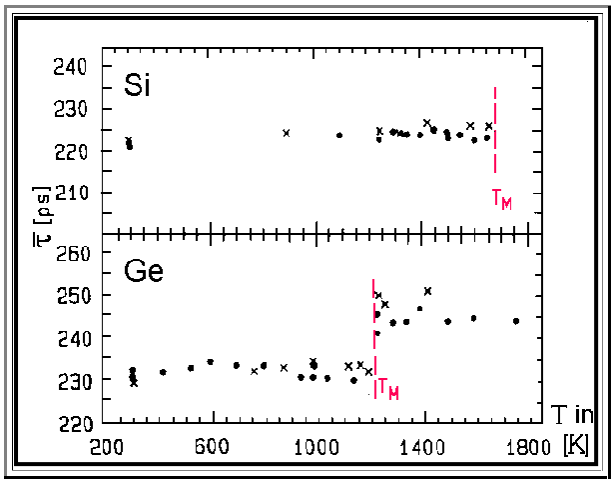


- From Dr. Wolf; one of the researchers in the group of Prof. Faupel in Kiel
  - The "S-curve" is clearly visible; the linearly rising part is simply due to thermal expansion which decreases the electron density and thus increases the positron lifetime.

Positron Life Time in Si and Ge

Illustration

Hiere is the experimental demonstration that you do not find point defects with positron life time spectroscopy in **Si** and **Ge**.

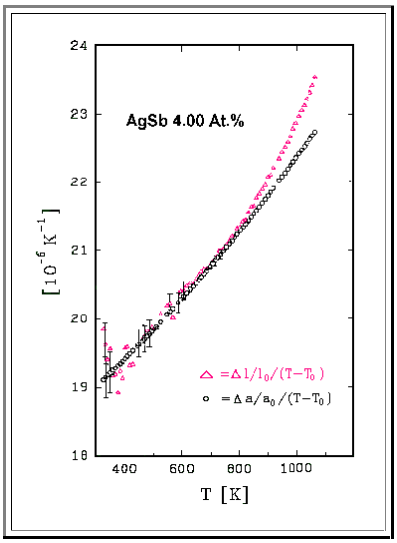


***T*M** denotes the melting point. There is no discernible influence of temperature on the positron life time.

Example for ∆l/l - ∆a/a Measurements

Illustration

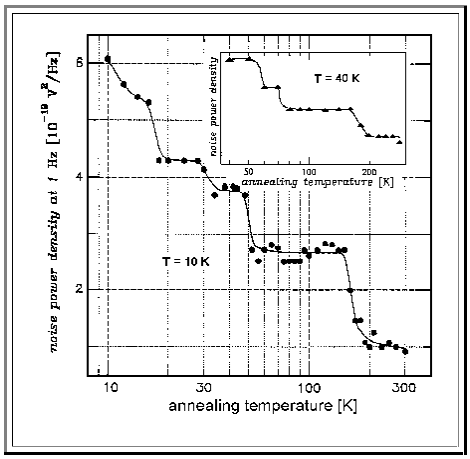
Here an example for a differential dilatometry measurement.



Example for Noise Measurements

Illustration

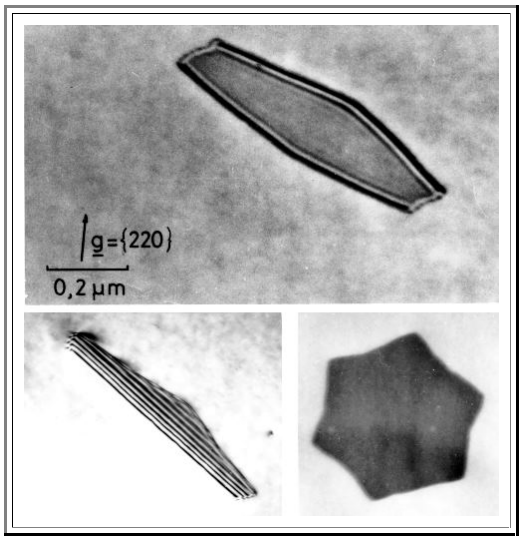
An example for a noise measurement showing definite recovery steps due to point defect annealing.



Interstitial Agglomerates in Si

Illustration

- Shown are small stacking fault loops, i.e. small platelets of **Si**-atoms wedged in between two **(111)** lattice planes. This then might be considered to be a two-dimensional agglomerate of **Si** self-interstitials.
  - In the picture on the lower right the **(111)** plane is in the paper plane, in the other cases it is inclined; in the lower left case the (hexagonal) loop is truncated by the specimen surfaces.

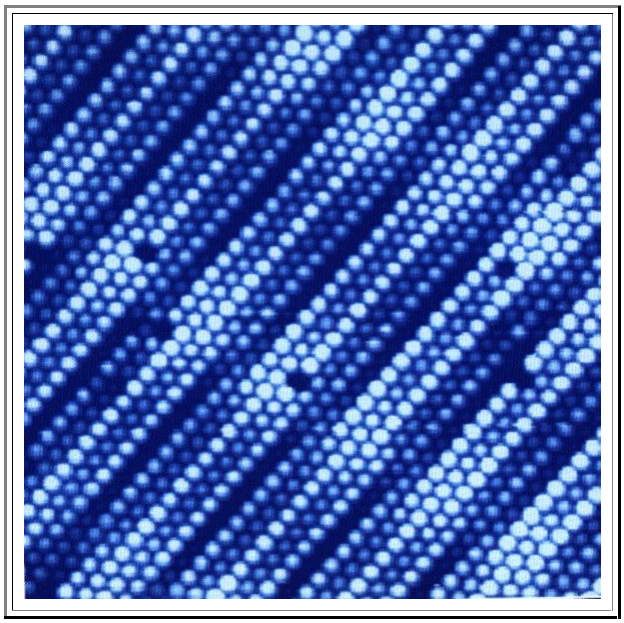


- For more details about electron microscopy refer to chapter 6.3.3

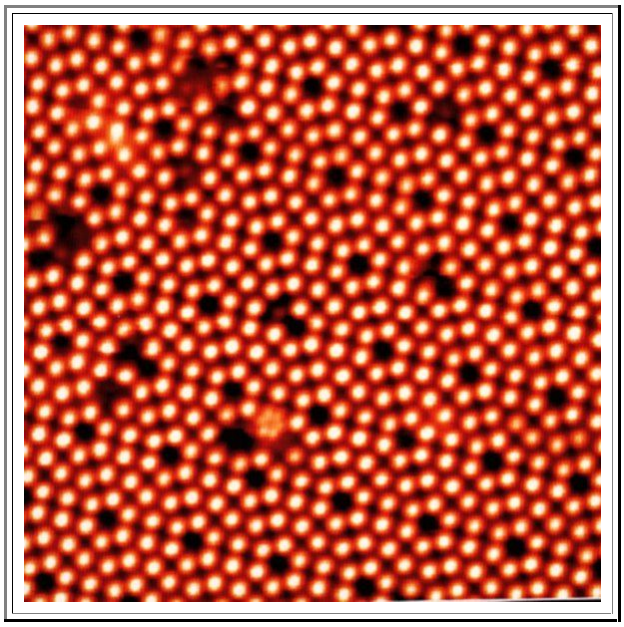
STM Images of Point Defects

Illustration

- This is a ***STM*** image of a **Pt** surface. Vacancies are clearly visible



- Next we see the clean **{111} Si** surface in ultra high vacuum conditions (otherwise the surface would immediately oxidize and we would see amorphous $\mathbf{SiO_2}$).

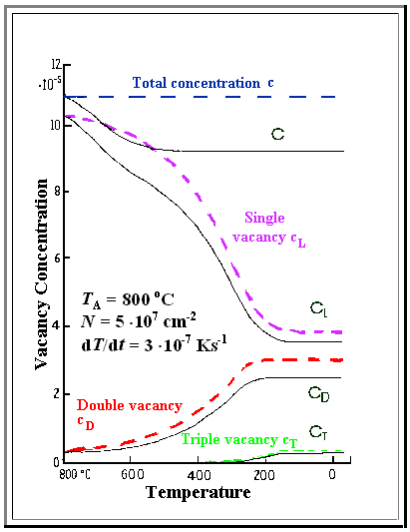


- The **{111}** surface looks not as one would expect from a straight model - it has "reconstructed". This means that the surface layer formed a two-dimensional crystal that is totally different from the **fcc Si** lattice (it has a so-called **7 × 7** symmetry). Still, point defects are clearly visible.
- These fantastic pictures are all over the world. I do not know whom I should acknowledge, sorry.

Calculated Vacancy Concentration After Quenching

Illustration

- Calculated changes in the concentration of single- double- and triple vacancies (c_L, c_D and c_T) and the total concentration $c = c_L + 2c_D^0 + 3c_T$ in **Au** during quenching from **800 °C** with $dT/dt = 3 \cdot 10^4$ **K/s** (after **Furuka**).
  - The colored dashed lines assume a dislocation density of zero (i.e. no sinks, $N = 0$), whereas the solid lines assume a dislocation density of $N = 5 \cdot 10^7$.

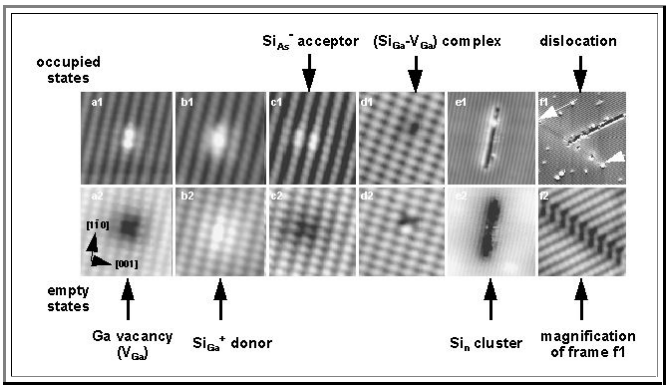


  - Without sinks, the *total* concentration c of vacancies does not change is required (since no clusters with more that **3** vacancies are allowed). The concentration of single vacancies, however, changes considerably despite the large cooling rate.
  - The presence of sinks does change the picture somewhat, but not dramatically as we would expect for large cooling rates - there simply is not enough time to migrate to a sink.
  - For the migration energies ($E_{X,M}$) and the binding energies $E_{X;B}$ the following values were used: $E_{L,M} = 0.83$ eV, $E_{D,M} = 0.71$ eV, $E_{D;B} = 0.35$ eV, $E_{T,B} = 0.65$ eV.
  - See also chapter 10.2 in the "Physikalische Metallkunde" of **P. Haasen**

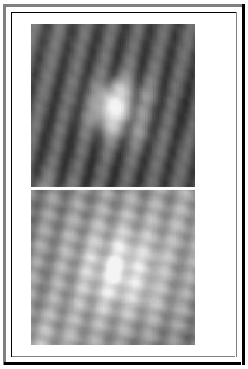
STM of Point Defects in GaAs

Illustration

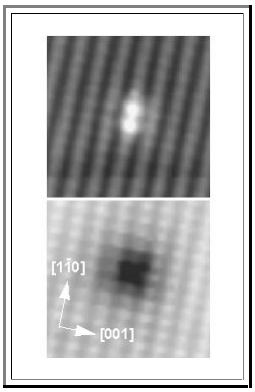
Here is a set of images coming from the work of the group of Prof. **Urban** (Research center KFA Jülich).



Various point defects are clearly visible; bigger defects, too. Below some defects at higher magnification:



Si donor in **GaAs**



Ga vacancy in **GaAs**

More information under http://www.kfa-juelich.de/iff/Institute/imf/imf-d.shtml

Values for the Formation Enthalpy of Vacancies

Illustration

- In this module a collection of vacancy data will be built up.
  - There might be several values for one and the same quantity, sometimes wildly different. Some colleagues would criticize the "uncritical" inclusion in a table like this.
  - Well, discussing incompatible values, always culminating at the conclusion that one's own measurements are superior to the measurements of the others, is one of the joys in the life of a scientist. And of course, there is only one value that can be correct *within the basic assumptions* (e.g. that we have just vacancies, not doing anything in particular except for migrating around a bit). So maybe some measurements were not so good, some evaluations relied on faulty assumptions - or, maybe, the basic premise of single vacancies is wrong.
  - Who knows - now. Eventually we will find out what is really going on. *This is the way science works* and students should be aware of this. The same comment made for the self-diffusion data in **Si** applies.
  - In **Si**, for example, the emerging point of view now (July **2001**) seems to be that you cannot simply consider just having vacancies, or vacancies and interstitials, but you must consider a complex system of **Si** vacancies, interstitials, oxygen interstitials, **C** substitutional atoms, and all kinds of recombination, pair formation and agglomeration phenomena that interact strongly and couple the point defect concentrations (oxygen precipitation, e.g. produces **Si** interstitial, **C** precipitation eats 'em up). Since the exact situation depends on many parameters, experiments may measure quite different values of just a single parameter - *and those measurements were perfectly correct*!

Element	**c_V at T_m** $\times 10^{-4}$ Various techniques		**H_F [eV]** from $\Delta l/l$ - $\Delta a/a$	**H_F [eV]** from positron annihilation		**H_F [eV]** from Thermopower	
Ag	1,7 - 5,2 17 - 24	Δl TE	0,99	1.31 0,89	L M	1,0	TP
Al	3 - 11 20 11 - 22 6	Δl TE C E	0,65	0,68 0,66	L M		
Au	7,2 14 40 7 5 - 20 3	Δl TE C Q E QM	0,92	0,89 0,89	L M		
Cd	5 -6,2 24 40	Δ TE Q		0,52	D		
Cu	2 - 7,6 13 50	Δ TE C	1,04	1,42 1,28	L M		
Co				1,34	A		
Cr				2,0	D		
In				0,56 0,55	L A		
Kr	3						
La						0,98	TD
Li	4						
Mg				0,9	M		

Mo	190 290 - 430	TE C		3,6 3,0 3,0	L M D		
Na	7						
Nb				2,65	D		
Ni				1,78	D		
Pb	1,7 20 - 23	∆l TE C	0,5	0,65 0,50	L A		
Pd				1.,85	D	1,7 1,5	TC TD
Pt	70 - 80 100 26 3	TE C Q QM		1,35 1,32	L D	1,45 1,45	TP TC
Ru						1,75	TD
Si			no values obtained	no values obtained			
Sn	<0,3 6 - 14 13	∆l TE C		0,54	D		
Ta				2,9 2,8	M D		
Tl				0,46	M		
V				2,07	D		
W	230 210 - 340 1 - 3	TE C QM		4,6 4,1 4,0 3,67	L M D QM		
Zn				0,54	A		

Some remarks

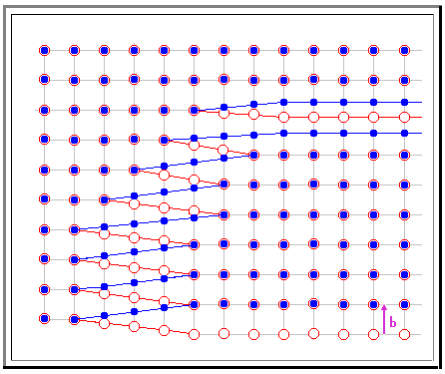
- There are more ways to obtain the formation enthalpy of vacancies from positron annihilation than just measuring the life time. In particular, measurements of
  - Angular correlations between the emitted γ-rays (abbreviated "**A**")
  - Doppler broadening (abbreviated "**D**")

  complement the lifetime measurements which also can be done in two modes (*lifetime spectroscopy* "**L**"; and *mean life time measurements* "**M**"
  Values given are from the compilation in Kraftmakhers book..
- Vacancies influence thermal conductivity "**TC**", thermopower "**TP**" and thermal diffusivity "**TD**" of metals, Clever measurements allow to deduce the vacancy formation enthalpy. Values given are from Kraftmakhers book.
- Vacancy concentrations at the melting point can be measured with various techniques. The abbreviations refer to
  - ∆***l*** **=** ∆***l*** **-** ∆***a*** method
  - **E =** stored enthalpy
  - **DC =** differential calorimetry
  - **Q =** quenching
  - **QM =** microscopic observations of quenched samples
  - **SH =** specific heat
  - **TE =** thermal expansion.

Movement of a Mixed Dislocations

Here is again the dislocation from chapter 5.1.2

- The picture is animated; the dislocation can be seen as it moves out of the crystal, thus reversing the cut-and-displace procedure that created it.



Advanced

Internet (and Other) Literature to Damascene (and Other) Techniques in the Production of Iron and Steel

Advanced

Following are the notes I took while reading through some Internet papers and which I will share with you.

Note: There is no guarantee that the direct Internet addresses always given are actually working. They worked, however, in May **2000**. The links in the headlines for each contribution therefore lead to the original version of the contribution *stored in a file* within this hyperscript.

- I do not want to infringe on copy rights. Since all articles are (or at least were) available in the the net, I assume that the authors actually want their papers widely distributed and read. Since this hyperscript is available in the Internet without charge, I trust that I did not violate any copyrights.
- Check also the following modules of the Hyperscript:
  Damascene Technique in Metal Working
  A cross-linked glossary of terms around the history of metal working
  History of Steel

But before we go into the Internet "literature", I must mention the wonderful book by **Manfred Sachse**: "*Damaszener Stahl - Geschichte, Mythos, Technik, Anwendung*" (Verlag Stahleisen, Düsseldorf; I'm told, it also exists in English)

- This book is not only only remarkable because it was written by an actual practicing smith, but because of the wealth of (first-hand) information it contains. Nothing directly available in the Internet comes close.

Steel in Ancient Greece and Rome

E. A. Ginzel

- Direct Internet Address: http://www.mri.on.ca/steel.html
- Note: the number in brackets in the paper do not refer to the list of references at the end of the article, as is common in scientific papers, but to some footnotes not included in the Internet version.
- Purports to show that the ancient Greeks and Romans knew more about steel that credited for so far. Includes a short but informative discussion about what steel is.
- Sees wootz steel as the source of the raw steel from about **500 BC** and describes two ways of how it was produced in India. Wootz steel made with the second method had an Fe content of about **1,5% - 2%**, this came close to the region of (gray) cast iron. The **C**, however, was precipitated as cementite (**Fe_3C**) and not as graphite as in cast iron.
- One of the main points in this article is that ancient smiths in the Mediterranean (in other words: Not the Indians themselves) found out how to forge this wootz steel into a material with "amazing strength and toughness" by hammering at a specific temperature.
- This forging technique also is supposed to explain the "swirl coloration" (otherwise known as "water pattern") of the Damascus steel and traces its origin back to **330 BC**. Damascene technique here thus does not rely on the weld forging of two different kinds of steel.
- The romans, however, did not exploit the Damascus steel technique or tried (or succeeded ?) in emulating the wootz steel production.

Metallurgical Heritage of India

S. Srinivasan and S. Ranganathan
Dept. of Metallurgy, Indian Inst.. of Science, Bangalore

Direct Internet Address: http://www.metalrg.iisc.ernet.in/dept/heritage.html

- A brief review of the major aspects of mining, smelting and working all major metals in antiquity with particular emphasize on India (which actually has a really outstanding record of early metal technology).
- A brief account of the importance of **wootz steel** for western technology
- Some interesting quotes from antiquity and a glimpse of how the efforts to unravel the mysteries of wootz steel and damascene blades prodded along western technology in the **19th** century.

WOOTZ STEEL: AN ADVANCED MATERIAL OF THE ANCIENT WORLD

S. Srinivasan and S. Ranganathan
Dept. of Metallurgy, Indian Inst. of Science, Bangalore

- Direct Internet Address: http://metalrg.iisc.ernet.in/~wootz/heritage/WOOTZ.htm
  - An expansion of the article cited above, unfortunately without pictures.

The Key Role of Impurities in Ancient Damascus Steel Blades

J. D. Verhoeven, A.H. Pendray, and W.E. Dauksch
Iowa State University, Materials Science and Engineering Department
This article appeared in the journal: JOM, 50 (9) (1998), pp. 58-64.

- Direct Internet Address: http://www.tms.org/pubs/journals/JOM/9809/Verhoeven-9809.html
  - The authoritative article about how to make "true" Damascus blade with wootz steel.
  - Quite informative, but does not comment on the properties of the emulated blades, however. (Could they be bent into a semicircle and so on?)
  - Stresses the role of trace impurities for the generation of "true" Damascus blades.

The Road to Damascus - Sorting Modern Pattern Welding from Myth and Legend

Kevin R. Cashen
Master Blacksmith, USA

- Published in the magazine "Sword Forum"; Direct Internet Address: http://www.swordforum.com/forge/roadtodamascus.html
  - The authoritative article about how to make "folded" Damascus or pattern welding today by a real blacksmith. One of the articles I enjoyed most.
  - Gives a good overview and does away with some myths surrounding damascene technique.
  - Comments especially on "quality vs. appearance". Damascene blades (in the meaning of folding or pattern welding) are **not** superior to good homogeneous steel!

WATERED STEEL, WOOTZ AND TRUE DAMASCUS

Lord Mikal Isernfocar called Ironhawk (???)

- Published in: The Vigil; Barony Of Middle Marches; Volume XVIII Issue V AS XXXII September 1997
  Direct Internet Address: http://www.ald.net/middlemarches/vigil/1997/V_9709.htm
  - Thank God, there will always be an England! (to quote the "new Yorker"). The article above, though published in surroundings probably not recognized as serious, peer-reviewed scientific journal, puts forward, without a trace of uncertainty, a completely different technique for producing "true" damascene blades!
  - "Low carbon wrought iron was hammered into very thin sheets. A stack of these sheets was wired together in a tight bundle. A batch of high carbon cast iron was heated until molten. The bundles of low carbon wrought iron were plunged into the vat of high carbon caste iron. The cold wrought iron would 'suck' the molten cast iron into the spaces of the bundle by a process called capillary action. This would partially remelt the wrought iron, 'welding' the bundle together into one solid mass. This mass was forgeable for a short time, so it was hammered into rough shape while it was still hot".
  - Otherwise, the article contains much of what has been stated elsewhere.

The Serpent in the Sword: Pattern-welding in Early Medieval Swords

Lee A. Jones
the persons who keeps a really interesting site in the Internet (try "home" at the end of the article)

- Direct Internet Address: http://www.vikingsword.com/serpent.html
  - Everything you want to know about pattern welding in the European history.
  - Explains in detail how the pattern emerge upon twisting and grinding.
  - Many pictures of real swords as well as very informative drawings.

Some papers to specific points

Ancient carburization of iron to steel: a comment
Donald B. Wagner
Department of Asian Studies
University of Copenhagen
Direct Internet Address: http://donwagner.dk

- Can you produce steel from wrought iron by heating it in charcoal? D. Wagner, citing a book from 1790 and doing some calculations thinks you can - refuting a 1989 claim from another scientist that you cannot! A nice illustration to the rapidly growing field of archeometallurgy!

Early progress in the melting of iron
V.H. Patterson and M.J. Lalich
This paper was presented at the 44th International Foundry Congress, held in Florence in 1977
Direct Internet Address: http://members.tripod.lycos.nl/cvdv/historycastiron.htm

- Short history in the development of cast iron from the Chinese to the tricks of the British in the 15th century. Did the British defeat the Spanish armada because of their superior iron technology?

Damascus Steel - A Brief History
Motoyasu. (Edited by WarAngel)
Direct Internet Address: http://www.angelic.org/highlander/metallurgy/damascushistory.html

- Very short and very concise! Contains most of everything, even the specific Japanese development.

Blade Patterns Intrinsic to Steel Edged Weapons
Several authors; the link brings you to a starting page, or use the
Direct Internet Address: http://www.vikingsword.com/ethsword/patterns.html

- Contains examples from all over the world (Bali, Japan, Philippine, India, ..). Otherwise like the "Serpent in the Sword"

History of Swords from Toledo
? Some agency to promote tourism
Direct Internet Address: http://www.intercom.es/espadas/history.htm

- Some Hyperbole about the great swords from Toledo

From Rapier to Langsax - Sword Structure in the British Isles in the Bronze and Iron Age
by Niko Silvester
Direct Internet Address: http://www.vikingsword.com/smithy/seax.html

- A short but rather clear history of the development of sword types (with some remarks to their making) in the British Isles.

"If it works, use it"

Motto of the old smiths
(and modern medicine men)

Damascene Technique in Metal Working

Advanced

A version translated into Romanian (by Alexander Ovsov) can be found in this link

A version translated into Indonesian (by Fira Widagdo and ChameleonJohn.com company) can be found in this link

A version translated into Russian by a "translator group" can be found in this link

A version translated into Hindi (by Dealsdaddy) can be found in this link

A version translated into Ukrainian (by UKessay) can be found in this link

A version translated into Punjabie (by the Bydiscountcodes Team) can be found in this link

A version translated into Danish (by Philip Egger) can be found somewhere in this link

A version translated into Dutch (by Arno Hazecamp) can be found in this link

A version translated into Sindhi (by Samuel Badree) can be found in this link

A version translated into Italian (by Ahsan Soomro) can be found in this link

Disclaimer:
I cannot guarantee for the accuracy of the translations nor for anything else for these links.

Helmut Föll

Far more about the subject can be found in the Hyperscript: "Iron, Steel and Swords"

A personal remark

- Before I started this page, I thought I had a sufficiently clear idea of what "damascene technique" meant: The forge-welding of **steel** and iron, or more generally, two types of steel. I also believed that this produced superior swords and mail, and - for obvious reasons - that this technique was pioneered in **Damascus** in ancient times.
- I also had a notion that this damascene technique was also used in **Toledo** (Spain) in ancient times, so when I visited Toledo in the spring of **2000**, I looked for some remnants of the famed Toledo sword smiths.
  - Indeed, there is a store selling swords, knifes and other metal stuff at about every corner. However, their merchandise are mostly "fantasy items" like the sword of Conan the Barbar, probably mass produced somewhere in the far east - you can find that everywhere in the world.
  - Then there was "artistic" stuff (e.g. ornamental dishes and plates) and especially jewelry done in what the Toledans called "**damascene technique**". What this meant was that some darkish metal was inlaid with silver or gold to obtain rich ornaments as shown on the right. It certainly was not what I had in mind when I searched for "damascene" technique

- Accidentally, I run across a shop connected to a real smithy[1)] - the last one left in Toledo, as the owner said.
  - There they made swords the old way; and with old he meant that they used to do this already for the *Romans* (so he said). At least, you could watch some rather special forging techniques and try the swords produced: They could be bent to a considerable degree without breaking or deformation - I actually bought one.
  - However, there was no damascene technique in this sword or in anything else in evidence; it was simply a solid piece of (hopefully) very good steel. Obviously, I got it all wrong, so I started to investigate a little. I chose the Internet rather than the science library because this is a "on the side" activity for me.

- Surfing the net for a few evenings, first created a tremendous confusion, because the word "damascene technique" seems to be used for many different things (see below). Now, I'm a lot less confused, but there are still some questions. This is no wonder, considering that steel was one of the main technical issues for about **2000** years all over the world, that its historical development would fill a small library, and that there ares still plenty of unresolved issues.
  - So some questions seem still to be open - there are no reliable answers or at least I couldn't find them. Interestingly, a lot of people, including serious "archeometallurgists" seem to share my interest - there appears to be an increasing number of publications and investigations in the last **10 - 20** years.
  - Most interestingly, even nowadays, steel technology seems to hold some mysteries and promises. And that brings us right back to the properties and the manipulation of defects in crystals.
- On this and some other pages I will share my confusion and my findings with you; as time progresses, *this material may become clearer*. I include a lot of documents, mostly found in the Internet, for those who want to investigate on their own.
- The material did become somewhat clearer, indeed. I include a few more remarks based on my present (May **2001**) understanding of damascene techniques: always indicated by a yellow triangle or button
  - But beware! Everything below or in the links thus represents **my** present knowledge and interpretations; it may well be wrong - take care!

Starting Point

- For me, the term "damascene technique" until recently had the following detailed meaning:
  - The manufacture of iron-based artifacts, especially knifes and **swords**, from *two kinds of steel*. You got it by hammering together (at high temperatures of roughly **800 ^{o}C** or so; called "forge welding") a package of several sheets of the two kinds. The sheets will fuse or weld as a result of solid state reaction and diffusion - a solid "compound material" is formed. The layered package of two kinds of steel is frequently folded over; the resulting structure is similar to the cross-section through a folded and twisted cake made from two different doughs (e.g. chocolate and vanilla).
    Not wrong, but only covering a small part of what is meant with "Damascene".
  - The two types of steels were
    1. *soft iron*, relatively low in carbon content, called ***wrought iron***; the basic product of early iron production by solid state reactions at temperatures well below the melting point of iron.
    2. *carbon-rich iron*, often from an source in India (that had a monopoly for many centuries) called "**wootz steel**".
    You may find some basic information about the development of iron and steel technology (including wrought iron and wootz steel) in the link.
    Mostly wrong.
  - The resulting sword combined the positive properties of the two constituents while avoiding the negative ones. It was hard, but not brittle, could hold a sharp edge, did not deform easily, but could be bent to a considerable degree.
    Mostly wrong.
  - This "damascene technique" was invented, or at least brought to perfection, in Damascus and Toledo in ancient times.
    Totally wrong
  - The old Celts, Germans, Vikings, Anglo-Saxons and so on, imported their damascene swords from the south (in exchange, maybe, for ***amber*** or blond women); or at least some raw materials.
    Totally and inexcusably wrong
- As indicated, this view has a few correct points, but it is often totally wrong; it needs to be modified and enlarged. In what follows I give a brief outline of my present (May **2001**) understanding (which includes a lot of open questions and most likely some misunderstandings, too).
  - On two other pages are (commented) lists of articles which I found interesting and a cross-linked glossary of some issues I was looking for in the Net. *Use with care*.

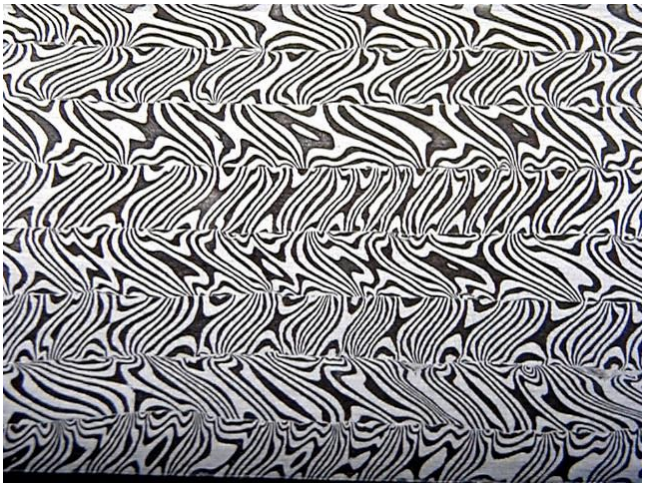
Some Variants of "Damascene" Technique

- As it turned out, "damascene technique" means quite different things to different people; but even within the defintion given above, there are many variants.
  - The "steel" part could consist of iron which is rich in Phosphorous and not necessarily Carbon (especially, maybe, in northern Europe?).
  - The **forge welding** could be done by folding over the same basic material which, however, may have been quite inhomogeneous. Lots of folding and forge welding created a homogeneous looking material - this is the **Japanese way** (horribly abbreviated).

- The welding technique was not only continued (and somewhat irregular) folding and hammering, but a more complicated technique, called "**pattern welding**". The result could look like this:

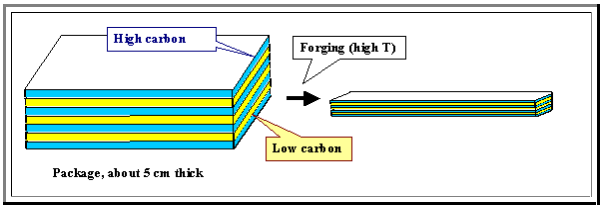
Part of a very old damascene (=pattern welded) sword blade. (from the Internet article "Blade Patterns Intrinsic to Steel Edged Weapons)" from Lee. A. Jones

- It also could look like the picture below. This is a photograph of a real piece of damascene steel recently made by the German master smith **Manfred Sachse** (whom we will encounter again) and taken from his Home page

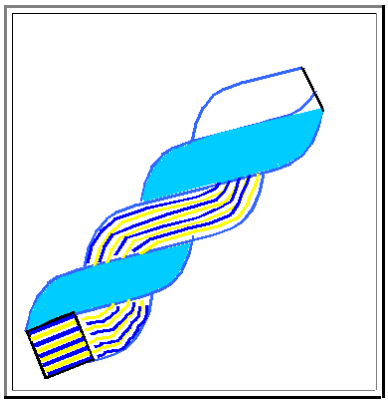


From the (by now abandoned) homepage of Master Smith Manfred Sachse with his gracious permission.

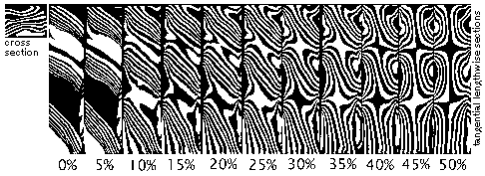
- In the "**Württembergisches Landesmuseum**" in Stuttgart I saw a very impressive sword from the time of the Merovingians (around **500 A.C.**) that was made by pattern welding (and found in **Ingersheim** - direct neighbour of the town of Geisingen where I grew up). This **sword from Ingersheim** was reproduced by the modern smith - **Manfred Sachse** mentioned above - as follows:



- The loose stack of steel plates is banged into a rod with a cross section of about **1cm^2** - some work! Several of those rods, about **1 m** long are produced. The labelling "High carbon" in the drawing could, perhaps, also mean "high Phosphorous".
- Next, these rods are twisted and ground flat on two sides. The twisting is hard to draw, but you get the idea.

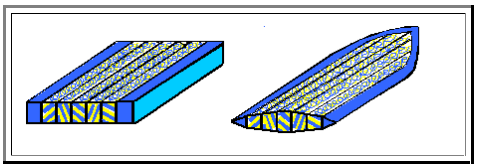


- What it looks like on the surface if you now grind the twisted rod to increasing depth is this:
  (From the Internet article of Lee A. Jones: The Serpent in the Sword: Pattern-welding in Early Medieval Swords)



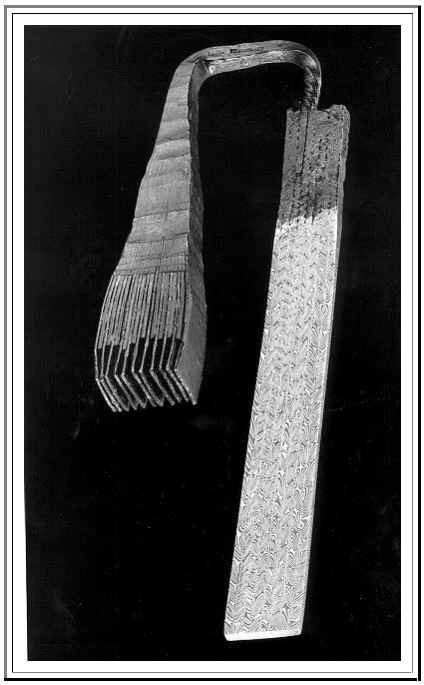
Original Fig. Captions:
Patterns disclosed by successively grinding a facet along the length of a twisted rod demonstrated in a clay model of a rod composed of sixteen alternating layers prepared by bladesmith Dan Maragni. The rod has been mildly squared and joined along side another rod, shown only focally at the edge. The rod was progressively ground and photographed at each interval, reduced in overall thickness by the percentage shown in the scale below. Further leveling of the rod will reverse this trend, as a mirror image of the patterns first disclosed.

- Several of those rods were than forged welded; with possibly a pure steel rod on the outside. Banged into shape, and ground to a sharp edge, we have a fine sword, it represented about the value of a car in todays currencies.



- only it was even more complicated: Two independent layers were used for the center part, so that the front and backside od the sword looked different; and the twisted regions were alternated with non-twisted regions to form specific pattern down the lenght of the sword. Well, look at it yourself.
- This is probably as close as you can get to a **magical** or simply **famous sword** like **Notung** (**Wagner's** sword for **Siegmund** and **Siegfried**), **Excalibur** (**King Artus**), **Balmung** (What Siegfried made from the Notung pieces in the **Nibelungen saga**, **Tourendal** (**Roland** saga), **Mimung** ("**Wieland der Schmied**" made it for his son **Wittich**), **Eckesachs** and **Nagelring** (**Dietrich von Bern**), **Colada** and **Tizona** (**El Cid)** - and so on.
- More about "magical swords" can be found in this (German) link.

Alltogether, in a model showing all process stages and also on display in the "Landesmuseum" mentioned above, it looks like this:



- The picture is from the wonderful book: Manfred Sachse: "Damaszener Stahl - Geschichte, Mythos, Technik, Anwendung" (Verlag Stahleisen, Düsseldorf) and reprinted here with permission of the author.

It seems that pattern welding and **P**-rich steels were especially popular in northern Europe; but some kind of "damascening" or pattern welding can be found all over the world.

- It would be totally wrong, however, to credit ancient smith with the invention of something very sophisticated. The truth is: They had no choice but to come up with some kind of pattern welding!
- The reason is that nobody could melt wrought iron or mild steel with a melting point of **1550 ^{o}C** during the first **2000** years of iron technology. Only **cast iron** (eutectic melting point at about **4% C** is **1130 ^{o}C**) could be molten (and was used in large quantities in ancient China)
- Everybody had to work with small lumps of iron out of a "***bloom***" obtained by a solid state reaction. This small lumps needed to be forge-welded, i.e. banged together at high temperatures, to obtain large pieces. Invariably, the little lumps had on occasion different **C** or **P** content; the forge welded blades showed some structure. Iron blooms obtained in different regions from different ore deposits also would be different; with a little trading it could not escape notice that forge welded parts showed structures, and that some parts were hard and others soft.
- It is then a small step to first forge-weld some kind of iron to relatively homogeneous stuff, then some other kind (easily distinguished by color or hardness, produced in some special way, or traded from some other smiths) - and having two kinds of iron plus knowing about forge welding, pattern welding is something that does not need a big innovation.

Even so, it took almost **1000** years of forge welding and simple pattern welding before welding reached its zenith around **700 - 800 AD**, producing extremely complicated and certainly very beautiful and valuable works of art (the performance in real fights was probably no better than that of simpler swords, however).

- And, to be clear, the whole process was not simple at all! It took a lot of knowledge, experience and practice, to produce a "good" pattern-welded sword! Those ancient and medieval smiths were not barbarian brutes but highly educated and skillful man!

First questions come to mind; some answers are contained in the commented list of articles

- Who did it when (and how)? Which cultures just copied, and which ones invented or improved?
- Were those pattern-welded swords really much better than "regular" ones? Or was the whole thing more for show, a status thing? Was damascening or pattern welding a major innovation or something you couldn't avoid discovering?
  Crude Answer: The better pattern-welded swords were superior to swords from plain iron (or soft, inhomogeneous steel), but inferior to swords from good homogeneous steel. See the table below for data on "true" damascene swords.
- What were the ingredients? How where they obtained? How did different types of starting materials influence or determine the forging process and the final result?
- What exactly was the role of Damascus or Toledo?
- What exactly were the famous Damascus blades? How were they made, and how good were they really?

The "True" Damascene

- The last question seems to have an answer:
  - "True" Damascene blades were made from **wootz steel** only. The Damascene (or water) pattern comes from a striated precipitation of **Fe_3C** particles and not from folding and welding two kinds of material.
  - The "secret" art was how the high carbon wootz steel (coming close to cast iron) was treated to yield a highly flexible and extremely sharp blade - check in the commented list of articles.
  - It now appears that it was cricial to have traces of Vanadium (or somethimg similar) to enable proper nucleation of the **Fe_3C** particles - see the latest article to this subject
- What seems less clear, however, is how good these blades really were. Obviously, the crusaders, wielding quite respectable swords themselves, were mightily impressed.
  - Trying to forge similar blades lead European sword smiths astray, however. They believed that these blades were composed of two types of steel and re-invented the "old" pattern welding technology in new variants - seemingly without much success. The explanation given above seems to be a pretty recent discovery!
  - How good "true" damascene blades were was something an early metallurgist actually did find out to some extent. Prof. Zschokke (from Switzerland) was lucky enough to get a few true damascene blades for (destructive) investigations (this is quite unusual because these blades ar valuable and museums and collectors do not easily agree to have some destroyed).
  - Some of this results (taken from the book of M. Sachse) were

General composition					
Sample	**[C]**	**[Si]**	**[Mn]**	**[S]**	**[P]**
1. Knife	1,677	0,015	0,056	0,006	0,086
2. Knife	1,575	0,011	0,03	0,018	0,104
3. Saber	1.874	0,049	0,005	0,013	0,127
4. Saber	0,569	0,119	0,159	0,032	0,252
5. Saber	1,324	0,062	0,019	0,008	0,108
6. Saber	1,726	0,062	0,028	0,020	0,172
7. Modern welded steel (Solingen)	0,606	0,059	0,069	0,007	0,024
8. Modern cast steel (Solingen)	0,499	0,518	0,413	0,038	0,045

Properties						
Sample	**3**	**4**	**5**	**6**	**7**	**8**
Bending toughness	13,4	15,2	11,5	14,5	21,6	30,0
Work to bend	94	221	55	63	361	622
Angle of bending	27	59	19	17	69	78
Hardness	216	233	193	248	347	463

  - What ever the numbers mean (no units were given), the modern blades always "win". Otherwise blade **No. 4** is best. In any case - the properties of what was (and is) traded as "true" damascene vary widely, there are very good and very lousy specimen.

The new "High-Tech" Damascene Technique

- The word "Damascene Technique", if uttered in a gathering of materials scientists dealing with functional materials, will carry yet another meaning, completely different from everything above.
  - Here we mean a special technique for the production of structured Cu connections - in the **0,5 mm** range - on **Si** integrated circuits. Check for yourself in the Link.
  - The naming "in honor of the metallurgists of old Damascus" is a bit misleading, however, because the damascene chip technology is related to what is called damascene today in Toledo.

All the Meanings of Damascene Technique at a Glance

- The expression "Damascene Technique" thus has a lot of different meanings. In the listing below I include some techniques that are not "officially" listed as damascene, but follow the general idea. The adjectives used in differentiating the diverse techniques are mostly my invention
- **Folding Damascene**; two kinds of steel:
  - Folding over a stack of different steels several times; gives many layers with beautiful, but irregular patterns. The kind of damascene that many modern smiths do today.
  - This technique was to some extent re-invented in the West after encountering "true" Damascus swords in an attempt to emulate these famous weapons
  - Many myths abound. In truth, the finished product has a rather homogenized **C** content (i.e. is not a real "compound" material from two different steels and no better that a homogeneous blade from good steel). But the damascene pattern obtained after suitable treatment (etching) gives this Damascus steel a special beauty and appeal. It does not so much reflect the different C concentration of the layers, but probably (I'm not so sure about this) the different amount of other impurities, especially P (which, supposedly, does not diffuse as fast as **C**).
- **Folding (Damascene)**; one kind of steel
  - What the Japanese did to get sword material. Not usually called Damascene, but not so different, because the small lumps of iron or steel selected from a bloom were, after all, rather different in composition and contained slag inclusions and other inhomogeneities. The speciality of the Japanese was a lot of folding and hammering; the finished product therefore does not show evidence of folding to the naked eye - it is now quite (but not totally) homogeneous on the outside.
- **"Simple" pattern welding or laminating**
  - This could be, e.g. just some mild steel in the middle and hard steel at the edges; or a core of soft iron surrounded by harder stuff. What (maybe) the Romans had and the early Celts. There is no particular pattern besides the simple geometry of the design.
  - The Japanese also used this technique; their swords consisted of up to three different kinds of steel (each one obtained through multiple folding as described above) welded together or laminated in quite complicated arrangements.
- **"Decorative" pattern welding** - like the technique shown above.
  - While the twisting had also technical advantages compared to forge welded untwisted rods, its main purpose - at least in later - times, was the decorative effects possible with this technique. Designs much more complicated than the one shown above were in use.
  - In later times - lets say around **1000 BC** - when smiths had learned how to make swords form homogeneous steel (especially in Toledo, it seems); the blade may still have been adorned with a thin layer of pattern welded foil only for the look of it!
- **"True" Damascus**
  - Swords and other implements made from one kind of steel - the famous **wootz steel** - obtained from Indian sources from sometime before **300 BC** up to the **7th** century **AD**. After that, the people in Damascus, in Toledo and probably other places also, could produce this high-carbon steel themselves.
  - Treated the right way, **Fe_3C** (cementite) forms in striations, producing the special "damascene" pattern (often referred to as "water pattern", too). These were the swords of tall tales that emerged when the crusaders met the arab owners of these beauties.
  - One recent scientific paper reproduced the ancient technique successfully and claims it needs three thing to produce "true" Damascus sword:

    - The right combination of time/temperature firing during ingot making
    - the proper thermomechanical sequencing during the forging process.
    - and the right chemical composition (especially minor element additions, e.g. V in sufficient concentrations

We are right back to point defects in crystals!

- At least one more modern smith works on "true" damascene made from wootz steel (he send me an e-mail). You may jump to an article about him and his art by activating the link. (if the link does not work any more, here is the stored version)

"Mysterious" Damascus

- There are some people out there, who honestly believe (more or less based on a scientific background) that everything was either quite different or that the technology is truly lost.
- But then there are also those, who cook up some pseudo-scientific bullshit including some magic - usually in the attempt to sell their "magic" product.

"Inlay" Damascus (what they sell in Toledo)

- While this is certainly a technique to intimately combine two metals (not necessarily by forging, but e.g. by soldering); it is not a technique usually associated with the making of swords, knifes, armor or other "functional" products. It may have been used in Toledo for adornments of the sword hilts, though.

"Microelectronic" Damascene technique.

- "Damascene technique" (even "double damascene") has become a common name in microelectronic technology; everybody in this business knows what it means.
- It has, however, nothing to do with all the variants given above that could produce a sword, but is a kind of "inlay" damascene technique, albeit on a **<1µm** scale.

Whow!!

How wrong can you be? But then, how confusing can it be? My "quick" attempt to figure out exactly what "Damascene" really means, in order to include nothing wrong in this (rater unimportant) addition to the "Defects" Hyperscript, took several evenings and weekends! But there were rewards: I learned a lot of very interesting things about the history of technology, including some points which I always wanted to know a little better. Then there were a few unexpected but rather interesting finds:

- Most of the serious knowledge comes from recent to very recent times. A whole new field of research is developing: **Archeometallurgy**! Some of its findings already changed the way we look at ancient history, and promises are that there is much more to come.
- There is a lot of interest in these issues out there - at least in the anglo-american world. Try any search engine with keywords like "steel" "Damascus" or "swords" and you will get an overwhelming response.
- Then try it with the German equivalents: Essentially you will end up with Karl **May** and the bible. This can be seen as a comment on the attitude to technology in these cultures!

---

1) Here is the name and the address of "my" smith:

Mariano Zamorano; Fabrica de Espadas y Armas Blanca ; 45002 Toledo; C/. Ciudad, n.º 19

Discovery or Invention?

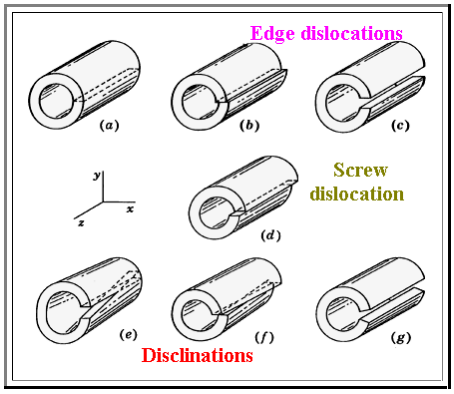
Advanced

- Did the fathers of quantum mechanics, of statistical thermodynamics, of the theory of relativity, of dislocations as the source of plastic deformation, and so on, did they **invent** their theories, or did they **discover**, did they *find* them?
  - Are all those theories simply human inventions, intimately tied to carbon-based life on this planet, or are they absolute, invariant truths completely independent of the existence of humans?
- The answer, somewhat surprisingly for **scientists**, does not seem to come easily to the **philosophers**, who are the people who worry about those things.
  - *Invention* means that the laws of nature are nothing but an outgrowth of human activities; other thinking beings at other places or times may invent completely different systems fitting their peculiar needs. At best, we may come up with some approximation to something intrinsically intangible, because there are no absolute truths. This statement, of course, must be an absolute truth, which opens a different can of worms labeled "**Gödel's** theory".
  - *Discovery* means that the laws of nature exist in a defined form, totally independent of humans or anybody else below the level of an almighty being, and that there is a possibility to discover them in total (if there is a finite number of natural laws) or at least in parts and to describe them in some language (including the language of mathematics). Maybe we find only parts, or we see the laws coarse-grained (i.e., in some approximations), but it is out there to be discovered.
- We move quickly to **metaphysics** this way, to the *theory of science* with all its changes, developments, and idiosyncrasies.
  - And if you think that there can be little doubt that we scientists *discover* truths and do not *invent* them, you should take note that this position is in total opposition to the current beliefs in modern philosophy, especially in the branches known as "post-modernism" or "positive realism".
  - If you are interested in this, read, e.g., **John Horgan**: "The End of Science", which gives a well written, if not outright exciting account of the various metaphysical developments in the last **100** years or so.

Volterras Tubes

Advanced

- How can we obtain an arbitrary deformation of an arbitrary body by just repeating and combining some basic deformation prodedures?
- The illustrations shows Volterras answer to this question: Take a cylinder of a material, cut it along some wall, shift the surfaces of the cut in all ways that - after welding the walls together again (including taking out or adding material) - will lead to different deformation states.
  - As Volterra showed, there is a limited and rather small number of possible independent cuts + shifts. All other cuts plus some deformation can always be expressed as a linear superposition of the elementary cuts.
  - Here are the elementary cuts. The first one just shows the cut, the next three ones correspond to dislocations - i.e. a real dislocation produces exactly the strain field generated by the cut and shift procedure.



  - The last three cuts corresponds to special defects called **disclinations** that are more elementary than dislocations, but are not observed in real crystals (except, maybe, in grain boundaries). They do however, appear in two-dimensional lattices, e.g. in the **flux-line lattice** of a **superconductor**.

A Brief History of Steel

This moduel is also availabe in Romanain languga, thanks to **Irina Vasilescu**. Here is the link

This is a much embellished translation of an earlier version written in German (it can be found in the Hyperscript "Matwiss I") and with some footnotes added later.
If you really want to know about the history of iron and steel use this link.

Advanced

In order to make steel not accidentally, but conscientiously, you obviously first need to make **iron**. In contrast to the noble metals like gold, silver or platinum (and the occasional find of pure copper), iron is never (?) found as an element but practically always as an oxide.

- However, in contrast to other metals found as oxides (especially **Cu** and **Sn** oxides needed to make bronze), the temperature of a "normal" fire is not sufficient to reduce iron oxide *and* to make the elemental iron liquid - the melting point of iron is T_m**(Fe) = 1535** ^{0}**C**; far above the **(1000 - 1100)** ^{0}**C** that the ancients could produce (?).
- For **Copper** (**Cu**), e.g., it is different - its melting point is T_m**(Cu) = 1083** ^{0}**C**. Throw some copper minerals in a nice hot fire made with plenty of charcoal (producing **CO** which is great for reducing oxides), and liquid copper will result almost automatically.
- This happened *and* was noticed probably a good **6000** years ago, when early potters tried to adore their pottery with nice green **malachite** - a copper mineral known in antiquity and used as a gem stone. What a surprise, when one day in a particularly hot fire, instead of decorated pots they found an ingot of pure - and then extremely precious - copper in their oven. Copper was otherwise only found in small quantities (much less frequent then the (then) ubiquitous gold) in mountain ranges and river beds.

This was a decisive discovery for mankind: Precious and shiny metals could be made from dull stones. Things could be changed from one seemingly immutable form into a completely different one - **alchemy** has its roots right here, and the yearning for "transmogrification" has never stopped since.

- Early metal industry and the short-lived "**copper age**" began to be replaced rather soon by the **bronze age (Cu + (5 - 10)% Sn** and often some **As**); and the bronze age lasted more than **2000** years (it was not abruptly replaced by the iron age, but coexisted for about **1000** years).
- Here we first encounter the importance of impurities: A little bit of **As** as an impurity atom makes bronze "harder", it doesn't deform so easily any more. Of course, nobody knew this. All that was probably known was that some sources of copper and tin ore, together with all kinds of tricks (including some magic or prayers, of course) produced superior bronze.
- It is quite natural that tin and other metals were discovered shortly after the momentous discovery of copper ***smelting***. Once you saw that precious copper could be made from some kind of rock, everybody not completely stupid would of course try what you could get with other rocks.
- We also have the beginnings of an environmental disaster, because for metal smelting you need tremendous quantities of **charcoal**. First in order to obtain high temperatures but, just as important, for reducing the metal oxide according to


Auf Köhlers Spuren

Bis ins 16. Jahrhundert war Schleswig-Holstein ein Waldland. Dann wurde abgeholzt, um unter anderem Holzkohle herzustellen. Wie das geht, zeigt Stefan Brocke aus Sophienhamm bei Rendsburg jetzt im Loher Forst. Seite 6

From the "Kieler Nachrichten", front page, one day after after I wrote this paragraph. It says:
On the Track of Charcoalers
Up to the **16th** century, Schleswig-Holstein was woodland. Then the trees were felled to produce charcoal (among other things). How that is done will be demonstrated by Stefan Brocke in the Loher woods.

$$MeO + C \Rightarrow M\ e + CO$$

- About **100 kg *charcoal*** are needed to smelt **5 kg** of copper.

- Besides shipbuilding, charcoal production is responsible for the disappearance of large parts of European forests (the disappearance of ***yew*** trees (which were ubiquitous in antiquity) from present day forests, by the way, is due to the middle age bow-and-arrow industry - nothing beats a yew bow!). Charcoal production was a major industry and the source of the many ***charcoaler*** ("Köhler") stories in fairy tales and folklore.
- Beside **Cu** and **Sn**, **Pb, Hg, Ag**, and of course **Au**, were known and produced on an industrial scale - especially by the Romans. But the Romans (and the Chinese, and the Indians, and the ...) had also **Fe** - but still no fire hot enough to melt it.

Early experience with the smelting and melting of other metals did not help in producing iron - it first came into use about **1000** years later than bronze. This must have been a kind of puzzle, because the ancients *did know* that iron existed. It was extremely rare and precious - because it fell from the sky in exceedingly small quantities.

- **King Tut**, matter of fact, had a little iron dagger made from meteorite iron right on his breast - obviously his most precious object. In old Sumeria, iron was called "*sky metal*" and the pharaohs in old Egypt knew it as "*black copper from the sky*".
- Of course, only pictures of his less precious and useless but more showy gold dagger are easy to find. The picture on the right shows both.
- The **Eskimos** in **Greenland**, matter of fact, made their iron tools for hundred of years from a large (**30** tons) meteorite.
- Some American explorer (**Admiral R. Peary**) finally stole it (he wouldn't have expressed it that way, though) in the **1890**s and had a hard time to transport it to the Natural History Museum in New York. Here it is:

King Tut's daggers (Internet source "Stacey")

Meteorite stolen from the Eskimos

We may safely assume that the old materials scientists tried everything to smelt iron from suitable stones. They did have tricks to raise the temperature of a fire - in a **4500** old mastaba in Egypt, I took a picture of a relief showing six gold smiths (probably rather their **Ph.D**. students) blowing into the fire with hollow reeds. But just blowing with lung power will not do the trick for iron - maybe you get **1200 ^{o}C**, but that's it.

- So in a typical fire with temperatures well beolw **1500 ^{o}C** you do not get *liquid* iron - but you do get *solid* iron because reduction does take place - in a solid state reaction. What you get is an **iron bloom** ("Eisenblüte" in German), a mixture of fine iron particles, unreacted iron oxide, slag and charcoal residue. Here is an actual picture of some ancient bloom (from around **600 AD**; I actually "found" this myself (in some museum).

- The iron in the bloom was rather pure (and thus comparatively soft) because a solid state reaction produces *only* iron - carbon or other impurities have to diffuse in from the outside (if the iron would be liquid, it would just dissolve the dirt up to the solubility limit).

The early iron smiths (probably being **Hethites** of some form) could "wring" the iron from this bloom by separating the iron from the rest mechanically and repeatedly hammering together what was left at high temperatures (about **800 ^{o}C**; some of the slag then is liquid and gets squeezed out) with, no doubt, proper prayers to the respective gods and many (magical) tricks.

- What they finally obtained was "**wrought iron**" ("Schmiedeeisen"), i.e. a lump of rather pure iron consisting of small pieces welded together, with plenty of small inclusions (small, because of the hammering that breaks up large pieces of slag).

- Extreme care was necessary - from the selection of the iron ore, the reduction process and the hammering business. If you were careless, the iron oxidized again (it really "burns" at temperatures in excess of about **800 ^{o}C**), and if you kept your reduction process going too long, carbon diffuses in and you may end up with **cast iron** (**C** content about **3% - 4%**; melting point as low as **1130 ^{0}C**). Then you actually got it liquid - "casting" was possible - but cast iron is brittle and useless (for weapons, that is).
- Somewhat later, with larger furnaces and increased experience, the bloom obtained may have contained some high-carbon melted parts on its top layer. It then consisted of a whole range of iron-carbon alloys - from rather pure wrought iron to cast iron with *good steel* - say **0,5 % - 1,5%** carbon - in between. The art of the smith than included to pick the right pieces. This was a highly developed skill, we know about it especially from Japan; but that does not mean that the Kelts or others did not do it just as well.

But beware. The art of making iron and steel, developed over **2000** years in many civilizations, cannot be contained in a few lines, not to mention that very little is known about that story - iron, after all, rusts (see the link showing an old sword), and not much has been found that gives detailed knowledge about how the old romans, Indian, Chinese, etc. made their steel and iron products.

- Nevertheless - the early smiths, starting with the Greek god **Hephaistos** (the roman Volcanos) and containing many fabulous figures like the Nordic "Wieland the smith" or "Mime" in Wagners "Ring des Nibelungen", could produce articles, especially swords, from the iron bloom that were much better than the customary bronze stuff (and than of course "Magical" swords). In other words, they sometimes succeeded in making good steel.

What was their secret? It is rather simple - looking at it retrospectively: You need the proper concentration of **C** in the **Fe bcc** lattice at room temperature (some other impurities are helpful, too; while others - especially **S** and **P** - were harmful). Raising the about **0,1% C** in wrought iron to an optimal **0,7 -0,9%**, raised the hardness (or better the yield point) threefold! But if you got too much - say **2%** - you were on the road to brittle cast iron not useful for swords.

- Not being able too melt iron (and thus not being able to throw some magical stuff into the brew) the only way to get carbon (or on occasion **N** which also "works") into the **Fe** lattice was diffusion via the surface. What you needed to do was to "roast" you iron (possibly the whole sword) for the right time at the right temperature in a charcoal fire. Magic and praying helped - it did indeed: How do you keep track of the time without a watch? You utter a long prayer that you learned from your master - the right ones "worked"! The rest of the magical ritual was helpful in providing reproducible conditions.
- Of course the old practitioners had no idea of what the really were doing; if they thought about it, they felt that were purifying the iron in the (more or less holy) fire. This erroneous believe (like so many others) goes back to the (from a materials science point of view somewhat questionable) philosopher **Aristoteles** who certainly asked the right questions about life the universe and so on, and is righteously famous for that. His answers, however, were invariably wrong - even in the few instances where he could have known better.
- Well, we have made but the first step to steel. We now must make a few more steps for good homogeneous steel - or we delve into a fascinating world of its own, the various **damascene techniques**, one of which is blending different kinds of steel into a compound material. More to that in the link.

Here we look first a bit on what happens in heating up and cooling down your material. *We* know, after all, that going up in temperature, iron changes at **910 ^{0}C** from the **bcc ferrite** phase to the **fcc austenite** phase.

- Carbon feels much more at home in austenite - its solubility is higher than in ferrite. If the smith kept his iron in a good fire very long, he now might have had a rather carbon rich austenite in the outer layers of his sword. So what happens upon cooling down?
- Well, it depends. If the iron cools down s l o w l y, the carbon rich austenite will change to carbon rich ferrite. If there is more carbon in the austenite than the ferrite can dissolve, carbon will precipitate, forming a new **Fe - C** phase called **cementite** (with a quite complicated lattice). We now have cementite particles in **fcc** ferrite; usually in a very typical structure - both phases appear like a stack of plates. This kind of structure is called **perlite** because, looking at it under a microscope, it has a luster like pearls..
- Perlite, the mixture of ferrite and cementite, however, is not much better than bronze as far as its mechanical properties are concerned. So you must prevent the phase change from austenite to perlite if you want to keep your sword "magic"! In other word, you must not allow enough time for the carbon atoms to diffuse around during cooling as would be necessary for forming precipitates. In other words: *You must cool down rapidly* (hopefully you did the proper exercise for calculating how fast you must cool down).

Here we have the next *big trick* - after making bloom, extracting wrought iron, and carburization: **Quenching** - often the big secret of master smiths (there is a whole Japanese mythology to this subject). The hot sword is stuck in a liquid for some time and thus quenched - and only very unimaginative smiths would have taken common water at room temperature for that.

- If the cooling time was too short to allow **Fe-C** precipitate formation, we now have a supersaturation of **C** in the ferrite phase which then will have a strongly disturbed lattice structure. A kind of mixture between **fcc** and **bcc** phases will prevail which has its own name: "**Martensite**".
- *Now you did it:* Martensite has the fivefold "strength" of wrought iron!
- Unfortunately - if you got martensite at all, it tends to be brittle! Now the *next bag of tricks* is needed: Heat up your sword again - but keep the temperature moderate.

- Some of the *defects* that make martensite brittle anneal out and its ductility goes up. Bang it (i.e. deform it plastically), and you produce *dislocations* (hey, that's were we started from some time back!). Now you are manipulating a second kind of defect for optimizing mechanical properties!
- But now we stop (so does the smith). If you *really* want to know much more about this, use this link.
- Anyway, if everything worked, you now have a very good (and of course magical) sword which was far superior to the bronze stuff of your opponents. In particular, you could make it *longer* without having to worry that it might break in battle (which was about the worst health hazard imaginable then).

And don't think that an increase in strength by a factor of **4 - 5** is not all that much. The old Gauls, **Asterix** and **Obelix** notwithstanding, were conquered by the Romans not least because their swords bent and needed straightening (over your knee) after a forceful blow - something the Roman swords did not need. (Haha - don't you believe all this Roman propaganda!)

Well, making a good steel sword was lots of work, lots of knowledge, and lots of luck. Considering what could go wrong, it is quite remarkable that the old smiths actually did produce superior steel swords now and then. Of course, probably more often then not, only the outer layer was steel, while the inside was still soft wrought iron - the sword was made from compound materials, in fact.

- This gives us (and possibly also the old smithies) the idea of doing that from the start: Weld together soft and hard layers, carefully picked from the bloom or made by carburization, and hope that the result will combine the positive properties of both materials. We are talking *damascene techniques* here.
- However, the word "**damascene techniques**" is a collective identifier of several very different technologies. Most people associate it with a kind of compound technology where two different kinds of steel were put together in layers and then forged into a sword or whatever. While this is something that was done - especially by the **Kelts** and other North Europeans - it was *not* what the guys in Damascus did, the purported source of the famous damascene blades.
- As far as we know today, the "true" damascene technique actually worked with a famous kind of steel, so called "**wootz**" which was produced in **India** for maybe a **1000** years in a kind of closely guarded monopoly. Wootz was rich in carbon (about **2%**; there was a secret carburization technique) and the trick was to precipitate the surplus carbon in a pattern of fine **FeC_3** precipitates.

A fascinating world unfolds behind the catch word "damascene technique", if you like you can browse the following links

- Damascene Technique in Metal Working
- Literature to Damascene (and Other) Techniques in the Production of Iron and Steel From the Internet
- A Cross-Linked Glossary of Some Terms from the History of Metal Working

Steel technology was not confined to the Mediterranean and the European North West. India may well have been at the apex of steel technology and **China** had its own technology centered around **cast iron**, used not so much for warfare but for civil objects like pots and pans.

- And lets not forget the **Haya**, a people who lived in what is now **Tanzania**. They had a highly developed **Fe** technology and used it for beautiful sculptures, too. Their myths and fairy tales contain many stories relating to the making of iron, using a vocabulary that was heartily enriched with expressions relating to the making of humans.
- There is even some evidence - collected recently (and, of course, being discussed controversially), that the old Africans had the highest temperatures of all, even reaching the melting point of iron some **2000** years ago (long before everybody else did)

Whatever happened whenever and wherever, during the millennia, and despite the many difficulties, iron and steel became common materials. At some time in the middle ages or Renaissance, the melting temperature could be reached, but the mass production of good steel still had to wait for the **19th** century. Before, only "thin" objects - the paradigmatic "sword" or katana, scimitar, saif, shamshir, tachi, tulwar, yatagan,.. - could be made by in-diffusion of carbon.

Charcoal was replaced in the **17th** century with coal, but not without unpleasant surprises. Iron that was smelted with coal instead of charcoal was very brittle and completely useless. We now know, of course, that minute amounts of sulfur in the **Fe** lattice - it segregates in grain boundaries - are sufficient to make **Fe** brittle, and **S**, like other harmful impurities, is contained in regular coal in rather large concentrations.

- The solution to this problem, surprisingly, did not come from the military related strata of society, but from the second most important enterprise dear to the hearts of men: **beer** brewing. Brewers had tried to use coal instead of charcoal for roasting the barley - and produced a stinking abominable brew. Thusly coke was invented: Roast coal in an environment deprived of oxygen - the stinky stuff will evaporate and what remains is clean carbon - called **coke** - which could not only be used to brew beer, but was also usable for the iron smelting industry.

- The beginning of the industrial revolution was severely hampered by the lack of a large-scale process for the production of good steel. (Just imagine how the **Si** revolution would have fared without large dislocation free and rather perfect **Si** crystals). The (at least in German and French) paradigmatic **Eisenbahn** (chemin de fer in French), the rail road, needs rails; with regular wrought iron or cast iron the rails had to be renewed every **6** month because they deformed under the load (or cracked). Accidents were frequent and often catastrophic.
  - The production of large amounts of iron was common by then - the essential part was blowing large amounts of air into the fire with the aid of mechanical bellows powered by steam engines. The leading British production accounted for **2,5** million tons of iron in **1850**, but the production of steel was still a cumbersome and expensive business, accounting for a few percent of the total production.
  - It was also known for sure since 1**786** that steel had something to do with carbon; the first person suspecting this was one **Tobern Bergmann** in **1774** (other sources, however, refer to **Vandemonte**, **Berthollet** and **Monge** from France).
  - Still, all efforts to produce iron with the proper carbon content (and the right structure) "from scratch", were in vain. Sometimes things worked, sometimes they didn't - there was no large-scale, reliable, and reproducible process. And thus no big bridges, sky scrapers, safe railroads, big ships, efficient engines, and so on - one rarely reflects how much cheap steel changed the world!
- This time, however, progress came from the military industrial complex. It became simply too embarrassing that the big canons (made from cast iron) had a tendency to explode. Something had to happen.
  - It was Henry **Bessemer** who was especially interested in good steel for big canons, because he had just invented a new kind of projectile that received some spin even from smooth bore guns (and thus was harder to destabilize during flight). Unfortunately, the canons couldn't take the additional pressure building up while the projectile was building up spin as well as speed- they exploded more than ever. So Bessemer was looking for large amounts of cheap steel.
  - He was then the first person (so it was believed for a while) who had the genius idea of making steel by getting carbon *out* of cheap, carbon rich *cast iron*, instead of using the cumbersome way of getting carbon *into* low-carbon wrought iron. The way to "*drive out*" the surplus carbon was to blast large amount of oxygen through the cast iron melt (which, by the way, definitely needed the **steam engine**; quite hard to do this through a reed). **CO** will form in the melt which not only burns off to $\mathbf{CO_2}$ upon hitting the air, but by doing this supplies the heat to increase the temperature of the melt because the melting point will go up with decreasing carbon content. If you stop at the right time (looking at the color of the flame), you will be able to adjust the carbon content of a large amount of iron to just the right value and thus produce large amounts of good steel.
- Mr. Bessemer, who was not exactly unknown before (he already had some fame as the inventor of the "lead" pencil (which in reality contains graphite), after publishing his finding on Aug. **12th, 1856** became very famous - and very rich - quickly; everybody wanted his process. The London Times went as far as printing the whole paper two days later.
  - But *point defects* were fighting back. The industrial realization of the **Bessemer process** with large quantities of ore and coke yielded a big and very unpleasant surprise: **Bessemer steel** from large size production, in contrast to the Bessemer steel from "laboratory" experiments, was brittle and not fit for anything. Bessemer felt like "being hit by a flash of lightning from the blue sky"; the descend from the Olympic heights of top inventors to desperation was quick and brutal.
  - But Bessemer was a good materials scientist and engineer; if it worked once, it must work again. There must be reasons for what happened, and with diligence, one can find out what is going wrong. What had happened?
- Well, Bessemers work, and the work of many others, supplied the (here much simplified) answer. Bessemer used *Swedish iron ore* for his experiments (you always use the best in lab experiments), while his industrial country fellows used *English ore* - and this stuff contained some phosphorous. The Bessemer process (possibly in contrast to the old-fashioned steel making process) did not remove the phosphorous, and small amounts of **P** are sufficient to render steel brittle. As we know now, **P** segregates in the grain boundaries and changes the local properties in a detrimental way.
  - Phosphorous had to be removed (if you lived in merry old England, out on a conquest to assemble an empire, you did not want to have your steel production depend on the supply of Swedish iron ore). Two cousins, Sydney Gilchrist **Thomas** and Percy Carlyle **Gilchrist**, found the way in **1875**: Take (among other things) chalk stone for the lining of the **Bessemer *converter*** and even add some to the melt. The phosphorus would react with the **CaO** of the burnt chalk and end up in the slag which could be skinned from the liquid steel, or stuck to the lining.
  - There were plenty of other problems - on occasion, e.g., some oxygen remained in the steel and rendered it useless. Mr. **Mushet**, another Englishman coming to the aid of his country, found the solution: Add some "Spiegeleisen" (an iron - manganese alloy found somewhere in Germany) and your problems are gone. The **Mn** reacts with the surplus **O** and forms slag. It also neutrlizes any sulfur in the mix, which would otherwise create real trouble.
- So besides Bessemer, many people were involved in bringing large scale steel production to fruition. And, as it practically always will turn out with great inventions, somebody else *did it before*. In this case it was one Mr. **Kelly** from the **USA**, who had the "Bessemer" idea **10** years before Bessemer himself. While he made a mint over patent hassles, the name Bessemer remains attached to steel, and Kelly is quite forgotten as a materials scientist.

- After the Bessemer process was sufficiently debugged, steel production took off and became supremely important strategically.
  - **Siemens** in Germany and **Martin** in France developed the "Siemens-Martin process" and so on and so forth. The world production of steel grew exponentially (like Si or chips today): **22 kto** in **1867**, **500 kto** in **1870**, **1 Mto** in **1880** and **28 Mto** around the turn of the century. Today we are in excess of **500 Mto** a year.
  - In **1970** politicians generally still believed, that the wealth of a nation (and thus its power to subdue others) was directly coupled to its steel production (and thus to the degree of the nations prowess in manipulating point defects in **Fe**).
- You may feel now that we are talking chemistry here, and the typical urge of the chemist to produce pure substances. Nothing could be farther from the truth. We are exclusively discussing the dramatic influence of **point defects** on certain properties of a crystal lattice, like its resistance to the generation and movement of dislocations.
- If you would like to read more about this subject, refer to the splendid books of S. Sass, I. Amato und R. Hummel.

- Polybius was the guy who wrote about those bending swords of the gauls. The gauls as all other celts, unfortunately, did not write anything. "Publish or perish" is not a new invention
  - That the swords of the gauls / celts were inferior to those of the romans is about as believable as the existence of wepaons of mass destruction in Iraq **2000** years later: It was and is propaganda, stupid!
  - It probably was the other way around. The celtish long sword made from damascene steel was far superior to the roman short sword, and eventually (around **300 AD**) was adopted as the roman "spatha".
  - One is tempted to generalize: maybe the famous roman technology was mostly adopted from other folks? Be that as it may, the way the Romans *used* technology - based on discipline, organization and large-scale production - was unprecedented and instrumental in conquering most everybody.

- Here are most modules dealing with the subject as a list:
  - Steel from a Materials Science and Engineering point of view
  - Details to Damascene Technologies with many links to more sites.
  - A "magical" Sword
  - In German: Magische Schwerter (und japanische Schwerter)
  - In German: Gruselige Schmiedegeschichten (mit Magie).
  - In German: Der Ring des Nibelungen Zur Schmiedekunst und Siegfrieds Schwert

A Cross-Linked Glossary of Some Terms from the History of Metal Working

For a very good overview of the various types of steel in todays terminology activate the link

You may also want to check the following modules of the Hyperscript

Damascene Technique in Metal Working

History of Steel

A commented Internet literature list to the history of metal working

Advanced

These are my personal notes, reflecting **my** major points of interest while perusing the Internet data in May **2000**. I may update them occasionally.
You are welcome to share these notes with me, but I cannot guarantee for their scientific soundness (in contrast to almost everything else in this Hyperscript).

Bloom

- The iron-rich spongy stuff left over after smelting iron ore at temperatures of roughly **(1100 - 1200)^{o}C** - well below the melting point of (pure) iron of **1550^{o}C**.
  Steel in Ancient Greece and Rome by: E.A.Ginzel **1995**
  Metallurgical Heritage of India by S. Srinivasan and S. Ranganathan

Cast iron

- Everything with more than about **2% C** content; low melting point of **1100^{o}C** (eutectic composition) to **1200^{o}C**, depending on the **C** conc.
  Steel in Ancient Greece and Rome by: E.A.Ginzel **1995**
- First produced in China.
  Short history including furnace design:
  Early progress in the Melting of iron by V.H.Patterson and M.J.Lalich

China and steel

- First producers of cast iron; but seem to have used it mostly for agricultural and home keeping equipment "on a very large scale". (Now here you have really cultured people!)
  Steel in Ancient Greece and Rome by: E.A.Ginzel **1995**
- Produced cast iron **800 - 700 BC**
  Early progress in the Melting of iron by V.H.Patterson and M.J.Lalich

Damascus steel and blades

- Something made from **330 BC** on from wootz steel (and only from wootz steel, i.e. not with forge welding two kinds of steel). Shows "swirl coloration" and is of "amazing strength and toughness".
  "Could be bent at right angles and still snap back"
  Steel in Ancient Greece and Rome by: E.A.Ginzel **1995**
- Source for Damascus steel was wootz steel; modern reconstruction yields blades with superplasticity (?).
  "One blow of a Damascus sword would cleave a European helmet without turning the edge or cut through a silk handkerchief drawn across it" (from the crusades)
  Blades have "water pattern... whose wavy streaks are glistening - it is like a pond on whose surface the wind is gliding" (from a **6th** century writer).
  Metallurgical Heritage of India by S. Srinivasan and S. Ranganathan
- "Wootz" is the true Damascus steel.
  Damascus Steel - A Brief History by Motoyasu. (Edited by WarAngel)

- The term Damascus steel can refer to two different types of artifacts, one of which is the true Damascus steel from wootz steel (with the water pattern) and the other is a composite structure.
  The mechanical properties of the traditional Damascus blades and the degree of exploitation of the unique properties of the steel are less well understood.
  Comments on the major paper of Verhoeven et al.
  Wootz steel: An Advanced Material of the Ancient World by S. Srinivasan and S. Ranganathan
- Two types from before **500 AD**: "Normal" (two kinds of steel) and "oriental" (or otherwise "true") Damascus from wootz steel.
  Pattern from alignments of **Fe_3C** particles
  Last high-quality blade produced around **1750**; even low-quality stopped in early **1900**.
  Highest quality wootz blades from **16th- 17th** century.
  Chemical analysis of old blades.
  Reproduces "**Mohammeds latter**"
  Note "It is relatively easy to make an ingot that will not pattern on forging"
  Art was lost because change in impurity content of wootz (no more V traces)?
  The Key Role of Impurities in Ancient Damascus Steel Blades by J.D. Verhoeven, A.H. Pendray, and W.E. Dauksch
- Verhoeven summarized his findings in an Scientific Aerican article in Jan **2001**
  (*The Mystery of Damascus Blades*, John D. Verhoeven
  Essentially; Verhoeven together with the black smith Pendray could reproduce "true" damascus including specific patterns.
- The Romans were not impressed by the (early forge welded) blades of the Celtic tribes (they bent easily and broke).
  (Folded) Damascus steel is far superior to homogeneous (ancient) iron, but inferior to good homogeneous steel.
  The Road to Damascus - Sorting Modern Pattern Welding from Myth and Legend by Kevin R. Cashen
- Third technique to create Damascus blades (immersion of a wrought iron package in liquid cast iron)
  Story of Saladin and Richard Lion-Heart)
  True Damascus steel was no longer produced after the Tartar conqueror Timur Leng raided the city in the 14th century and took all blacksmith with him
  Watered steel, wootz and true Damascus, by Lord Mikal Isernfocar called Ironhawk
- In the **7th** century the Syrians in Damascus came up with their own version of wootz steel
  Damascus sword of later times in forging two-metals-technique.
  "Hummels" book

Defects and steel

- Relationship between iron and steel (the role of carbon) first described by **Torben Bergman 1781** in his "Disseratatio Chemica de Analysi Ferri".
  Steel in Ancient Greece and Rome by: E.A.Ginzel **1995**
- Relationship between iron and steel (the role if carbon) first described by Torben Bergman **1774**
  Steel in Ancient Greece and Rome by: E.A.Ginzel **1995**
- Ferrosilicon introduced around **1810**
  Early progress in the Melting of iron by V.H.Patterson and M.J.Lalich
- Relationship between iron and steel (the role if carbon) first described by the Swedish chemist Torben Bergman **1774**.
  The carbide banding mechanism (forming the water pattern in Damascus blades) was found to be assisted by the addition of P, S along with **V**, **Cr**, and **Ti**.
  Wootz steel: A Advanced Material of the Ancient World by S. Srinivasan and S. Ranganathan
- The type of impurity elements (especially **V** and **Mn** besides **C**) in the wootz steel is the most decisive element for true Damascus blades.
  True damascene depended on trace impurities. It may have be decisive were the ore came from!
  The Key Role of Impurities in Ancient Damascus Steel Blades by J.D. Verhoeven, A.H. Pendray, and W.E. Dauksch
- Could **W** (tungsten) have played a role in true Damascus steel?
  Watered steel, wootz and true Damascus, by Lord Mikal Isernfocar called Ironhawk

Europe (after the Romans) and steel

- Significant progress only in late medieval times; due to the use of coal for improved blast processes
  Relationship between iron and steel (the role if carbon) first described by Torben Bergman in **1781** in his "Disseratatio Chemica de Analysi Ferri".
  Steel in Ancient Greece and Rome by: E.A.Ginzel 1995
- Development from the Catalan forge (**8th** century) on:
  Early progress in the Melting of iron by V.H.Patterson and M.J.Lalich
- Description of the diverse periods (Bronze, Hallstatt, La Tene, Celtic, ...) and their swords.
  From Rapier to Langsax - Sword Structure in the British Isles in the Bronze and Iron Age
  by Niko Silvester
- British, French and Russian metallography developed largely due to the quest to document this structure (water pattern in Damascus blades from wootz steel).
  Wootz steel: A Advanced Material of the Ancient World by S. Srinivasan and S. Ranganathan
- Invention of the Catalan furnace was the end of ancient pattern welding (but came back later in an effort to emulate "true" Damascus steel?).
  The Road to Damascus - Sorting Modern Pattern Welding from Myth and Legend by Kevin R. Cashen
- Highly developed pattern welding technology in Europe from about **3rd** to **5th** century AD.
  The Serpent in the Sword: Pattern-welding in Early Medieval Swords by Lee A. Jones
- Toledo was the center of steelmaking from pre-roman times! Repeats what my smith told me. I have no idea about how much of it is halfway accurate.
  History of Swords from Toledo from some tourist agency

Japan and steel

- In Japan, around **600 A.D.**, smelting technology was introduced from China and Korea. The Japanese speciality was the mass production of (impure) steel, which was folded so many times and forge welded again that all the impurities were driven out of the steel and the carbon became as evenly distributed as modern steels we have today.
  Damascus Steel - A Brief History by Motoyasu. (Edited by WarAngel)
- "Tamahagane" steel from selecting suitable pieces from a bloom; much folding and hammering homogenized and carburized the steel.
  The Road to Damascus - Sorting Modern Pattern Welding from Myth and Legend by Kevin R. Cashen
- Some Japanese samurai had their swords made in Toledo!
  History of Swords from Toledo from some tourist agency
- There are many ways to compose a Japanes sword from different types of steel.
  Japanese Sword: Blade lamination methods

Hittites and steel

- First culture to produce iron (wrought iron?) in quantities ; about **1500 BC**. Had a monopoly for some time.
  Metallurgical Heritage of India by S. Srinivasan and S. Ranganathan
- Hittites vanished into oblivion around **1200 BC**, being overrun by the "sea people". This may have caused the scattering of the iron working skills throughout the Mediterranean.
  "Hummels" book

Pattern welding

- Pattern welding is about as old as iron and steel. Vikings were best at it (**500 AD**).
  Pattern welding in the West fell into disuse (around **1000 AD**, when full steel blades could be made) until around the time of the Crusades, when the knights brought back Wootz blades, and the smiths began pattern welding again to duplicate the appearance of the watering patterns found on Wootz Damascus blades.
  This seems to be the reason for the wrong assumptions that Damascus blades were obtained by forging together two kinds of steel.
  Damascus Steel - A Brief History by Motoyasu. (Edited by WarAngel)
- Relatively primitive before **500 AD**; but used by the Celts much earlier. The trick was (among many things) the twisting of the single rods.
  Later it became an art form (around **1000 AD**).
  Pattern not necessarily due to difference in C content (homogenizes considerably), but other impurities, mainly P.
  The Road to Damascus - Sorting Modern Pattern Welding from Myth and Legend by Kevin R. Cashen

- Pattern welding from **3rd - 10th** century; before that more simple techniques; zenith in the 6th and 7th century
  Patterns due to different kinds of iron, not necessarily only in **C** content, could be **P** or slag or whatever.
  Started for better quality, in the end purely decorative.
  Many swords with names of maker, but counterfeiting must have been rampant!
  The Serpent in the Sword: Pattern-welding in Early Medieval Swords by Lee A. Jones
- The Romans used pattern welding.
  Patter welding used less by around the 9th century.
  From Rapier to Langsax - Sword Structure in the British Isles in the Bronze and Iron Age
  by Niko Silvester

Recent issues; open points and contradictions

- Can you get "good" steel by roasting wrought iron in a charcoal fire?. J. Rehder **1989** said you can't, but D. B. Wagner in **1990** shows that you can.
- Is wootz steel (or Damascus blades) showing superplastic properties or shatters on impact at high temperatures? It is all a matter of having the temperature right!
- Did anybody in modern times ever made a blade which could be bent at right angles (or tried with an ancient blade)? Not mentioned anywhere.
- Was the art of damascene technique lost? Certainly not the two-steel folding kind; possibly the true (wootz) kind.
- How about true Damascus from soaking bundles of wrought iron (or mild steel) in molten cast iron?
- Was **W** (tungsten) important in creating true Damascus blades?
- When did true Damascus disappear? In the **14th** century or around **1750**?
- Did the Romans use pattern welding? Why were their swords superior to the (pattern welded?) swords of the "Franks"?
- What really happened in Toledo before the **7th** century or so?

Toledo and steel

- About **1000 AD**, a form of this technology (= Wootz) made its way up via the Moors to Spain - this technology allowed the Spanish smiths to create small amounts of smelted steel, which vastly improved the quality of their blades (this is the origin of the reputation of Spain, and the city of Toledo in particular, for manufacture of high quality blades - far better then the pattern welded blades.).
  Damascus Steel - A Brief History by Motoyasu. (Edited by WarAngel)
- Invention of Catalan furnace crucial to development of steel technology in Europe.
  The Road to Damascus - Sorting Modern Pattern Welding from Myth and Legend by Kevin R. Cashen
- The Catalan furnace, invented in **1300**, produced enough good steel and pattern welded blades went rapidly out of style.
  Watered steel, wootz and true Damascus, by Lord Mikal Isernfocar called Ironhawk
- In the **7th** century the Spaniards in Toledo came up with their own version of wootz steel
  "Hummels" book
- "There are stories of how the wrought iron swords of the Gauls bent during their battles against the Romans legions armed with **Toledo steel blades**, which were <<so keen that there is no helmet that cannot be cut by them>>. The hapless Gauls had to stop and straigthen their blades after each blow before continuiung fighting"
  "Sass'" book, p. 96.

Wootz steel

- Carbon rich steel produced on a consistent base in India from about **330 BC** up to the renaissance. Two methods are quoted.
  Steel in Ancient Greece and Rome by: E.A. Ginzel **1995**
- Anglized version of "ukku", denoting steel
  Still exported to Europe, China, the Arab world and the Middle East in the 12th century (and supposedly still a secret).
  Source for Damascus blades with "water pattern".
  Played a major role to the development of metallurgy (together with the "secret" of Damascus steel). Everybody in the **19th** century, it seems (incl. **Michael Faraday**), tried to figure out what it was and how it was made.
  Was the first "advanced" material, used in three continents for well over a millennium - unparalleled by anything else.
  Metallurgical Heritage of India by S. Srinivasan and S. Ranganathan
- Wootz was (the source of) the true Damascus.
  Damascus Steel - A Brief History by Motoyasu. (Edited by WarAngel)

- Superplasticity and other mechanical properties of wootz steel. "Superplastic material essentially comprise a two-phase material of spherical grains of extremely fine grain size of not more than **5** microns at the working temperature".
  Details on production techniques.
  High tech material of the ancient world.
  Wootz steel: A Advanced Material of the Ancient World by S. Srinivasan and S. Ranganathan

Wrought iron

- "What you get upon pounding the bloom. Relatively pure iron. Soft, easy to weld, cannot be hardened. Around since about **1500 BC**.
  Steel in Ancient Greece and Rome by: E.A.Ginzel **1995**
  Metallurgical Heritage of India by S. Srinivasan and S. Ranganathan

Most Important Technology of Mankind

- If you doubt that metallurgy is still the most important technology of mankind, play the old game:
  - What would you take along if you would be banned to an isolated island for a few weeks with nothing but your cloths and one piece of equipment of your choice?
- If you do not pick a *metal* object (knife, axe, ...) you are either a fool, extremely religious, or suicidal.

Advanced

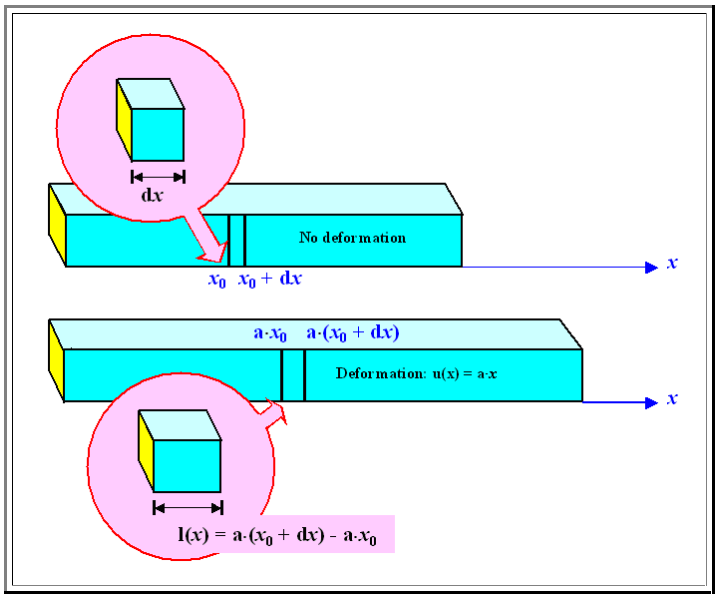
Hollow Dislocation Cores

- Dislocations with hollow cores actually do exist! In **GaN** and in **SiC** such defects have been observed.
  - The last word to hollow dislocation cores or "micropipes" is not yet out.
  - Some more informations can be found in a module discussing **SiC** in the Hyperscript "Semiconductors"

Advanced

Displacement and Strain

- While the relation between the displacement field $\underline{\boldsymbol{u}}(\boldsymbol{r})$ and the local strain tensor ϵ_{ij} is rather elementary, it does not hurt to recall the decisive points.
  - Let's take the simple example from the backbone and consider a rod that is uniformly elongated; i.e. $\underline{\boldsymbol{u}}(\boldsymbol{r})=u_x(x)=\mathbf{a}\cdot x$; $\mathbf{a}$ is some constant.
  - In other words, the vector $\underline{\boldsymbol{u}}$ *only* has a component in x-direction, which *only* depends on x as variable. The geometry than looks like this:



  - At any point in the rod a little cube will be deformed into a ***cuboid*** - the side in x-direction is somewhat longer than the others.
- What kind of strain do we have to put on a cube positioned a x, to produce the cuboid?
  - Well, since there is only strain in x-direction, we simply write down the elementary formula for strain

$$\epsilon_{xx} = \epsilon_x = \frac{l - l_0}{l_0} = \frac{u_x(x + dx) - u_x(x)}{dx} = \frac{du_x}{dx}$$

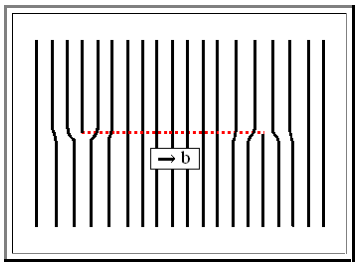
- If we deform in all three directions, we get corresponding expressions for ϵ_{yy} and ϵ_{zz}.
- Since we also might have displacement components in x-direction that depend on y or z, e.g. $u_x(x, y, z)=\mathbf{a}\cdot y$, we may, in general, also form mixed (partial) derivatives; e.g. $\partial u_x(x, y, z)/\partial y$. What do those derivatives signify?
  - Shear stresses, of course. A little less easy to see, perhaps, but there can be no doubt about it.
  - You may want to try to show that for yourself with the simple displacement field given above and the equations in the backbone as a guideline for what you are looking for.

Exercise 5.1-1

Sign of Burgers and Line Vectors

Illustration

Lets look at a dislocation loop in cross-section. After the "cut" along the red line, the lower half was moved to the right by ***b***. Two edge dislocation are visible if we look at a cross-section taken through the middle of the loop. A Burgers circuit now would give Burgers vectors of *different* signs - *or does it*?



We can ask the same question in a different way: From the Volterra construction we know that the Burgers vector - including the sign - must be the same everywhere. But the dislocations shown in the cross section look "reversed" - we would certainly assign different signs just looking at the picture. How is this contradiction to be solved?

Link to the solution

Exercise 5.1-2

Find the Mistake

Illustration

- Animations in Hyperscripts are bit tricky, because they are often done by some graphic experts who do not understand the subject. As a result, they may look very professional, but may lack precision - in other words, they may be very nice illustrations of something that is wrong.
  - Here is a version of an animated edge dislocation movement adopted from a hyperscript from *Edward Goo*.
  - The original link is http://www-classes.usc.edu/engr/ms/125/MDA125/defects/index.htm
- Can you tell what is wrong with this animation? There is a major mistake and a not so terrible one.
- Here is yet another animation, faithfully redrawn from the **CD-ROM** "Materials Science on **CD**" (Chapman and Hill; version 1.1, **1996**).
  - There is a slight mistake here, too. Can you identify it?

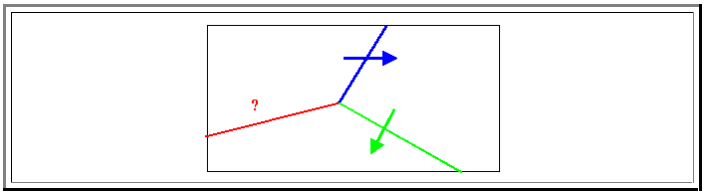
Link to the solution

Exercise 5.1-3

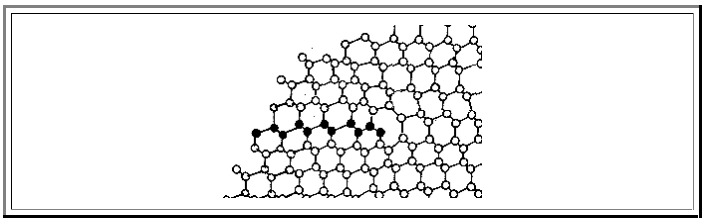
Quick Questions to 5.1: Dislocations - Basics

- Here are some quick questions:
  - The *answers* are sometimes (and possibly only indirectly) contained in the links.

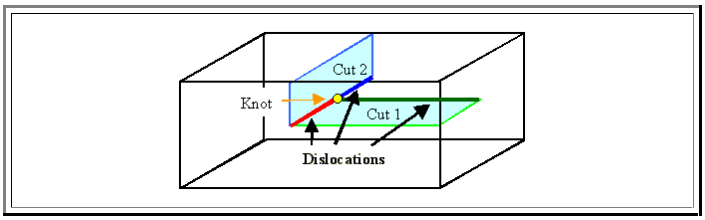
- Draw a schematic lattice fringe picture of a screw dislocations by sketching the planes above and below the dislocation line
- Produce the dislocation arrangement shown in the picture by Volterra cuts and determine the Burgers vector of the third dislocation.
- Enumerate at least 5 basic properties of dislocations.
- What do you know about the free enthalpy and so on of a dislocation? What is the source of the enthalpy (or energy) of a dislocation? Give a number and discuss consequences for (global and local) equilibrium.
- What is the difference between an edge and a screw dislocation?



- Determine the Burgers vector of the dislocation shown. Here are some hints.
  - Try to identify the unit cell first.
  - The picture shows a projection of a **fcc** lattice along a <**110**>direction
  - The crystal is of the diamond type
  - If all else fails - use this link



- The thee dislocations shown (in black, red and blue) were made by two successive Volterra cuts.
  - Three dislocations = three Burges vectors. How can you determine the three Burgers vectors by the properties of the two cuts?
  - Can you obtain this basic geometry by just *one* cut?
  - If yes, what would kind of Burgers vector would you find in this case for the red line?



- Dislocation loops
  - Draw a schematic cross -section and a top view of an edge type dislocation loop. Draw in the Burgers vector. Discuss apparent inconsistencies
  - What is the glide plane of a dislocation loop with edge type character (make a drawing)
  - Can you draw a screw-type dislocation loop?
  - Produce an interstitial type and a vacancy type the dislocation loop with the Volterra construction
  -
  -

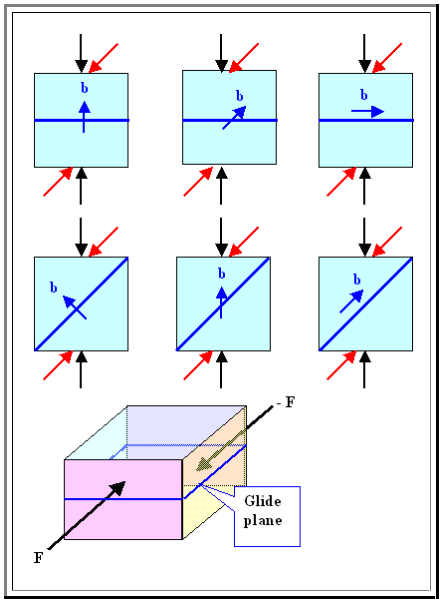
Exercise 5.2-2

Forces on Dislocations

Illustration

The drawing shows a simple crystal containing a dislocation in the **6** configurations given; we are always looking at the glide plane.

- Two kinds of forces act on the crystal, one case (black arrows) is illustrated in the three-dimensional perspective view in the lower picture.



Use the figures and draw in;

- **1.** The force (direction and rough magnitude) acting on the dislocation with black and red for the two cases.
- **2.** The position of the dislocation line after it has moved some distance.

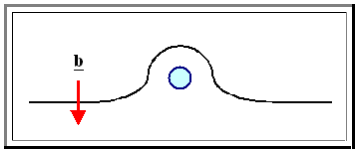
Link to the solution

Exercise 5.3-1

Dislocation Movement Around Obstacle

Illustration

A dislocation, moving in the direction of its Burges vector, encounters an obstacle and starts to move around it as shown



Use the figure and draw in:

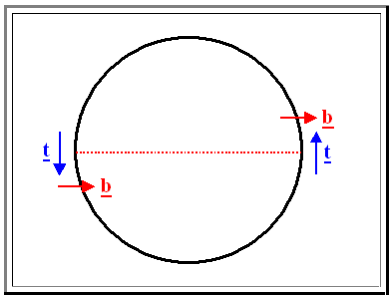
- **1.** Line direction ***t*** and the force ***F*** (direction and rough magnitude) acting on the dislocation for the case shown
- **2.** The position of the dislocation line after it has moved well beyond the obstacle (make a new picture with sufficient space).
- **3.** Draw in and discuss the forces on the screw-type segments produced. How do the segments react to the forces? Can the forces have different signs despite the same Burgers vector everywhere? If yes, why? What will the final result be?

No solution given

Solution to Exercise 5.1-1 "Sign of Burgers and Line Vectors"

Illustration

- The problem is solved easily by doing one simple thing: Look at the dislocation loop from above



  - After assigning a direction of ***t***, it is defined for the whole loop. At the places where we took the cross-section, it is actually the sign of ***t*** that is reversed! The Burgers vector thus must be "the other way around" if it is to be constant for the local ***t***.
- It is important to realize that we only can be unambiguous if we know that we are looking at *one and the same dislocation*. The cross-section by itself does not tell us that fact; it just as well could show two unconnected single dislocations. In this case we would assign Burgers vectors with different signs because we "automatically" would take the line direction to be the same.

Solution to Exercise 5.1-2 "Find the Mistake"

Illustration

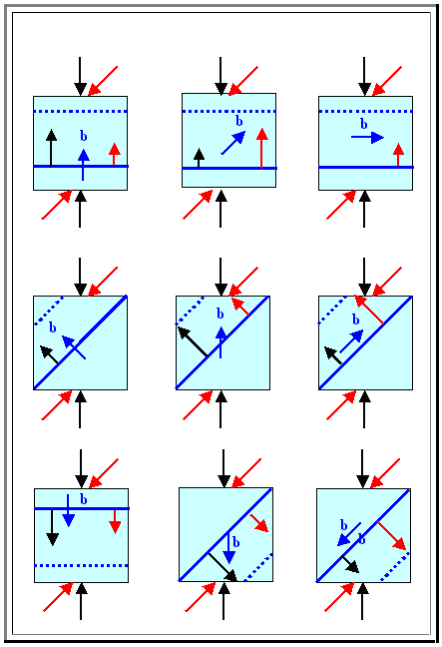
- The first animation contains two mistakes; a rather big one and a smaller one:
  - Small mistake: The lattice planes do not bent in the fashion shown while the dislocation is moving. To be sure, nobody knows exactly how they move, but the S-like shape is rather unlikely.
  - More serious: After the dislocation has left the crystal, there is no reason for any elastic deformation! The lattice planes would be absolutely straight and not curved as shown.
- The second animation contains a little mistake:
  - In the last two slides of the animation, the whole row of atoms just below the glide plane moves. This is not correct; regions far away from the dislocation core will not move perceptibly
  - However, this movement was obviously induced to achieve the strain-free state after the dislocation moves out. It corrects for the small deviation in all atomic positions which are not contained in the animation because the "artist" only changes the position of a few atoms around the dislocation core for every frame of the animation.
  - **Hint:** You can see this very clearly if you shift manually from slide to slide by moving the bar in the viewer menu back and forth with the mouse at the right position.

Solution to Exercise 5.2-2

Use the figures and draw in: **1.** The force (direction and rough magnitude) acting on the dislocation with black and red for the two cases; **2.** The position of the dislocation line after it has moved some distance.

- Here is the completed picture. The last row shows some additional cases with reversed Burgers vector.
- Of course, that the dislocations move "up" for the first two rows is an arbitrary choice as long as we don't define exactly how the sign of the Burges vector was obtained. Be that as it may, when we reverse the sign of the Burgers vector, forces and dislcotion movement also reverse signs.

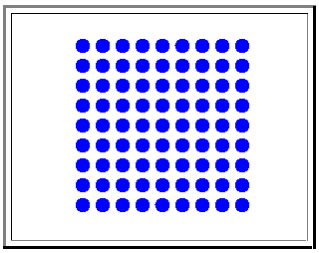
Illustration



Movement of an Edge Dislocation

Illustration

- Here is an animation giving a *schematic* view of the movement of an edge dislocation through a crystal.



- Dislocations move in response to an external stress σ. A
  - As soon as a **critical shear stress** is reached, the dislocation starts moving and deformation is no longer elastic but plastic, because the dislocation will not move back when the stress is removed.
  - The example shows the movement of an idealized edge dislocation in a cubic primitive lattice (which does not exist in nature). The grey lines show the projection of the lattice planes, the dislocation line (red symbols) is perpendicular to the screen and bounds the extra lattice plane.
  - The dislocation line moves on its **glide plane** and produces, upon leaving the crystal (and thus disappearing), an elementary step on the crystal surface. Note that after the dislocation disappeared, the crystal is completely stressfree.
  - For *macroscopic* deformation in three dimensions, many dislocations have to move through the crystal. The elementary process shown above thus has to be repeated literally billions of times on many (at least **5**) different planes of the lattice.

Illustration

Milestones in the History of Dislocations

Here are a few of the major milestones in the discovery of dislocations:

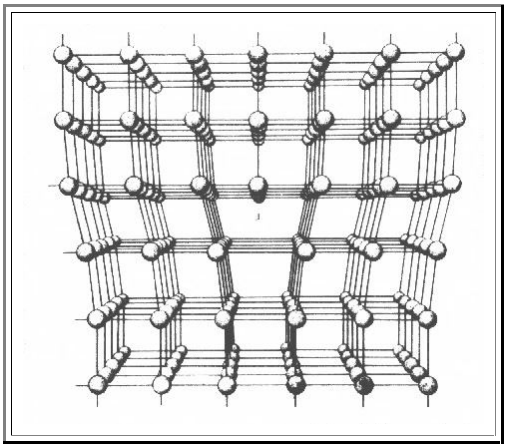
Theory of Elasticity	Timpe (1905) Volterra (1907)
Theoretical shear limit	Frenkel (1926)
"Verhakungen"	Dehlinger (1929)
Postulate of (edge) dislocations	Orowan (1934) Polanyi (1934) Taylor (1934)
Postulate of screw dislocations	Burgers (1939)
Multiplication mechanisms of dislocations	Frank-Read (1950)
Direct observation of dislocations (*TEM*)	Hirsch et al. (1956)
Dislocation free Si crystal growth	Dash (1960)

These discoveries solved the **3000** year old puzzles of metallurgy! But no Noble prizes were awarded and mankind hardly noticed that something big has happened.

Perspective View of a Dislocation

Illustration

- Here is a famous perspective view of an edge dislocation (from the times before a drawing like this one could be done easily on your **PC**)



- I am not quite sure who really did it, otherwise I would be happy to credit the author.

Illustration

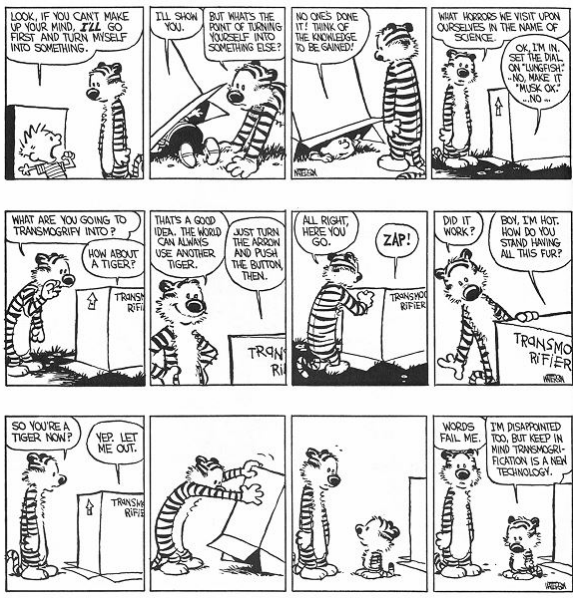
Steel, Defects, and Bullshit

- Some unedited excerpts from am Internet article from a source of "Knifes and Swords".
  - METALLURGY:
    Whereas most steels have a random crystalline structure, in *Living Steel* the crystalline structure becomes highly aligned, focusing its natural energy along the edge and towards the point. Under the hammer blows the steel becomes exceedingly dense. The hammer breaks the microscopic crystals as they form and forces them into a tighter and more highly aligned pattern. The grain boundaries become so small that it becomes difficult to distinguish one crystal from another, even with a microscope. Modern metallurgy refers to this as a super micro crystalline grain. In essence, the entire blade acts as though it were a single crystal of steel. This strongly effects the blade in two distinct ways, in its strength and in its magic.
  - STRENGTH:
    When a steel shatters, the breaks occur along the grain boundaries between the crystals. Smaller crystals in a more highly aligned pattern reduce the grain boundaries, making it more difficult for a break to occur.
  - MAGIC:
    Living Steel also gathers, focuses and transmits a low frequency electromagnetic energy similar to that which our bodies run on, similar perhaps to the way in which a ruby focuses a laser. This is a measurable phenomenon that can also be felt by the human body. In ancient times there was no explanation for this other than magic. It is still magic today.
- I sincerely hope that you can see why this is bullshit!

Transmogrification

Illustration

- Changing something into something else the easy way (without much effort and painlessly, of course) is an old dream of everybody.
  - All religions, of course, pr e a y on this fact. It is interesting (but useless) to meditate about the possible connections between the observation that the seemingly most unchangeable and common things - stones - could in fact be turned into something beautiful, useful and rare, and the evolution of believes in a "changed state of being after death" in religion.
  - Here is a present day expression of the need to dream about **transmogrification** from one of my favorite books dealing about children and how to raise them:

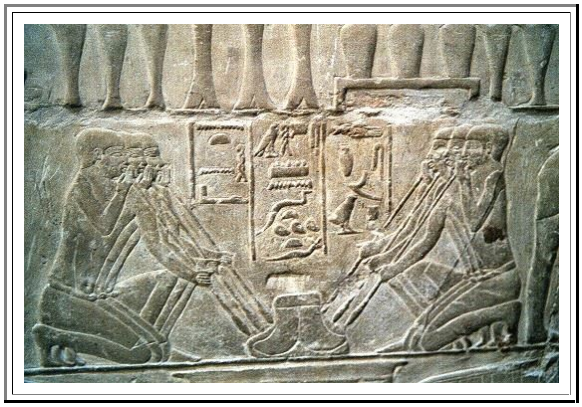


- Well, a working version of a transmogrifier has not been invented yet. So if you really want to change yourself into a person who knows about defects in crystals, you must still do it the hard way!

Egyptian Gold Smiths

Illustration

- Here is a picture (a relief, actually) from a mastaba in *Sakkara,* the necropolis of the early (and later) pharaohs and their underlings. It is about **4500** years old and shows how to raise the temperature of a fire.
  - It is a bit blurry, because it is dark in there and flashes are not allowed. In consequence, three out of the roughly ten words every Egyptian custodian or guardian seems to know are: "Flash no problem" - always uttered with a stretched out hand (palm up).
  - However, being a scientist who knows what light can do to pigments over time (we are talking defects here!), I kept my money and tried to live with highly sensitive film (another triumph of point defect and crystal engineering) and long exposure times.

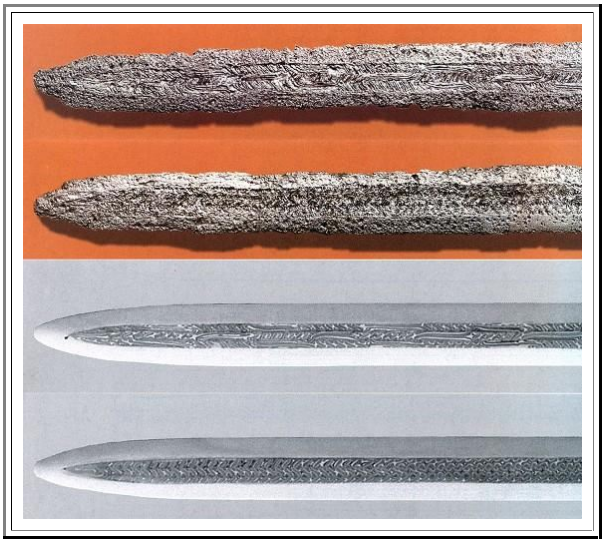


  - This picture was originally painted; but only traces of the colors are left by now.
- Note that blowing in the fire to raise its temperature does work - but not for large quantities of melt.
  - Even if having plenty of slaves it not the problem, you can't get enough of them close to the fire to reall go into mass production.
  - Serious metal industry thus had to wait for the invention of the steam engine; that is particularly true for the **Bessemer process** of mass steel production.

Illustration

Merowinger Damascene Sword

- Well, here it is - the Sword from Ingersheim from the **6th** century that got me all excited. While the original (about one half of the total length is shown) does not look too precious to the uninitiated, the reconstruction (below) is breath-taking.
  - Unfortunately the picture shown here does not do justice at all to the real thing. If you ever get to Stuttgart, don't miss to look at it.
  - Shown is the front side and the backside of about half of the length. Note that the pattern is different.



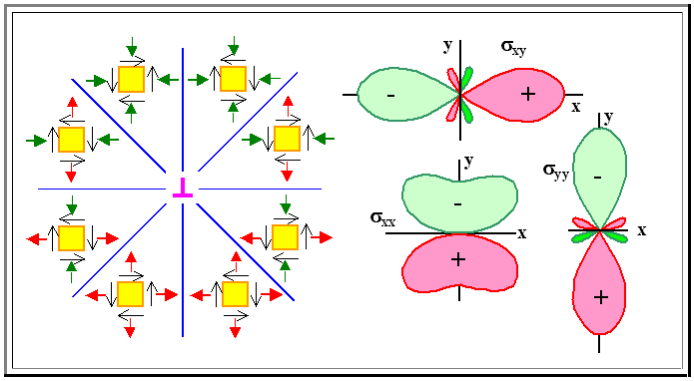
  - The pictures are taken from the wonderful book of *Manfred Sachse* (with his friendly permission).
- Here comes what the Museum has to say (translated and shortened) to this sword:
  - *The "**Spatha**" (= roman name for long straight sword) from Ingersheim represents a top achievement of the early medieval art of metal technology. It is more intricate and more complex in its structure then the celebrated blade from Sutton-Hoo (famous English site of two Anglo-Saxon cemeteries of the 6th and early 7th centuries, one of which contained an undisturbed ship burial including a wealth of artifacts of outstanding art-historical and archaeological significance). Especially remarkable is the exactly calculated change from torsion damask (the curlicue pattern) to linear damask (the straight part). More precisely, the core of the blade consists of two layers, each composed of three rods that change between being twisted and straight. The smith, quite obviously, chose a sequence of 5 successions, paying tribute to some antique number mythology, The number **5** always had a special meaning (pentagram!) and is known to have had a magical property for ancient smiths.*
    *The Sutton-Hoo blade also shows a succession of twisted and straight layers, but the individual rods all are patterned in parallel. The more sophisticated Spatha from Ingerheim surpasses that be the masterly interchange of twisted and straight portions on one side of the blade and yet another pattern on the other side, while always sticking to the **5** steps sequences.*
    A blade like this was strictly a status symbol. The Master, with the help of two or three apprentices, would need at least two full weeks to forge a blade like this one. The proud future owner would have to hand over at last 10 oxens - the present day equivalent of a merceds sedan.
  - You and me could not have afforded a weapon like that, and the alemanni nobleman who owned the blade would most certainly not have used it for lowly tasks like killing scum like you and me!
- Well, you find Ingersheim two miles to the North of the town I grew up in, right in the heart of Suebia in Baden-Württemberg. While we know that Suebians are presently a superior kind of people (Suebia produced Mercedes, Porsche and me - need I say more?), here we have incontrovertible evidence that Suebians were superior to those English lads even as early as **600 AC**!

Stress Field of an Edge Dislocation

Here is a well known representation of the stress field surrounding an edge dislocation

- On the left half of the picture, the stresses on the elementary cube are shown around the dislocation. Since there is no stress perpendicular to the image plane, a two-dimensional representation is sufficient.
- On the right half, contours of equal stress are shown for the normal component and the shear components of the stress tensor.

Illustration

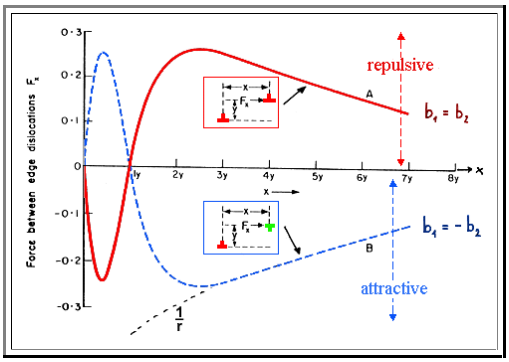


Forces between Edge Dislocations

Illustration

Shown is the force between edge dislocations of identical and opposite Burgers vectors as a function of their normalized distance.

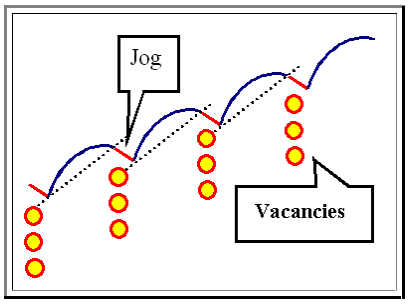
- The distance ***x*** between the dislocations is expressed in units of ***y***, the distance of the glide planes.
- The force changes from repulsive to attractive or vice verse for a distance ***x* = *y***; i.e. if the dislocations are at an angle of **45^{o}** relative to the glide plane.
- The **45^{o}** position is a stable equilibrium position for opposite Burgers vectors, because at this position ***F* = 0**, *and* **d*F*/d*x* < 0**.
- For dislocations with identical ***b*** vectors, the stable position is at ***x* = 0**.



Generation of Vacancies by Moving Jogs of Screw Dislocations

Illustration

- This picture shows the basic mechanism for the generation of vacancies by the movement of dislocation with jogs.

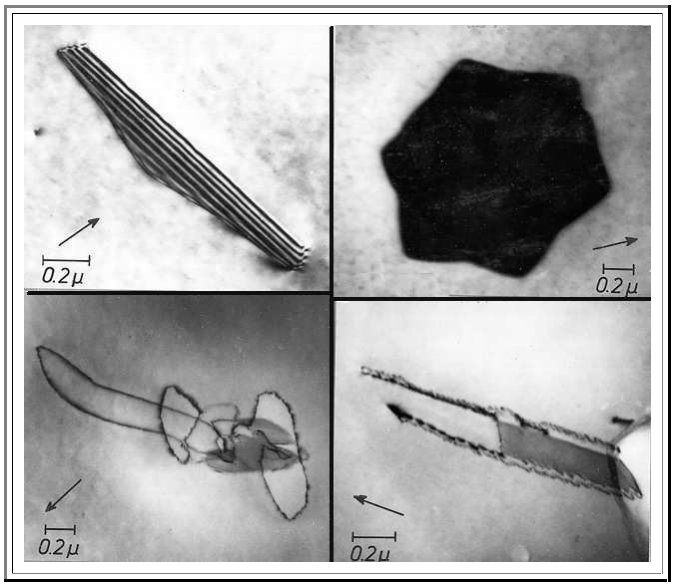


  - The screw dislocation is trying to move in the direction indicated by the bowing in response to the resolved shear stress on its glide plane which is assumed to be about perpendicular to the screen.
  - The jogs are short segments of edge dislocations; their glide plane would be the screen plan. The dislocation thus would be immobile.
  - However, if a vacancy is emitted, the jog moves one plane up (the inserted half-plane of the edge dislocation gets shorter). The jog thus has to "climb" to keep up with the rest of the dislocation
- In effect, the screw dislocation now moves as if it would experience some "friction" - but it still moves. At the same time vacancies are generated which may diffuse around and start to do their own thing.

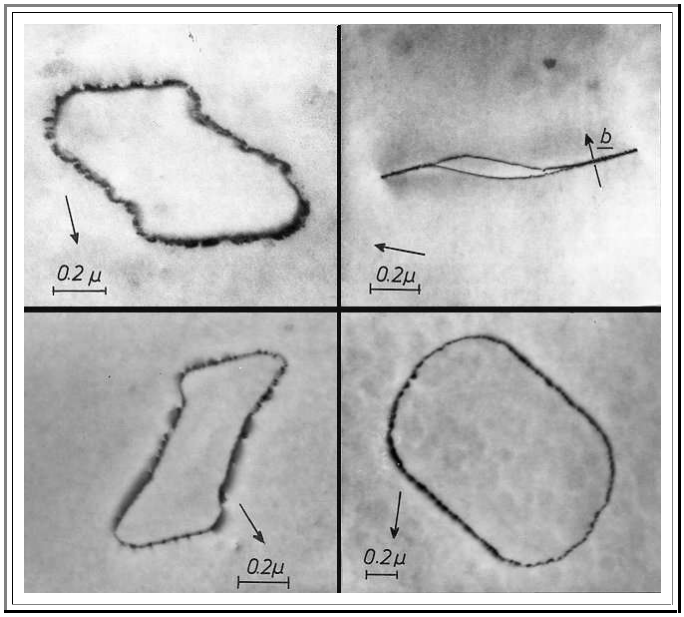
Generation of Dislocation Structures by Agglomeration of Interstitials in As-Grown Silicon Crystals

Illustration

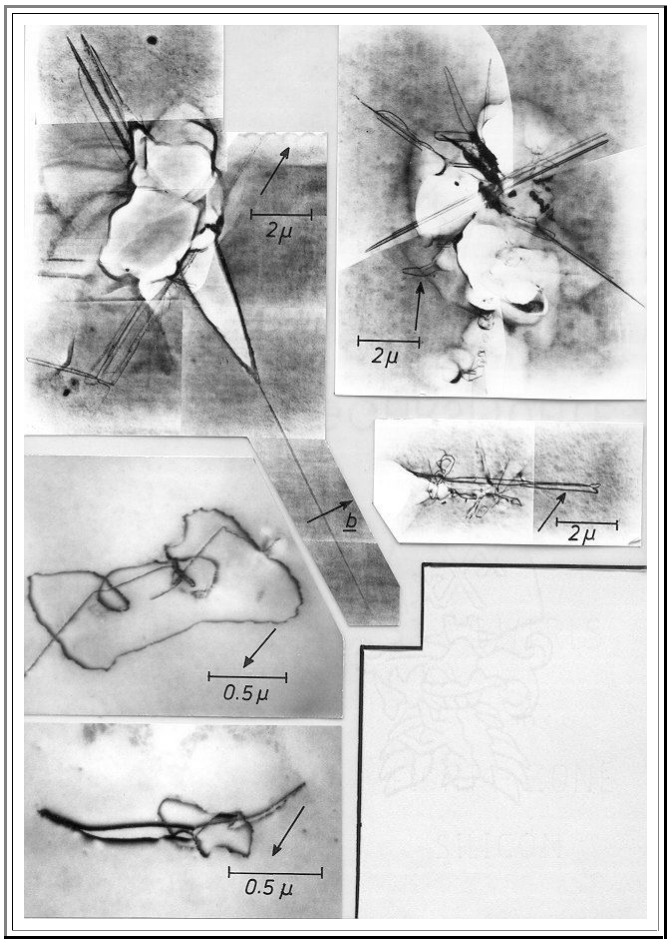
- The two top pictures, taken with a transmission electron microscope (**TEM**), show simple dislocation loops, bounded by Frank partials, which were generated by the agglomeration of interstitials. The stacking fault appears with characteristic stripes or at a brightness different from the background.
  - The loops are much larger than their equilibrium size - obviously the nucleation of the Shockley partial was not possible; maybe because the Frank dislocation line is decorated by impurity atoms.



  - The two pictures above show loop complexes. Some loops still contain stacking faults in parts oft their structures, but others are perfect and have started to move around.
- The pictures below show simple loops after the defaulting process. They are now bound by a perfect dislocation which assumed (more or less) hexagonal shape.
  - Two segments have started to move way. The "fuzzy" contrast of some dislocations may be due to impurity segregation or to tiny new Frank loops which nucleated at the dislocation core. This may happen because even after the primary loop has formed, there are still supersaturated interstitials which tend to agglomerate; but they now find efficient nucleation sites at the existing dislocations.



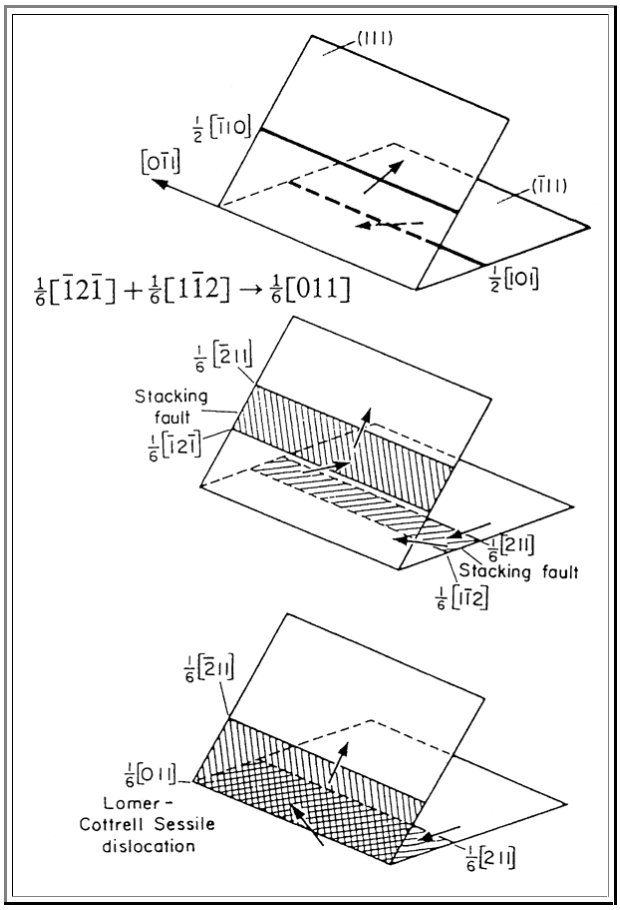
- Below are the end products. Complicated dislocation structures have formed; long dipoles were drawn out at some places. Add a little mechanical stress and you will have a crystal full of dislocations (and unsuited for integrated circuits).


2μ
2μ
b
2μ
0.5μ
0.5μ

Reaction Forming a Stair-Rod Dislocation

Illustration

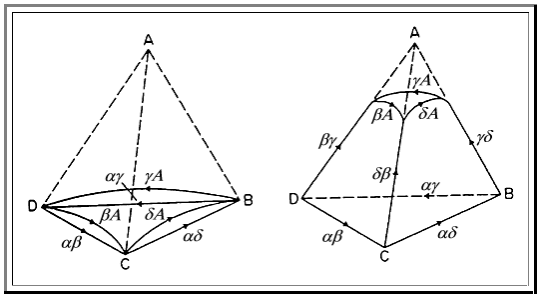
- The following sequence of pictures (taken from "Read and Hull") shows the formation of a Lomer-Cotrell dislocation.
  - Two perfect dislocations on different **{111}** planes split into **Shockley partials**, move until they meet, and react. The end end product of the reaction is a **stair-rod dislocation** with a **Lomer-Cotrell dislocation** at its apex.



Formation of Stacking Fault Tetrahedra

Illustration

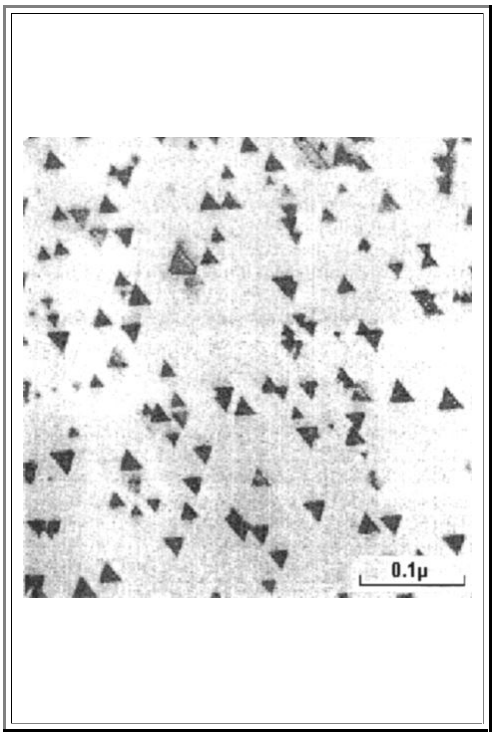
The picture, once more taken from "Read and Hull", shows the dislocation reactions and movements leading to the formation of a stacking fault tetrahedra in the short-hand notation of the Thompson tetrahedra.



Stacking Fault Tetrahedra in Gold

Illustration

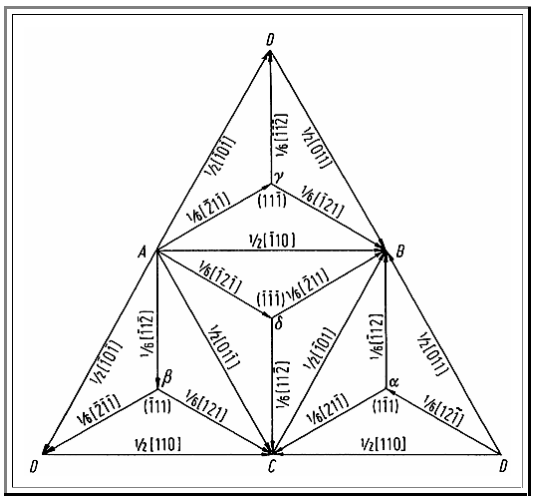
- This micrograph shows what you find when you quench **Au** from high temperature, trying to quench in the thermal equilibrium vacancies.
  - The vacancies had enough time to agglomerate to some extent, and their preferred form of agglomerate - as in many other **fcc** crystals - are **stacking fault tetrahedra**.



Thompson Tetrahedra

Here is the net for a Thompson tetrahedra.

- Use it to make one! You will need it.



Illustration

Defect Etching in Silicon

Advanced

- The big **Si** wafers are prime material for defect etching because they combine a really large area (**300 mm** wafers are now (**2001**) appearing in production) with a (hopefully) very small defect density, and sometimes very small defects (precipitates and point defect agglomerates).
  - In consequence, the main advantages that defect etching has to offer - large areas and high sensitivity - is exactly what you need.
- It is therefore not surprising that defect etching was and is the method of choice for getting an overview for what is going on in your sample. Main areas of interest are
- "*Native*" defects in **Si** crystals and unprocessed (or just pre-processed) wafers. Essentially, researchers hunt for:
  - The unavoidable *agglomerates of the point defects* that were present at thermal equilibrium at high temperatures. There are all kinds, and their classification is not too clear - in the seventies, people considered "swirl defects" (subdivided into **A**- and **B**-defects), then came **C**-defects, and more recently **D**-defects. The practitioners in the production lines prefer abbreviations like "***COP***" ("Crystal originated particles or pits") or "***LPD***" (for "light point defects", which does not mean light-weight point defects but "point defects detected by scattering light") that simply includes everything that some very expensive equipment detects on the surface (including etch pits).
  - The actual defects behind "**LPDs**", e.g., might be vacancy agglomerates ("**D**-defects") in the form of voids (which are, after all, small pits if cut by the surface of a wafer), particles that are stuck to the wafer surface, "precipitates" of organic molecules which will form on wafers kept a long time in a plastic container, or .... - just anything that scatters light.
  - Defects related to *oxygen* (i.e. oxygen precipitates or stacking faults caused by oxygen precipitation). This includes also oxygen precipitation intentionally introduced in the bulk of a wafer - and only in the bulk - by some special heat treatment for reasons of "**intrinsic gettering**".
- *Process induced defects*; meaning everything generated during the processing of an integrated circuit or other Si devices.
  - There are many defects that may occur: *Dislocations* produced by plastic deformation due to large temperature gradients, oxidation induced *stacking faults*, metal *precipitates* - the Hyperscipt is full of examples (check the matrix of modules)
  - Specialities include large power devices, where just *one* defect in a wafer can kill the whole device (in integrated circuit manufacture it would just kill one out of some **200** chips).
- Defects in *solar **Si***, i.e. "cheap" **Si**, mostly multicrystalline and full of defects of all kinds.
  - *Grain boundaries* abound, but more important are often the dislocations and precipitates of impurities. Optimized chemical etching together with plenty of experience allows to distinguish the different types.
- There are many defect etches for **Si**, They all rely on the basic chemistry of forming and dissolving an oxide. Since SiO_2 dissolution always requires **HF**, all defect etches contain hydrofluoric acid and thus are very dangerous chemicals.
  - The "trick" is to make the dissolution process difficult in general, so that it may become enhanced at defects. Why it should be enhanced is obvious on the one side - the bonds around defects are weakened after all - and rather tricky on the other side - electronic properties certainly may play a role, too.
  - The common defect etchants (or "etches") therefore restrict one of the crucial reactions, and that is usually the oxidation (after all, it is the Si that contains the defect, and not the SiO_2).
  - One of the first defect etches developed, the so-called **Dash** etch, therefore simply took the standard Si dissolution chemistry (always HNO_3 + **HF** + **HAc** (**HAc** = acetic acid)) but with far less HNO_3 than the standard solution. While it worked, it was not optimal in terms of selectivity and sensitivity and in producing etch features that allow to distinguish between certain types of defects. And, most disadvantageous, it needed etch times of **4 - 16 hrs**.
- The next approach then replaced HNO_3 by an oxidant that is still sufficiently strong to oxidize **Si**, but just barely so - CrO_3 (dissolved in water). Several etches were developed and are used, but we will see that there is still a "black art" component.
  - The first CrO_3 based etchant is called "**Sirtl etch**" after its inventor (Ms. **Adler**, who was a co-author (and probably did the work), has been forgotten by now). It employs **HF** + CrO_3 + H_2O in a ratio of **HF**(H_2O free) : H_2O **:** CrO_3 ≈ **1 : 0,4 : 0,2**.
  - The Sirtl etch works well, *however only on **{111}** surfaces*.
  - A less well known etch is called **Seiter etch**, employing (**HF** + CrO_3 + H_2O in a ratio of **HF :** H_2O ≈ **1 : 9** with **120 g** of CrO_3 dissolved in **100 ml** of the H_2O. Its peculiar behavior is that it etches defects very well, but (as far as is known) *only on **{100}** surfaces*.
  - A great etch used by many researchers is the **Wright etch**. It mixes **HF(conc)** + CrO_3 + H_2O in a ratio of **1 : 0,5 : 1** and throws in **0,5** HNO_3**(conc)**, some $Cu(NO)_3$ (**2** grans for **60 ml** H_2O) and **1** unit **HAc**.

- What the **Cu**-nitrate does is (relatively) unclear.
- The **Secco etch**, finally, using (**HF** + **$K_2Cr_2O_7$** + **H_2O** in a ratio of **HF : H_2O = 2 : 1** with **44 g $K_2Cr_2O_7$** dissolved in **1 l** of the of **H_2O**. *It etches defect on all surfaces*.

There are lots of more etches - most notably, perhaps, the "**Schimmel**" etch (a kind of improved Secco etch); and a literature search will easily find upwards of **50** papers dealing with defect etching in **Si**. No really good explanation has been offered for the strong dependence of the etch anisotropy on the composition.

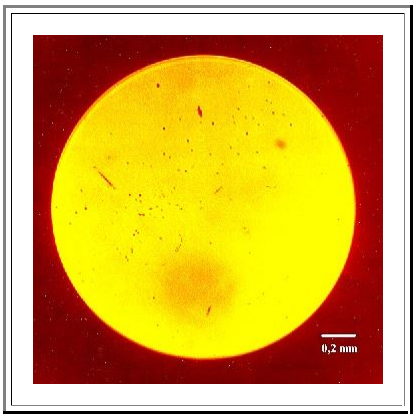
Seiter gives a short comparison of the major etches that is reproduced below.

Etch	Composition (Mol %)			Results on {100}	
	Solvent 1)	HF	Oxidizer 2)	line defects	"point" defects
Secco	67,6	32,2	0,17	pits	shallow pits or hillocks
Sirtl	71,2	26,3	2,5	pits or mound	-
Wright	78,5	16,1	5,4	pits	shallow pits
Seiter	78,5	5,9	15,6	mounds	mounds
1) H_2O + CH_3COOH (HAc); 2) CrO_3 + HNO_3					

- You see that the oxidizing part is the limiting factor indeed, and that on occasion you get mounds or hillocks and not pits, i.e. the dissolution is slower at the defect site and not faster.
- Try to make sense of this, and you will be a scientific hero.

But their are more puzzles:

- While a regular Sirtl etch used at room temperature etches defects always faster than bulk Si, it reverses its behavior to some extent at lower temperatures, say **10 °C**. Than a *hillock* may form instead of an etch *pit*. While the defect now is just as visible as with an etch pit, we have the dramatic difference that the defect is *still there*! This technique, together with a rather tricky specimen preparation for a subsequent **TEM** investigation was pioneered by **Kolbesen** et al.
- This was the key to finding the swirl defects shown in the link. Their density is so low that just blindly searching with the **TEM** would make it very unlikely of ever finding one. This becomes clear when looking at a typical **TEM** specimen:



- This is a **TEM** specimen where the whole area is transparent to the electron beam. It is so thin, that it is also transparent to regular light - what you see is a light optical micrograph.
- The small dark dots are the hillocks produced by defect - etching the front side. Under the hillocks are the dislocation loop defects shown in the link.
- The hillocks are visible in the **TEM** at very low magnifications (say around **5.000x**) and thus allow to find the defects.
- Without this guidance, you have quite a job of finding the defects: The magnification necessary to see the dislocation loops is about **50.000x**; i.e. the screen shows about **5 µm^2** of the sample. The area you have to scan is about **5 mm^2** - you must find the **20 - 50** defects by looking at about **1.000.000** screen pictures, i.e. your chances of hitting one "blind" are about **1 : 20.000.**

The original papers:

W.C. Dash, J. Appl. Phys., **27** (1956) 1193

E. Sirtl and Annemarie Adler, Z. Metallkunde **52** (1961) 529

H. Seiter, in "Semiconductor Silicon 1977", ed. H. Huff, E. Sirtl; Electrochem. Soc. Proc. Series, p. 187

F. Secco d'Aragona, J. Electrochem. Soc. **119** (1972) 948

Margarete Wright Jenkins, J. Electrochem. Soc. **124** (1977) 757

B.O. Kolbesen, K.R. Mayer and G.E. Schuh, J. Phys. E, **8** (1975) 197

A Warning

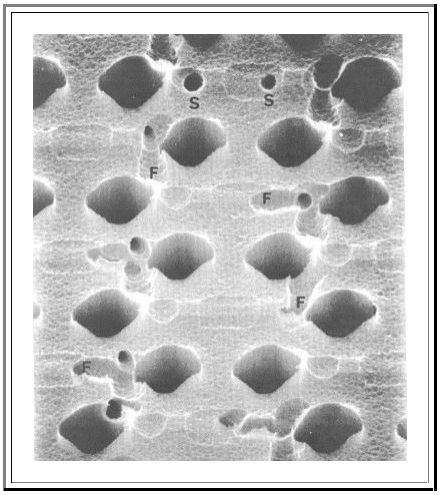
Working with **HF** without taking proper precautions may well severely injure or even kill you. Make sure you know what you are doing! Also make sure that you can identify **HF**, even in small droplets that might have been spilled somehow.

There is a little HF tester or in the market that you should always have in reach. Here is the address: Dr. V. Lehmann; FAX ++49 89 56826696; e-mail: vl@hf-acid-sensors.de; Bau & Vertrieb elektronischer Messgeräte; Geyerspergerstr. 53; D-80689 München

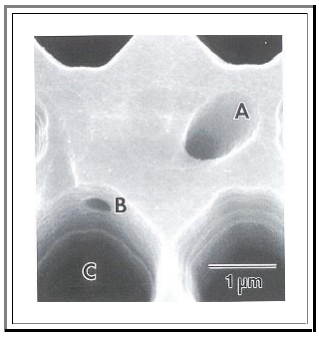
Detecting Dislocations in Trenches by Chemical Etching

Advanced

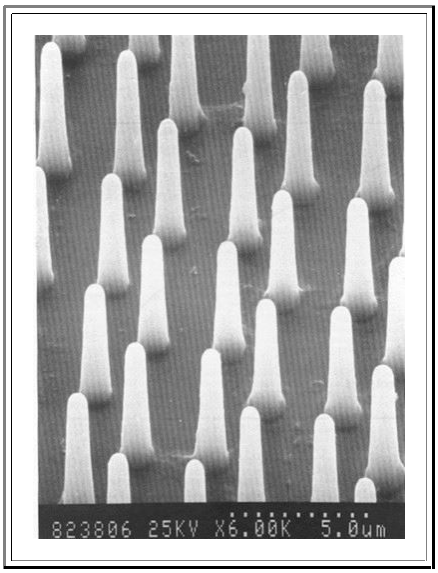
- With the development of the **4** Megabit Dynamic Random Access Memory (***DRAM***) and of the eighties, a new process was introduced into **Si** technology: Trench etching for trench capacitors.
  - A *trench*, contrary to the literal meaning of the word, is simply a small *hole* - typically **1 µm** in diameter and **(6 - 8) µm** deep. It is made to provide a large area for the capacitor while still using only about **1 µm^2** in "real estate" on the chip surface.
  - All kinds of problems were encountered with the performance of the trench capacitors, some - maybe - caused by dislocations ending inside the trench.
  - Since there are about **4 · 10^6** trenches on one chip, and some **100** chips on a wafer, looking with an electron microscope at a few trenches will not do much good - you will turn to defect etching,
- There is a clear question to the analytical people then: Are there dislocations inside the trench, and if yes, does his correlate with electrical performance? Well, *Wendt, Sauter and Kolbesen* of Siemens AG answered this question in an elegant, if tricky way with chemical etching. Here is what they did.
  - If you etch your wafer with some defect etch, you may obtain pictures like the following one



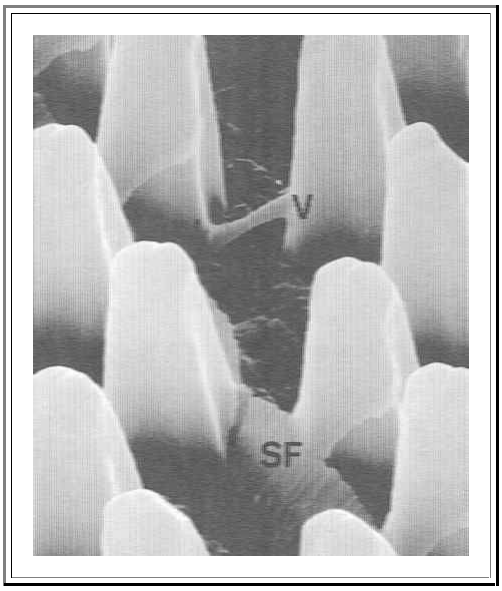
  - This is a **SEM** picture, because you would not see very much with a light microscope - the big holes being the trenches are just about **1 µm** across. In fact, while you would detect the relatively shallow dislocation etch pits marked by "**F**", you would miss the sharp little holes marked with "**S**".
  - Well, you see that there are dislocations ending at the surface. What you do not see is if there are dislocations ending on the surface of a trench; i.e. *inside* the hole. If you look real close, you might on occasion find something as shown below:



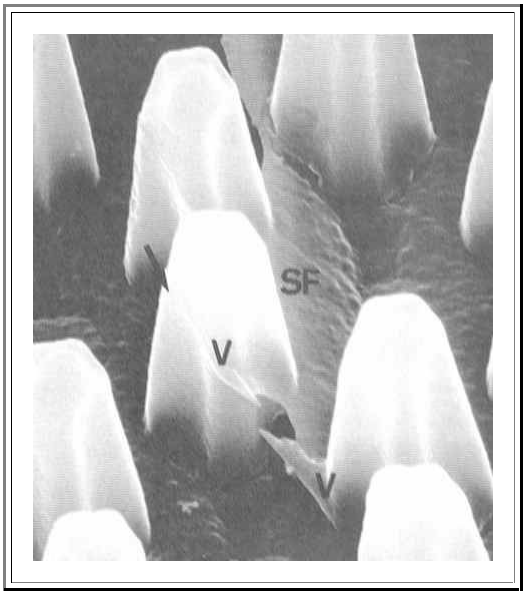
  - OK, here you have the etch pits of a dislocation that starts at the surface at "**A**" and ends obviously inside the trench at "**B**" - but still rather close to the surface.
  - How about deeper down in the trench? How do you look inside a **1 µm** trench (with any method)?
- Here is the solution:
  1. Etch the whole wafer, producing etch pits inside the trench if there is a dislocation.
  2. Coat everything with a thin layer of **Si_3N_4**.
  3. Etch off all of the **Si**, leaving only the **Si_3N_4** layer intact.
  4. Inspect the **Si_3N_4** layer. It is a kind of "negative" of the trench structure which now is easily inspected.
  - Pretty tricky (and by far not as easy as it sounds). Here are some results (the "stripes" are artifacts from image processing)



- The inverted trench structure - pretty perfect in this case.



- The structure marked "**V**" clearly results from a dislocation running from a trench to its neighbor.
- The structure marked "**SF**" actually shows a stacking fault.

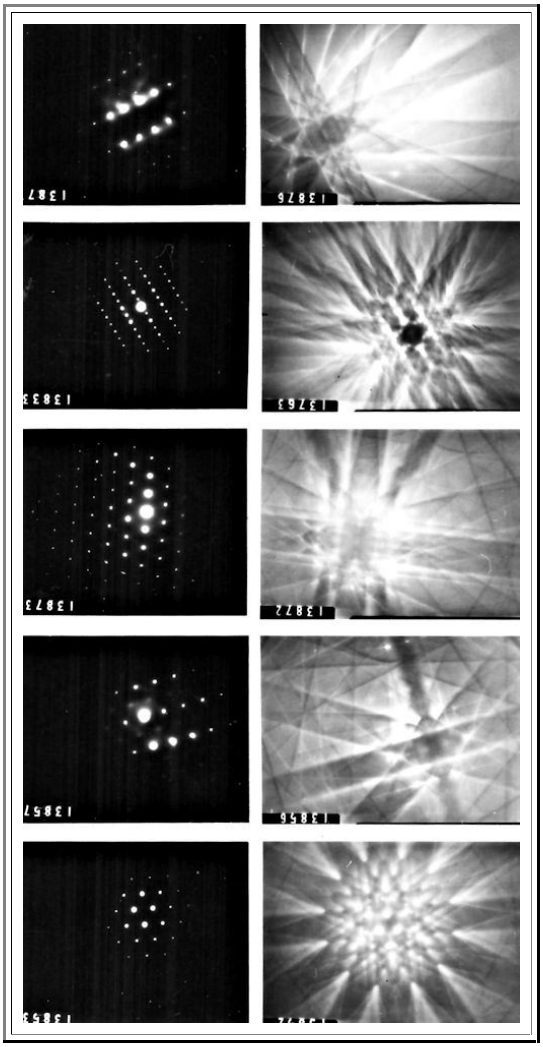


- Same thing here - a prominent stacking fault (and a dislocation).
- So we have stacking faults, not just dislocations at work here! This is a major finding, that would have been practically unavailable with other methods.
- It is a major finding, because now we have a pretty good idea where the defects are coming from: We most likely deal with a new kind of oxidation induced stacking fault, and that give us a clear idea of what needs to be done to get rid of those defects.

Diffraction and Kikuchi lines in the TEM

Advanced

- The wave vector of the primary electron beam is rather large compared to typical reciprocal lattice vectors; any small part of the the **Ewald sphere** is almost a straight line or plane in **3-D**, respectively.
  - For a thin foil the points in reciprocal space become elongated perpendicular to the foil; the flat Ewald sphere the can cut through many reciprocal lattice points - many reflexes are excited!
  - In very thin specimens where inelastic scattering is negligible, the diffraction pattern then consists of many reflexes with intensities that decrease as the excitation error increases. It is nearly impossible to establish precise diffraction conditions; e.g. a two-beam case with a defined excitation error.
- Fortunately, with a bit of inelastic scattering, electrons that are first scattered inelastically and then elastically, form a system of lines, so-called **Kikuchi lines**, which give a precise picture of the diffraction conditions.

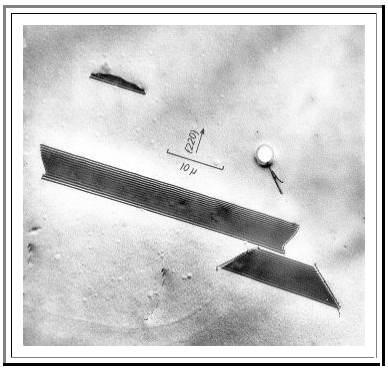


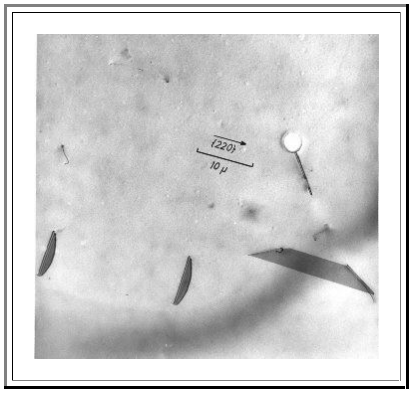
  - Shown are some diffraction patterns; on the left from "thick" specimens with Kikuchi lines, on the right from "thin" cases with the same orientation and without visible Kikuchi lines.
  - So, simply move to thick part of your specimen, where with some practice and the help of a "Kikuchi Map", it is easy to tilt the specimen to any desired orientation with high precision, and then go back to a thin part.

Stacking Faults and Micro Twins

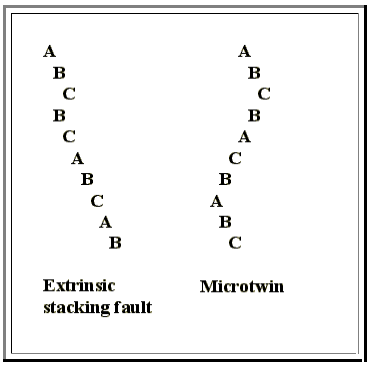
Advanced

- Shown is the same area of a **Si** sample imaged with two different diffraction vectors of the **{220}** type. The defects are the result of an epitaxial process used for making integrated circuits which was followed by diffusion /oxidation step.
  - Whereas some defects completely vanish with one diffraction vector and show strong contrast with the other one, the medium sized defects stay in contrast (the small dislocation ending at an etch pit, too).





  - Further analysis, using more special diffraction conditions, show the that medium sized defect is a **micro twin**. The difference between micro twins and stacking faults is shown in the graphic below:

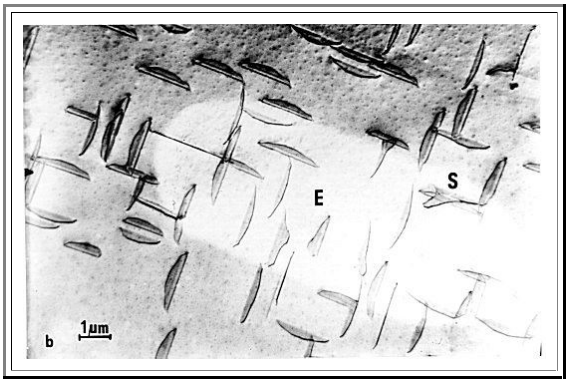


  - An interesting side observation was that the microtwins "killed" the devices, while the stacking faults did not. In other words, devices containing stacking faults still worked, while the ones with a microtwin inside were electrically faulty.
  - This is a correlation that cannot be obtained by defect etching or any method without "high resolution", because a microtwin and a stacking faults would be indistinguishable.

Oxidation Induced Stacking Faults in Silicon

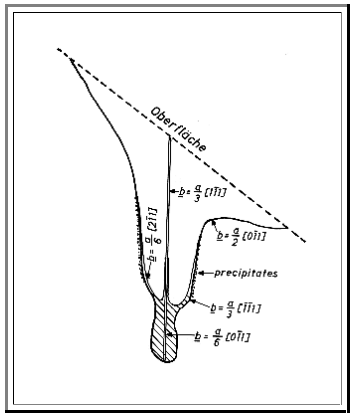
Advanced

- Oxidation of Silicon produces interstitials in supersaturation. These surplus interstitials tend to agglomerate in discs - i.e. stacking fault loops. The difficult part is the nucleation; it determines what will happen. We have to consider two ways of oxidizing **Si**, we first consider
- *Surface oxidation*: The surface oxidizes homogeneously by exposing it to an oxidizing atmosphere at high temperatures. This is the normal oxidation process. The emission of interstitials occurs at the interface; the interstitials diffuse into the bulk; the supersaturation decreases with the distance from the surface.
  - There is no easy nucleation for an interstitial type dislocation loop as long as the interface is defect free. If defects are present, most prominent small precipitates of metal impurities as, e.g. **Fe**, **Ni**, **Cu**, they may serve as nucleation center for the interstitials; a stacking fault penetrating in a semicircular fashion into the bulk is formed. If many precipitates are available, a large density of small stacking faults may be observed:

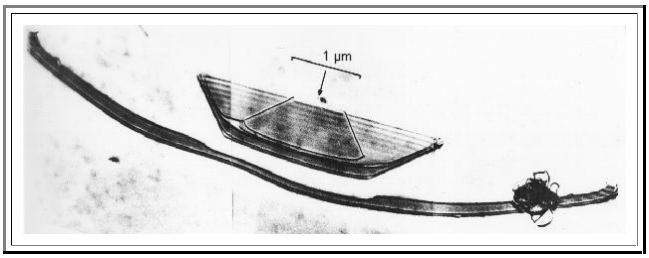


  - The oval shaped area with a lighter contrast is the emitter of a bipolar transistor. In preferential etching this would look similar to what was shown as an illustration for etching.
  - Some of these small stacking faults have a peculiar, "sailing-boat" like shape (marked by "**S**" in the picture above). Below, a detailed view of a "sailing boat stacking fault":

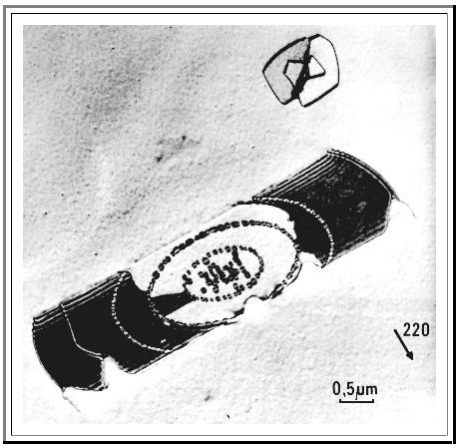
  - These "sailing boats" are formed whenever the nucleation produces two stacking faults on different planes (Connected by a pair of stair-rod dislocations). Obviously the diffusion of interstitial down the central dislocation dipole must be rather efficient. These stacking faults penetrate through the **pn**-junction and lead to a total loss of the transistor.
- In rare cases, "sailing boat" stacking faults started to unfault. For reasons still unknown, the unfaulting process stopped at a certain depth (maybe due to doping influence?); the resulting structure is remarkable, because it contains all types of dislocations that exist in an **fcc** lattice in one defect:



  - We have the perfect dislocation (**b = a/2<110>**), the Frank partial dislocation (**b = a/3<111>**); the Shockley partial dislocation (**b = a/6<112>**) and the stairrod dislocation (**b = a/6<110>**) in one defect.
- If there are only a few precipitates; they may nucleate a stacking fault many times. As soon as the first dislocation loop is too large, a new one will form. As a result, a whole system of overlapping stacking faults is seen (for every third one the contrast disappears because the sum of the displacement vectors is a lattice vector).
  - In this example the precipitate is still visible as a black dot in the center of the stacking fault system. This is usually not the case because the precipitate is incorporated into the oxide and etched off.



- If the **Si** contains some supersaturated oxygen (at high temperatures an equilibrium defect as an interstitial; "**O_i**"), we may observe **internal oxidation**.
  - A **SiO_2** precipitate forms by the agglomeration of **O_i**; but this may equally well be considered to be an internal oxidation of a small volume of **Si**. Again, interstitials are produced with the tendency for agglomeration.
  - In contrast to surface oxidation, nucleation is rather easy. The small **SiO_2** precipitate, especially if it is not spherical, has a stress field that helps to nucleate the stacking fault of the interstitials. We thus find oxide precipitates surrounded by large stacking faults.
- Both processes - the oxide precipitation and the stacking fault formation - occur simultaneously; new precipitates may be nucleated at the Frank dislocation and vice versa.
  - In the course of several high temperature treatments; the processes start all over again and complicated structures develop:



  - Several perfect stacking fault loops overlap (truncated by the sample surface, one of which has been preferentially etched; the etch pits at the dislocations are clearly visible). Some of the loops serves as nucleation sites for a second and third round of oxygen precipitation (shown as small coffee-bean like contrasts).

I-V-Characteristics of Junctions with Diffusion and Generation Currents

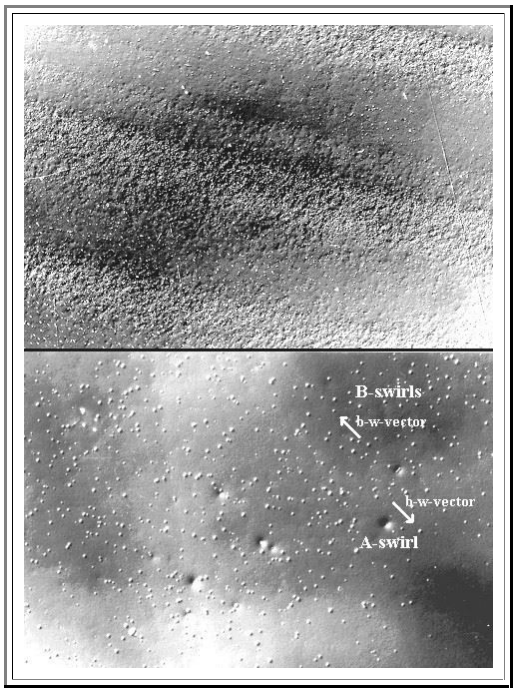
- ***I-V***-characteristics of **pn-**junctions should be basic knowledge. However, it is not a particularly easy subject.
  - The "ideal junction", without consideration of what happens in the space charge region, is not too difficult. Including diffusion and generation currents from the **SCR**, however, makes things quite messy. A proper treatment besides involving some knowledge of "Shockley-Read-Hall recombination theory" still must cut a lot of corners by using all kinds of approximations.
- If you are not too familiar with these topics, you may want to read up about it in the Hyperscript "Semiconductors". Here are the links:
  - Simple treatment of **pn**-junction (including **SCR** part)
  - More involved treatment of simple **pn**-junction
  - Shockley-Read-Hall recombination
- More links will be found in these modules.

Etch Pattern of Swirl Defects in Silicon

Illustration

- The micrograph shows a **100 mm Si** wafer after preferential etching. The wafer was cut from a large as-grown crystal and only polished before etching.
  - The crystal was grown with the float-zone technique and represented the state of the art in about **1972**. The typical spiral pattern of the small etch pits lead to the name "Swirl defects". These defects were extremely detrimental to the functioning of integrated circuits and power devices made from the wafer. It was thus of prime importance to learn about their nature so that they could be avoided.

- The picture was taken under "dark field" conditions. The wafer is illuminated at an angle; only light that is scattered at defects reaches the lens of the camera. Perfect areas are totally black. The defects must be due to agglomerates of the point defects (including perhaps the major impurities **O** and **C**) that were present a high temperatures - presumably in thermal equilibrium.
- The etch pattern at high magnifications as seen through an optical microscope reveals two types of defects (see also the pictures in the link) The first picture is at an intermediate magnification, the second one at high magnification:

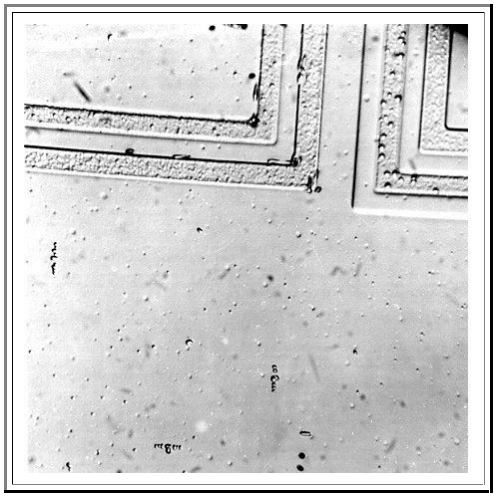


- Lots of small etch pits can be seen in a striated pattern - the swirl pattern. The inner areas of the wafer may only contain these "**B**-type" defects, whereas closer to the edge of the wafer, some large hillocks - the **A**-type" swirl defects are contained within the **B**-defects. Hillocks and pits give different signs of the black-white contrast (the vector from the black part of the contrast to the white part); this serves to distinguish between the two possibilities.
- The a-type swirl defects are dislocations loops and dislocation loop clusters of interstitial type - the loops shown before. This result was the first direct observation that showed that self interstitials play a role in Si. Etching techniques can not provide a result like that.
  - In fact, it was never possible to establish the nature of the **B**-type defects. They might be "fore-runners of the **A**-type defects - i.e. some kind of interstitial agglomerate - or small vacancy agglomerates; possibly small voids; but nobody knows for sure.
  - Since present day crystals are much larger and grown with different techniques, swirl defects are now longer seen. But other types of defects (called **C**- and **D**-defects) are present now and always first detected by optimized preferential etching solutions. **D**-defects meanwhile have been identified as small voids, i.e. vacancy agglomerates

Precipitates and Other Defects as Seen with Preferential Etching

Illustration

- Shown is a preferentially etched part of an integrated circuit. Many kinds of defects are revealed; the interpretation is not necesarily clear.
  - The big etch pits in the frames of device parts are due to dislocations.
  - In the structureless area we see pits and hillocks (distinguished because the "black-white vector", the vector from the black part of a small contrast to the white part comes with both signs) and a few very distinctive features consisting of a central pit with "satellites" along one direction.

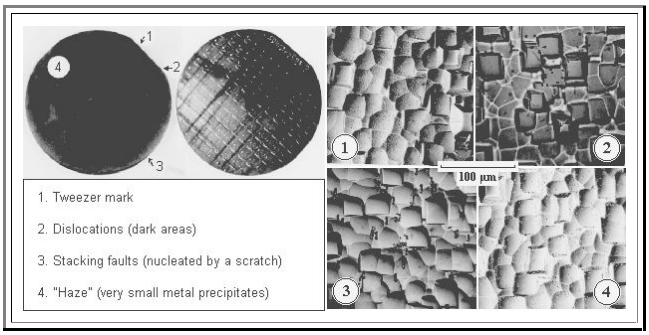


  - All these features are most likely due to precipitates. The rows of pits are caused by precipitates that produced a sequence of dislocation loops to relieve the stress in a process known as "**prismatic punching**".
- What prismatic punching looks like if imaged with a transmission electron microscope at high magnfication can be seen in the link

Process Induced Defects in Silicon Wafers

Illustration

- Shown is a wafer that has been processed to some extent in order to produce integrated circuits. Four kinds of defects were created that can be clearly distinguished, and (with some experience) identified as to their nature and cause of generation.

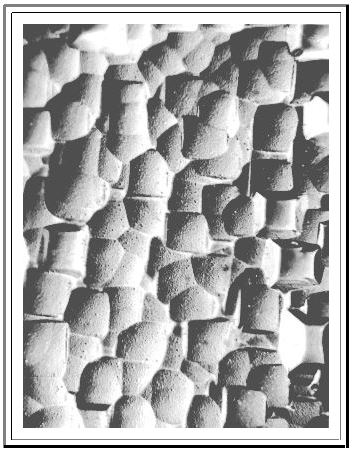


- The whole view of the wafer shows the polished front side (left) where not much is visible at this size. The backside (right) has been intentionally roughened by a **KOH** etch, this accounts for the large scale structure (the intersecting approximate rectangles) in the enlargements in the second half of the picture.
- The four micrographs showing the particular defects can be viewed at higher resolution. Click on the corresponding numbers.
  - Picture 1 and 2
  - Picture 3 and 4

Process Induced Defects: Large View of Tweezer Marks and Dislocations

Illustration

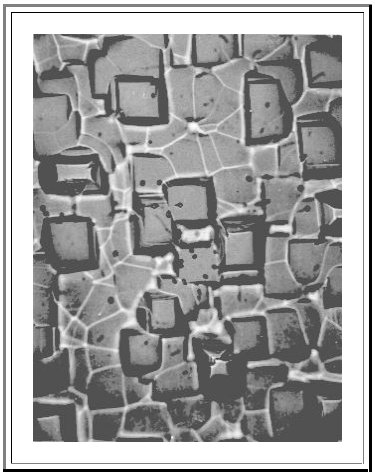
Tweezer Marks



- The tweezer marks consist of extremely small etch pits in high density - the hallmark of "**haze**". It is pretty safe to conclude that we are observing very small metal precipitates of some kind.

- Back to overview

Dislocations

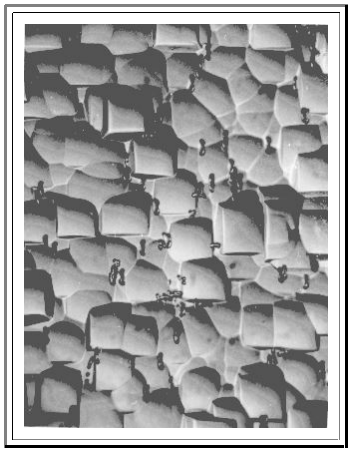


- The dislocations are marked by large and deep etch pits; sometimes slightly inclined. With a little experience in defect etching, they cannot be mistaken for anything else.

Process Induced Defects: Large View of Haze and Stacking Faults

Illustration

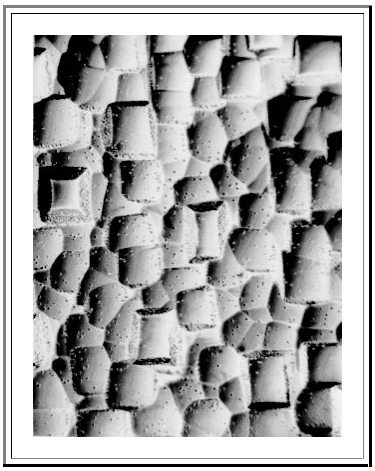
Stacking faults



- The stacking faults are shown by a grove along their intersection with the surface, always bound by two deep etch pits denoting the Frank partials bounding the stacking fault.

- Back to overview

Haze

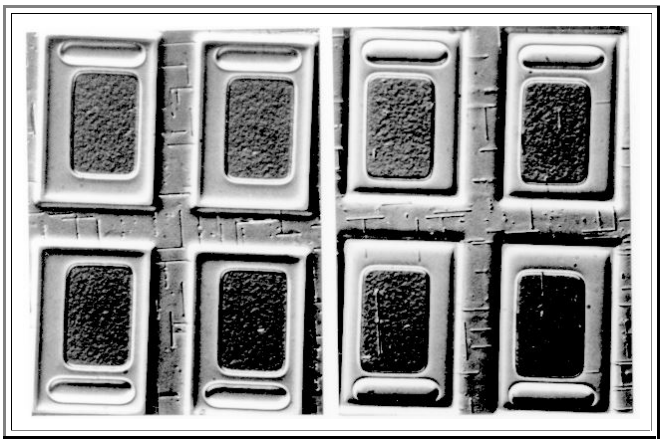


- Classical "Haze". Small etch pits in rather large density denote metal precipitates. The "Tweezer marks" in the other set of pictures are haze, too, but at the edge of the detection limit.

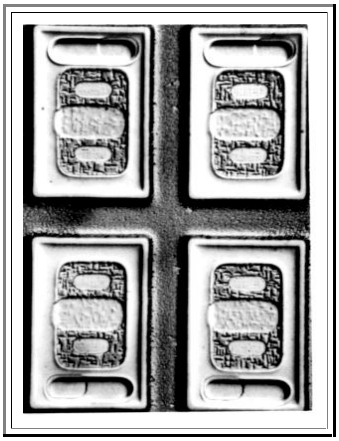
Defects in Transistors Revealed by Etching

Illustration

- Stacking faults revealed by the so-called "**Secco etch**" for **30** seconds in transistors. For one kind of process we can see different phases of the stacking fault generation by process-induced "forces" by looking at wafers from different stages of the process:



- Here we see bipolar transistors where just the collector contact (the oval part) and the base region (the rectangular part) have been defined. We have:
  - Small stacking faults only outside the transistor areas (the rectangles) on the left-hand side
  - Large stacking faults inside and outside the transistors on the right-hand side. These stacking faults may actually be rather complicated structures akin to the one shown in the backbone text.

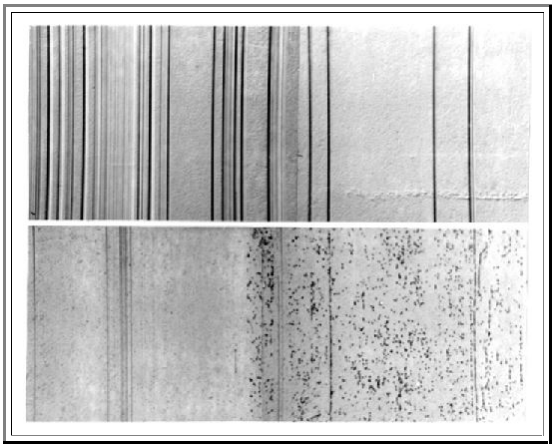


- Now we have progressed to the emitter (the two smaller ovals) the base contact (the bigger oval) and the highly doped contact area of the collector (the small oval in the bigger collector contact oval)
  - We have small stacking faults in high density only in the base region (including the contact). The others either vanished (unlikely) or were not produced in this wafer.
  - What that looks like at high magnifications as seen by transmission electron microscopy can be seen in the link.
- The question is: Will the transistors work? The answer is: It depends.
  - The transistors without any stacking faults will work, but their leakage currents may still be considerably higher than the leakage currents in transistors without any stacking faults in the neighborhood.
  - The transistor with large stacking faults in their interior will most likely not work at all. They will have a short-circuited emitter-base diode.
  - The transistors with small stacking faults will mostly work, (albeit not too well), but some of them will be dead. The likelihood of "death by stacking fault" increases with the stacking fault density. This puzzle could be solved by **TEM**.
- We also can see that the alignment of the structures to the **<110>** direction of the wafer is rather poor (the rectangles are not parallel to the traces of the stacking faults) and the alignment of the structures is not too good either (the base contact, e.g., is not at the exact center of the base).

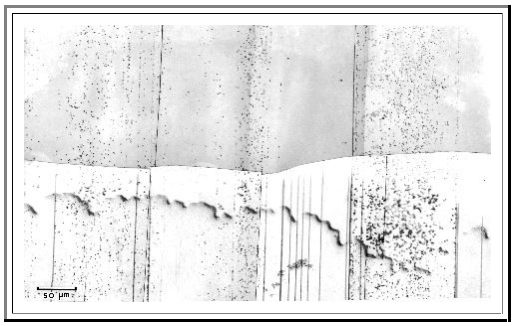
Anodic Etching of Defects in Silicon

Illustration

- The differential etch rate during anodic etching depends on the current density. At small current densities, defects may etch much faster than perfect Silicon; anodic etching than reveals the defects very clearly.



- Shown are adjacent areas of a **Si** specimen that was grown for solar cell applications. It contains many grain boundaries, preferably twin boundaries, and dislocations.
  - The upper pictures was obtained after etching with a rather large current density. Only grain boundaries can be seen; but this may be due to steps between different grains because the etching rate depends on the grain orientation.
  - The lower picture shows the area etched with low current densities. Many grain boundaries are no longer visible (despite the fact that we know they must be there), but a large number of dislocation etch pits is visible.
- Comparing anodic etching with chemical etching gives a similar result:

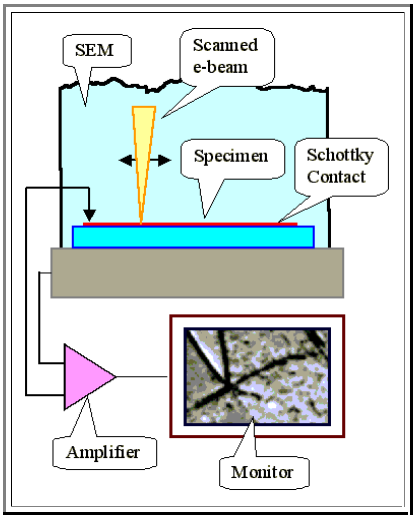


- The upper half of this sample was anodically etched, the lower with a purely chemical etch (**Secco-etch** in this case).
  - Evidently the anodic etch does not show *some* grain boundaries. From other experiments it became clear that anodic etching under these condition shows only *electronically active defects*, i.e. defects that influence electronic properties, especially the carrier life time.

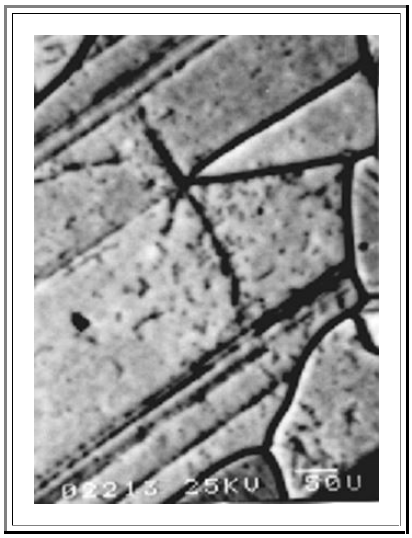
Principle of Electron Beam Induced Current Microscopy

Illustration

- The "**Electron Beam Induced Current** method (***EBIC***) employs a (***SEM***) on a sample with a thin electron-transparent Schottky contact (usually evaporated **Al**). The Schottky contact is biased in reverse, the leakage current is amplified and displayed on a monitor synchronized with the electron beam scan.
  - The elecon beam induces carriers; the minority carriers either recombine at defects or are collected at the Schottky contact as current with the resulting signal being displayed on the monitor.
  - The picture on the monitor thus shows the efffective minority carrier life time. Defects that are "electronically active" reduce the currents; they appear in dark contrasts.



Principle of **EBIC**

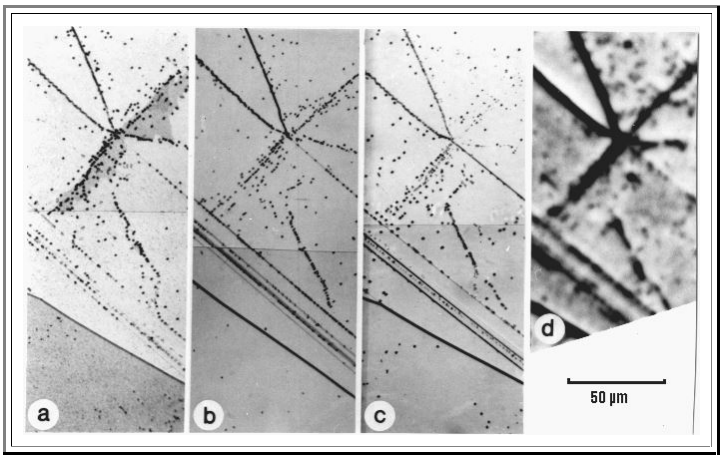


Typical **EBIC** picture, showing electronically active defects in solar-grade **Si**.

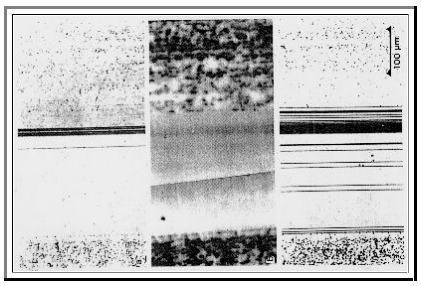
Comparison of Anodic Etching to Chemical Etching and EBIC

Illustration

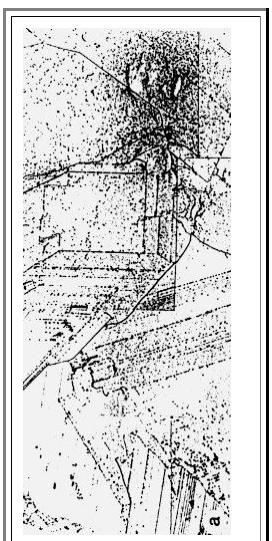
- The results obtained with **anodic etching** depend on the current density used. For small current densities there is a tendency to reveal only electronically active defects, whereas at higher current densities all defects are etched. This can be seen in comparison with "normal" chemical etching and with **EBIC**



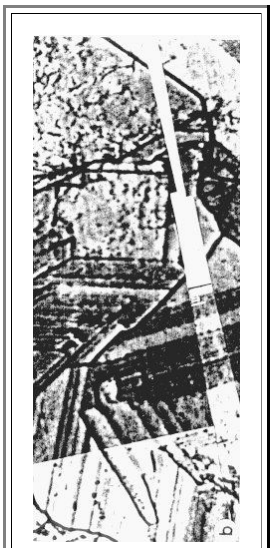
- The pictures show the same area of a solar **Si** sample (always repolished after one experiment):
  - **a)** Anodically etched at small current density. Only some of the twin boundaries at the lower half of the picture are faintly delineated.
  - **b)** Anodically etched at high current density. The twin boundaries at the lower half of the picture are delineated.
  - **b)** Chemically etched. The twin boundaries are partially delineated.
  - **c) EBIC** Micrograph. Upon close inspection, it is mostly compatible with **a)**.
- This gives the impression that anodic etching at small current densities reveals only electronically active defects whereas at higher current densities it shows all defects. This can be clearly demonstrated in another optimized comparison below

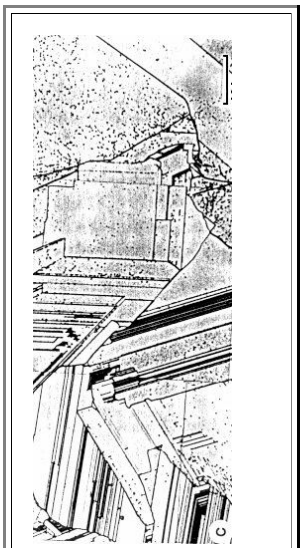


  - Left, the etching structure obtained at small, on the right with high current densities. The **EBIC** picture is shown in the middle. It is obvious that only a few twin boundaries show up at low current densities and in the **EBIC** mode.
- One more example confirms this result



Low current density



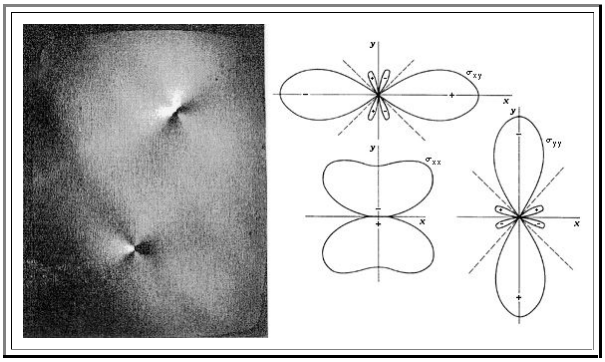



Low current density
EBIC
High current densitiy

Infrared Microscopy of Defects

Illustration

- The strain field of dislocations shifts the polarization plane of light. A specimen (with two polished surfaces) between almost crossed polarizers thus transmits more or less light than the bulk around defects

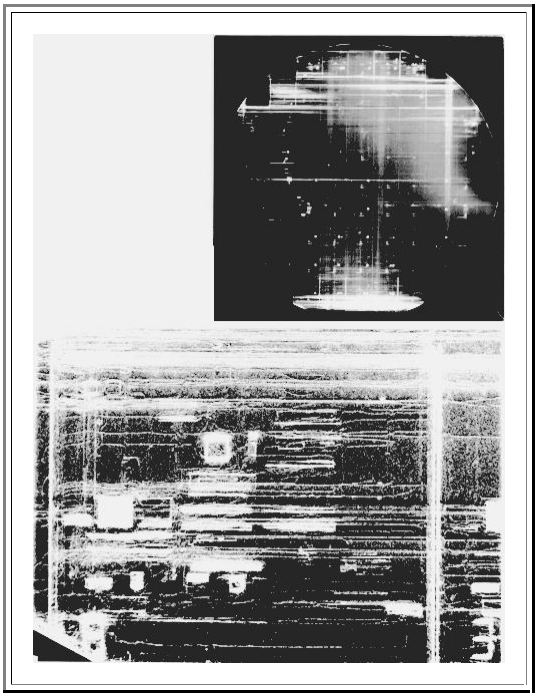


- The picture shows two edge dislocations in **GaAs** parallel to the optical axis. The contrast is directly proportional to the sign and magnitude of the strain field (or stress-field, which is shown for comparison)

X-Ray Topography

The big advantage of x-ray topography is that it can reveal the defect structure of large samples, in the case below whole (**100 mm**) **Si** wafers. If the negatives are enlarged, details on a **10 µm** scale may be seen.

Illustration



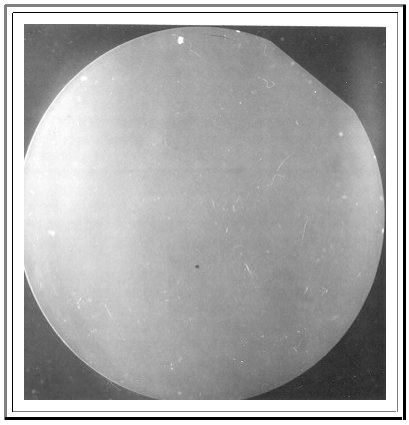
X-ray topography of a **Si** wafer showing "**haze**" (milky area in the upper right half) and dislocation structures. The isolated bright small rectangles are transistors full of dislocations

Enlargement of an area on the lower left. Some single dislocations are barely visible; the bright geometric structures correspond to device parts with high densities of dislocations.

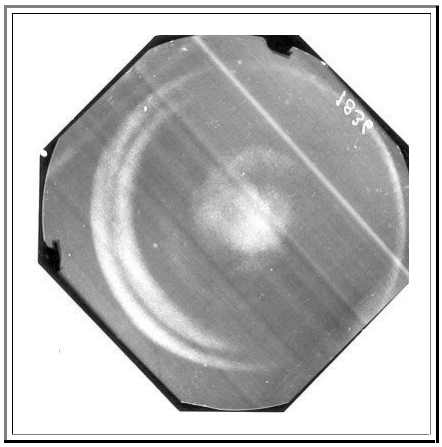
X-Ray Topography Case Study

Illustration

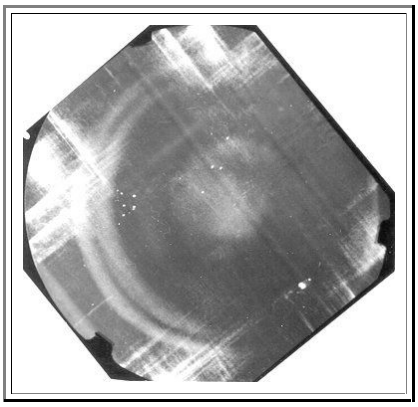
The following sequence shows **X**-ray topograms taken from the same wafer after major processing steps for bipolar devices.



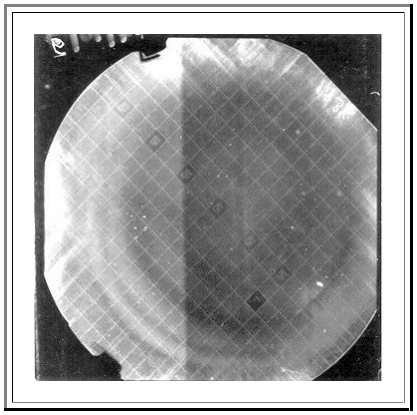
Starting wafer; no defect structures are visible



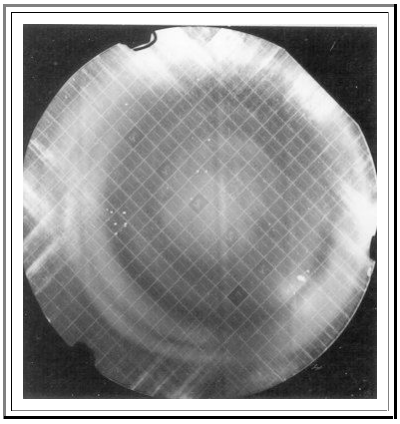
After "buried layer" diffusion; the first high temperature process.
The ring like structures are typical for oxygen precipitation.



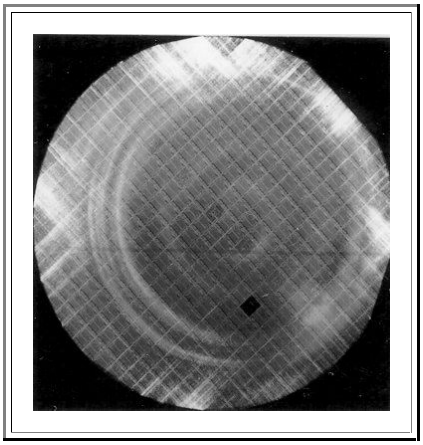
After epitaxial layer deposition. Very high temperatures are used,
in this case some plastic deformation produced dislocation arrays



After collector diffusion. The defect structure remains essentially unchanged, first device structures become visible.



After base diffusion



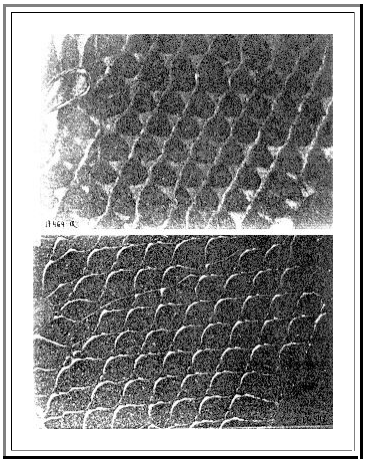
Finished wafer.

The sequence of topograms established that the crucial processes for defect generation are the buried layer diffusion and the epitaxy. The processes coming later may change the size and structure of the defects already present, but they do not generate new defects.

Weak-beam Image of a Network of Partial Dislocations

Illustration

- Shown are two weak-beam images of the same area of a dislocation network with threefold symmetry in a small angle grain boundary in Silicon.
  - Three sets of screw dislocations split into partial dislocations interact to form a network of partial dislocations bounding extended and constricted stacking faults.



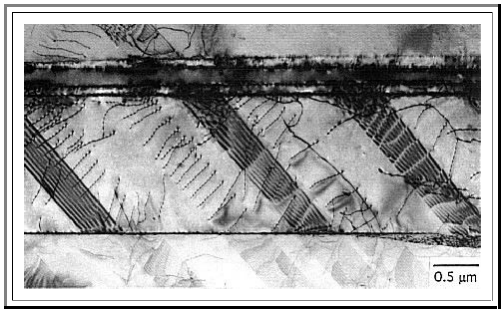
- Only one set of dislocations and one kind of stacking fault is visible
- Two sets of dislocations are visible; the stacking faults are invisible.

- With more images the Burgers vector of all dislocations and the displacement vectors of the stacking faults can be determined. The complete analysis of this network in given in chapter 7; here the images only serve to illustrate that what you see depends very much on the imaging conditions.

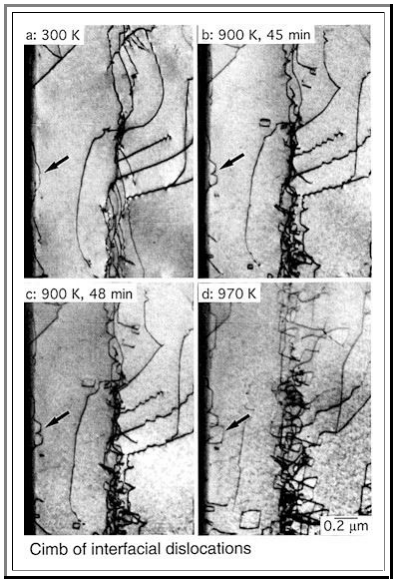
Dislocations in TiAl

Illustration

- Two more pictures from D. Appel and his group from the research center **GKSS** in Geesthacht.
  - The first one shows a medium magnification bright field picture of **TiAl**, just showing a lot of boundaries, stacking faults and dislocations
  - Not the parallel arrangement of some dislocations, which run right through the sample and move on the same glide plane, probably coming from some active source in the boundary.



  - The following pictures shows a time sequence of the same part of the specimen after successive annealing treatments. The changes in the dislocation structure (some pointed out by arrows) are due to **climb processes**.



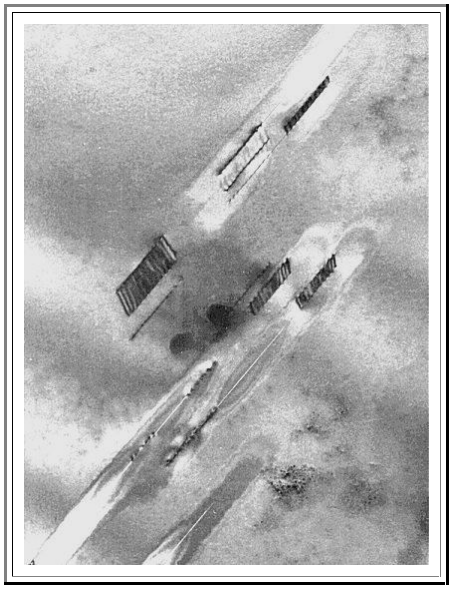
Cimb of interfacial dislocations

Unknown Defect in Silicon "Ribbons"

Illustration

On occasion, electron microscopists encounter defects that do not fit in any known category. That may be due, of course, to inexperience of the researcher. But some defects have been analyzed by many researchers without conclusive results.

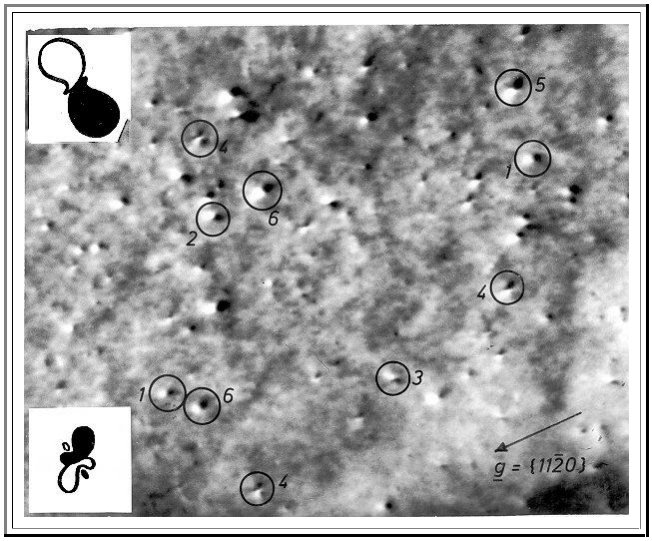
Below are some defects that were found in Silicon ribbons - **Si** poly-crystals grown as flat ribbons for solar applications - and that have not been identified. Their hallmark are the fringes in the "wrong" direction.



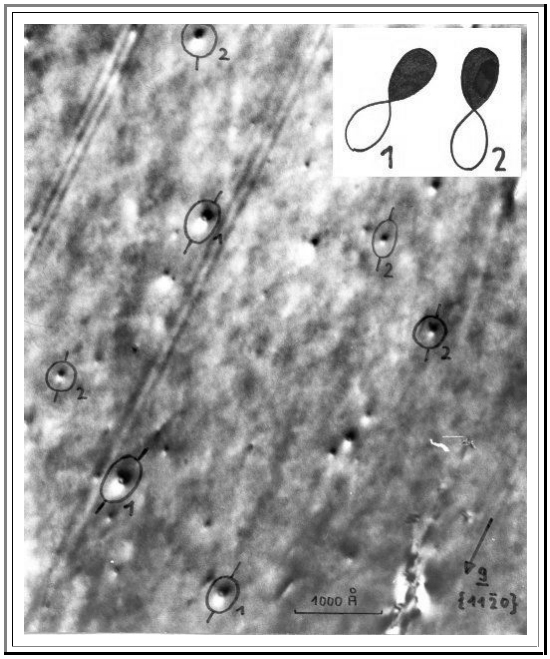
Radiation Damage in Cobalt

Illustration

- Metal irradiated with ions (in the following examples **Au** -ion with energies of some **10 kV**) will be heavily damaged; besides lots of Frenkel pairs, small vacancy type dislocation loops usually form some **10 nm** below the surface.
  - This kind of research was important for nuclear materials science and for ion implantation techniques in general.
  - The loops are far too small to be seen as loops in conventional imaging; at best they appear as black dots. However, if imaged with dynamical bright-field conditions, they give rise to so-called black-white contrasts with peculiar geometries.
  - The following picture shows black-white contrasts of dislocation loops imaged with a **{1,1,-2,0}** type of diffraction vector in a specimen with a **{0001}** orientation. Six distinctly different kinds of contrast are observed. Two calculated contrast profiles for a particular set of Burgers vector and normal vector of the loop are also included. The size of the black-white contrasts is about **20 nm**.



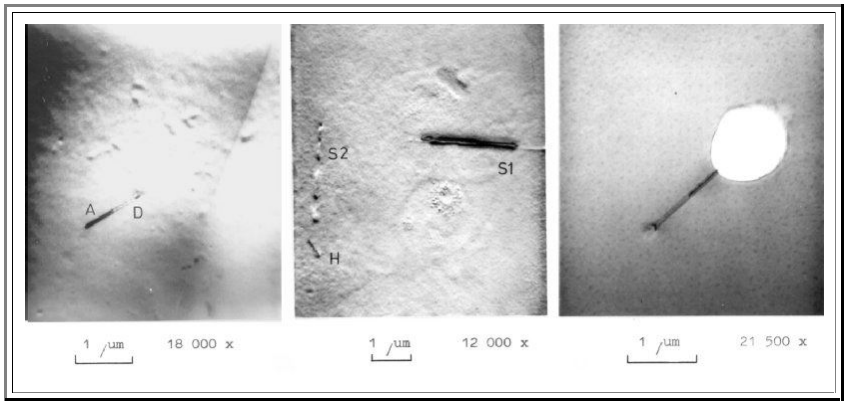
  - Same as before, but for a **{1122}** type specimen orientation. The observed contrasts match closely the calculated profiles for the types of dislocation loops assumed.



Iron Precipitates in Si Integrated Circuits

Illustration

- Iron is a major contaminant in integrated circuits because there is a lot of steel in contact with the wafers or with materials needed to process a wafer.
  - Iron atoms diffuse as interstitials; they are rather mobile. Since the solubility at low temperatures is low, there is a strong tendency for agglomeration. The small iron silicide precipitates in turn serve as nucleation centers for large defects, especially the huge oxidation induced stacking faults.
  - An iron concentration of well below **1** ***ppb*** thus may enough to kill all integrated circuits in the thus "contaminated" part of a wafer.

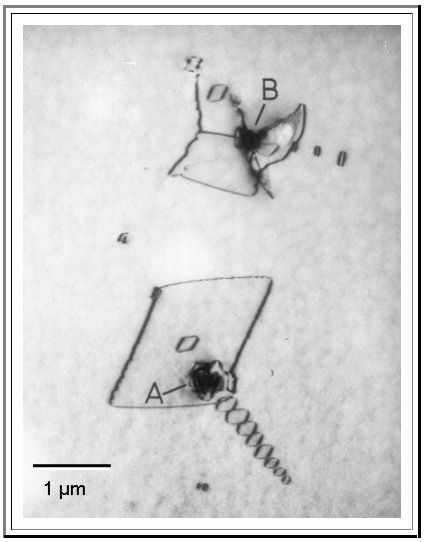


  - The defects shown are almost certainly **$FeSi_2$** precipitates, which often occur in "needle-shape". Some stacking fault and dislocation dipole components may also be involved. These needles are already very large; the defects labelled "**H**" may be a smaller needle.

Precipitates and Dislocations

Illustration

- Precipitates usually do not fit into the host lattice. The growing particle causes considerable stress that can be reduced by plastic deformation.
  - If the precipitate fits in one lattice direction, but not in others (a precipitate with an hexagonal lattice, e.g., may fit relatively well on the **{111}** planes of an **fcc** lattice) a compromise between a non-spherical shape of the precipitate and a system of dislocation loops in some direction may produce least strain energy. The precipitate-dislocation system then has a very specific structure; the process is known as "prismatic loop punching". An example is shown below on the left (taken under kinematic bright field conditions).



Precipitate with prismatic loops. An arrangement like that accounts for the peculiar etch features shown before

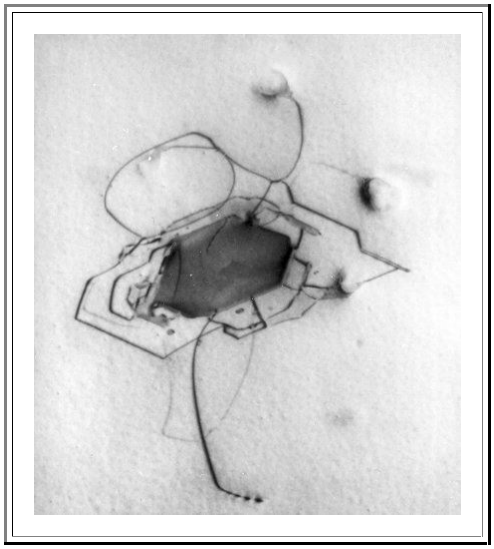


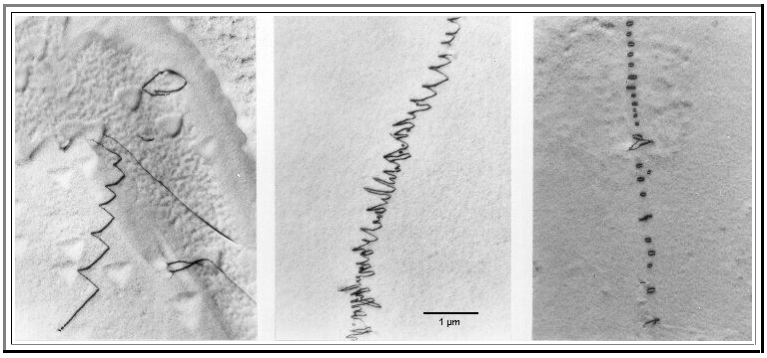
Plate-like precipitate (the dark grey feature) with dislocations relieving parts of the stress.

  - The two precipitates ("**A**" and "**B**") are seen as dark shapes; their nature is unclear, but they are probably **SiO_2**

Helix Dislocations

- Screw dislocations can climb, too. As a result they turn into a helix; a real "screw". The segments finally may collapse into a system of coaxial dislocation loops.
  - The images show screw dislocations in the emitter area of bipolar transistors in an early and advanced stadium of climb (right and center) and the final collapse into dislocation loops.

Illustration

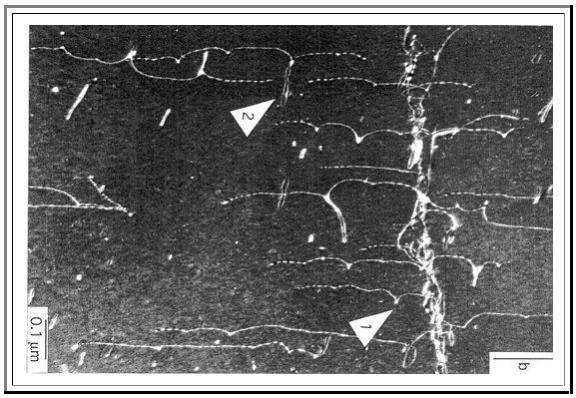


Dislocations in TiAl

Illustration

TiAl alloys are promising candidates for high strength and high temperature materials. A major concern for all structural materials are all mechanisms of plastic deformation, including creep and fatigue, especially at high temperatures.

- In the following picture dislocations in a **TiAl** alloy are shown. In contrast to "normal" pictures, they are heavily bowed out. This is special, because **TEM** specimens are no longer under the applied stress used while making mechanical tests and thus are expected to snap back to a rather straight line in the TEM specimen.
- In this case, **debris** from prior plastic deformation (visible as whitish specks), precipitates and possibly point defects keep the dislocations firmly anchored. At several point (e.g. at "**1**" and "**2**"), the dislocation can not overcome an obstacle and pulls out long dipoles.

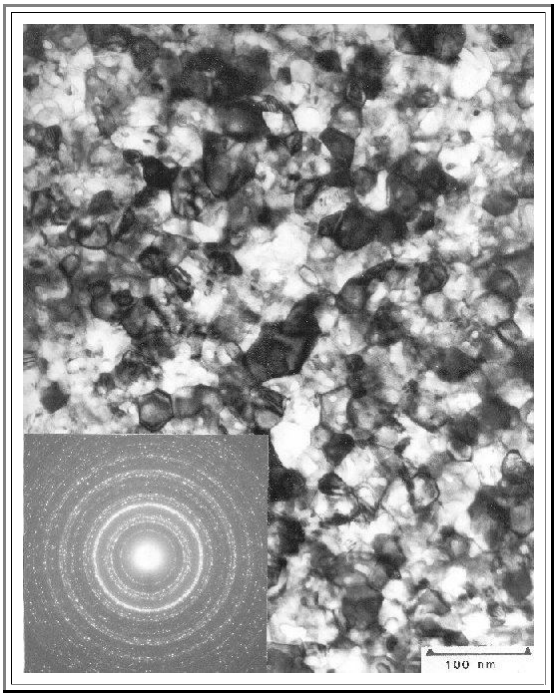


- This picture is from D. **Appel** from the research center **GKSS** in Geesthacht.

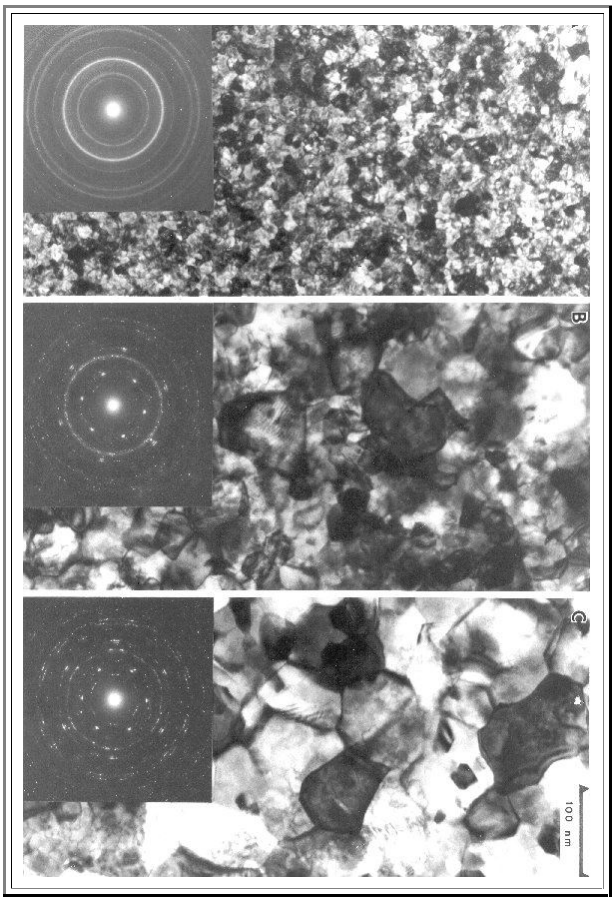
PtSi Silicide on Silicon

Illustration

- Metal Silicides play an important role in microelectronics. **PtSi** has been used in bipolar technology for quite some time; other silicides abound in **MOS** techniques.
  - Silicides are usually formed by evaporating a thin metal layer (here **Pt**) on a **Si** substrate, which is subsequently annealed at some high temperarture; say **800 ^{o}C**. Silicides form by solid state reactions, the picture below shows one result. A fine grained film of **PtSi** has formed in this case.
  - The picture illustrates that in polycrystalline materials the images are dominated by grain boundaries. The contrast conditions are pretty random and different in every grain. Not much can be seen.
  - The diffraction picture, shown as an insert, often provides more important information than the direct image. It consists of many reflexes arranged in rings; typical for polycrystalline materials. Every spot comes from one grain that happens to meet the Bragg condition for the particular reflex.



  - Increasing the annealing time or the annealing temperature makes the **PtSi** film more coarse grained; this is easily seen in the sequence below. But only the diffraction image shows that an epitaxial relationship to the Si substrate develops at high temperatures.

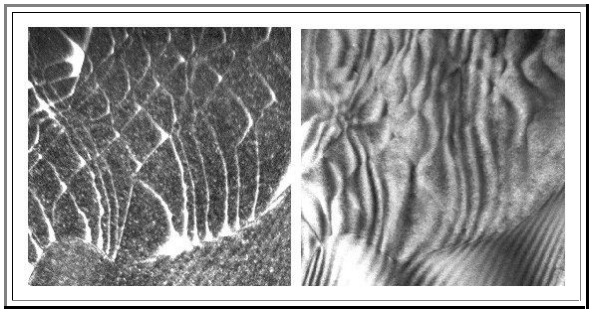


In the top picture the grains are so small that their diffraction pattern forms structureless rings. In the two lower pictures, however, some grains are still at a random orientation producing reflexes somewhere on the rings, but many grains have the same orientation producing strong spots at the same position -there is an epitaxial relationship to the substrate. This can be seen by closely inspecting the diffraction pattern: The spots from the epitaxial **PtSi** grains are almost coincident with the **Si** spots.

Comparison of Weak-Beam and Bright Field Conditions

Illustration

- The two pictures below show the extremes in resolution.
  - On the left hand side is a weak beam image of a dislocation network in a small angle grain boundary in **Si**; it has optimum resolution. The dislocation end at a **SiO_2** precipitate which shows faint fringes due to Moirée effects (the **Si** precipitate is sandwiched between **Si** crystals which are slightly misoriented).
  - On the right hand side is the same area imaged with (rather dynamical) bright field conditions. The dislocation lines are very broad and their images interfere with each other; it would be difficult to interpret this picture.



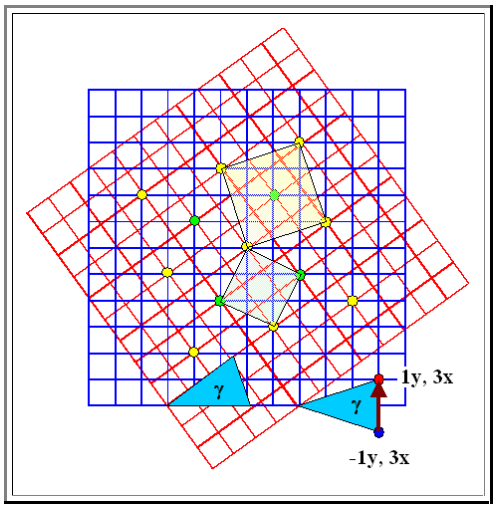
Σ is Always Odd

Advanced

- They most conspicuous issue in the **CSL** theory of grain boundaries is that there are no *even* values for Σ!
  - Try as you might - you will never find a **Σ = 2** boundary or any other *even* number in the literature. Now why is this? Mostly no explanation is given.
- A rigorous proof essentially needs the full power of the **O**-lattice theory, so it can not be easily given. But the general reason for this peculiar geometric fact can be envisioned as follows.
  - First, remember that *any* grain boundary can be obtained by generating grain **II** out of grain **I** by *one* rotation around a suitable axis with the rotation angle ϒ.
  - This means that we can produce all **CSL** orientations by looking at *one* rotation. We will do this for a square lattice, rotating around a **<100>** axis.
  - It is, however, not obvious that we can indeed produce all possible boundaries by this rotation, nor is it clear that the result will be valid for grain boundaries in non-cubic crystals. But it shows the direction of the argument.
- From all possible rotation, some will produce **CSL** structures. Which ones will do that is easily conceived:
  - The picture below shows a blue crystal **I**. Taking its origin at the apex of the blue triangle on the right, we see that we always will get a **CSL** orientation if we look at lattice points with the coordinates $(x, -y_0)$ which we may express as $(n, -1)$ if we set $x_0, y_0 = 1$, and than rotate the crystal so the the y-coordinate changes from **–1** to **+1**. The shift is indicated by the bold brown vector; we need to rotate an angle ϒ given by

$$\gamma = \tfrac{1}{2}\,\mathrm{cotg}\,\frac{y}{x} = 2\cdot\mathrm{cotg}\,\frac{1}{n}$$

  - The red lattice has been rotated by just the right amount to move the point **(3, –1)** to the position **(3, +1)**; the rotation center is in the middle of the crystals



- With this procedure we created the yellow **CSL** lattice.
  - Its **Σ'** value is given by its area divided by the are of a unit cell of the lattice; we have

$$\Sigma' = \frac{(x^2+y^2)^2}{x_0\cdot x_0} = \frac{(3x_0)^2 + (1x_0)^2)}{x_0^2} = (3^2+1^2) = 10$$

  - Its easy to generalize for **CSL** sites generated by moving the point $(nx, -y)$ on the $(nx,+y)$ position, we obtain for the Σ' values

$$\Sigma'(n) = n^2 + 1^2$$

- The result will be
  - **Σ'** is an *odd* number, if ***n*** is an even number (The square of an even number is even plus **1 = odd**)
  - **Σ'** is an *even* number, if ***n*** is an odd number (The square of an odd number is odd plus **1 = even**.)

So we can get even and odd numbers for Σ????.

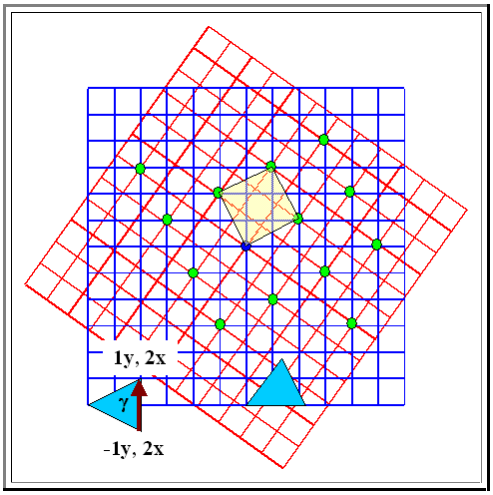
- Yes - but upon inspection you will find that for ***n* = odd**, there is *always* an additional coincidence point in the center of the lattice defined by the **CSL** points produced by the rotation, while for even numbers of ***n*** this is not the case.
- In the picture above this are the green points, and the lattice constant of the **CSL** lattice is now smaller. The Σ value in this case is simply

$$\Sigma = \frac{n^2 + 1^2}{2^{1/2} \cdot 2^{1/2}} = \Sigma'/2 = \text{an odd number}$$

- Instead of a **Σ = 10** boundary, we generated a **Σ = 5** boundary and there are no even Σ values.
- *q.e.d. (sort of)*

This, of course, is a far cry from a real mathematical proof, but it imparts the flavor of the thinking behind it.

- To complete this issue, the following picture shows the result for a rotation that tranfers **(2, –1)** to **(2, +1)**

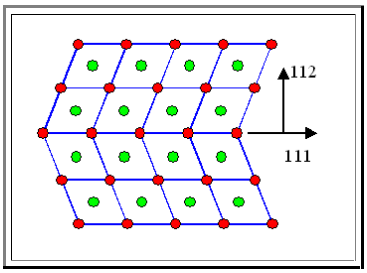


- There is no additional coincidenc point and we end up with a **Σ = ($n^2 + 1^2$) = 5** boundary, the same one as above

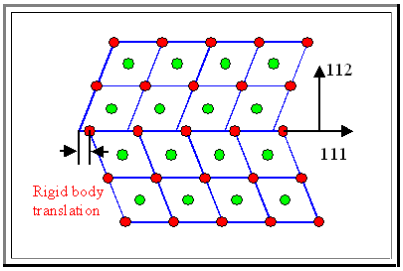
Rigid Body Translations

Advanced

- This is a somewhat special point; it only serves to illustrate that grain boundaries are complicated defects indeed.
  - Lets look at at twin boundary in a **bcc** crystal. The **bcc** geometry favors **{112}** planes for twins; a **<110>** projection of what you would expect would look like this:



  - However, what you get (according to calculations based directly on interatomic forces by **Vitek** in **1970** and subsequent **TEM** investigations) is something like this:



- There is some *rigid body translation* shifting one crystal with respect to the other one in the plane of the boundary.
  - The effect is due to the detailed nature of the interatomic potentials, but seems to be rather common. What will it do for structural considerations? Two things:
  - **1.** Besides the **5** parameters describing the geometry of the boundary that we encountered so far, we now need *three more*: The two components of the rigid body translation vector ***R*** in the plane of the boundary, and the (generally possible) third component perpendicular to this plane.
  - **2.** If there is some symmetry for ***R***, i.e there are several equivalent possible directions for the component in the boundary plane, different parts of the boundary may have rigid body translations along different directions. Wherever they meet, we need a *new kind of one-dimensional defect:* The boundary line between areas with different ***R***'s.
- OK - you get the drift: There seems to be some potential for complications - and mother nature certainly is aware of this!

Frank's Formula

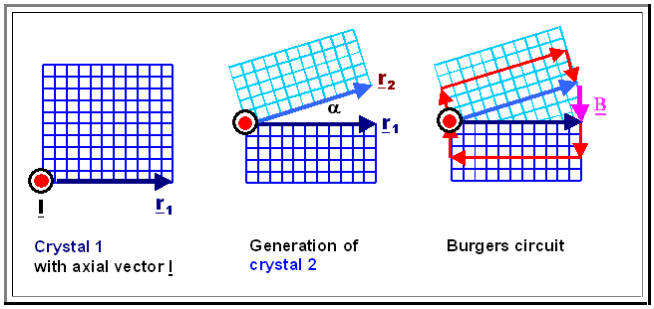
Advanced

Note: For ease of writing /reading in this module, variables are not in *italics*; instead vectors are underlined

- Franks formula relates **B**, the sum of all the specific Burgers vectors $\underline{b}_i$ cut by a vector **r** lying in the plane of the boundary, to the angle α with which one of the crystals is rotated with respect to the other one around the polar unit vector **l**. It is valid for small angles (say $\alpha < 10^o$) and given by

$$\underline{B} = (\underline{r} \times \underline{l}) \cdot \alpha$$

  - Note that we do not need three angles of rotation as required for a general grain boundary because we do not rotate around the axis of a coordinate system, but around the *polar vector* **l**.
  - Note also that the grain boundary plane (and thus **r**) is *not* required to be perpendicular to **l**. **r** thus can have *any* direction and length relative to **l**.
- For the derivation of Franks formula we consider a small angle grain boundary formed by rotating crystal **1** around an arbitrary axis **l** by α and thus forming crystal **2**. After that we join crystal **1** and crystal **2** on *any* plane.
  - A vector $\underline{r}_1$ in the plane of the grain boundary (to be) in crystal **1** thus gets transformed to a vector $\underline{r}_2$ in crystal **2**. Note that $\underline{r}_1$ does not have to be perpendicular to **l**.
  - Next, we make a Burgers circuit in the system with the small angle grain boundary and a reference circuit in the perfect crystal **1** (or crystal **2**). We will move along a vector $\underline{r}_1$ that is much longer than a lattice constant or the spacing of the dislocations that will make up the boundary.
  - In the perfect lattice we will start from the endpoint of $\underline{r}_1$ and move to the start of $\underline{r}_1$ in an e.g. counter-clockwise direction. In the crystal with the grain boundary, we do the same circuit, except that as soon as we switch over to grain **2**, we follow $\underline{r}_2$.
  - The whole procedure can be illustrated as follows:



  - There will be a closing failure **B** which must be identical to the sum of the Burgers vectors of all the dislocations contained in the circuit. Only the *components* of the **b**'s lying in the plane perpendicular to **l** are counted, of course.
  - For clarity, the vectors **r** are at right angles to **l** in the drawing, but this is not generally necessary.
- From vector calculus we know that a rotation can be described by an axial vector given by $\underline{R} = \underline{l} \cdot \alpha$.
  - The difference vector **B** between the two vectors $\underline{r}_1$ and $\underline{r}_2$ (with $\underline{r}_2$ produced from $\underline{r}_1$ by the rotation) than can be written as

$$\underline{B} = \underline{r} \times \underline{R}$$

$$= \alpha \cdot (\underline{r} \times \underline{l})$$

  - - and this is Franks formula from above.
- Note that there are two approximations in this. First, we assume *small angles* so that $\sin(\alpha) \approx \alpha$; and secondly, in the same vein, we assume $\underline{r}_1 \approx \underline{r}_2 = \underline{r}$.
  - Of course, we also assume that there is a smooth cross-over at the boundary (or that **r** is so large that to give or take parts of a lattice constant doesn't matter).
- This is a simple formula, but like most vector formulas, it has some hidden power. Before we look into the power of Franks formula a little more closely, we will consider what it *cannot* do:

- The formula gives the *net* content of Burgers vectors in a small angle grain boundary, but *not necessarily the arrangements of the dislocations*. It does not, of course, say anything about possible splitting into partial dislocations either. This means that there might be several arrangements of dislocations with the same **B**. The one that will be observed will be (most likely) the one with lowest total energy.
- No elastic distortion is considered. Between the dislocation the lattice is perfect; elastic distortion is present only in the core regions of the dislocations.

Bearing this in mind, lets look at some special cases. Since Burgers vectors are translation vectors of the lattice, *in general* three sets of non-coplanar dislocation will be required to produce the vector **B**. *Special cases* therefore are boundaries where only one or two sets of dislocations are needed.

If we have a boundary where *one* set of dislocations with Burgers vector $\mathbf{b_1}$ is sufficient, **B** can be written as

$$\mathbf{B} = N \cdot \mathbf{b} = \alpha \cdot (\mathbf{r} \times \mathbf{l})$$

- With **N =** number of dislocations cut by **r**

This obviously, looking at Franks formula, requires **b** to be perpendicular to **r** and **l**.

- The direction of **r** in the plane of the boundary is arbitrary; this means that **b** must be at right angles to the plane of the boundary or parallel to the normal **n** of the boundary plane and **l** must be at right angles to **n**; it follows that **l** must be contained in the boundary.
- If we now chose the particularly simple case of $\mathbf{r} = \mathbf{r_p}$ being parallel to **l**, we obtain $(\mathbf{r_p} \times \mathbf{l}) = 0$, which means that no dislocations are intersected by $\mathbf{r_p}$, implying that the dislocation lines must be parallel to the rotation axis **l**.

This leaves room only for the conclusion that *a boundary with only one set of dislocations must be a pure tilt boundary*.

- The spacing of the dislocations is obtained if we take $\mathbf{r} = \mathbf{r_{ra}}$ at right angles to **l** thus intersecting the dislocations lines at right angles, too. In this case we can write $\mathbf{r_{ra}}$ as $\mathbf{r_{ra}} = r \cdot (\mathbf{l} \times \mathbf{n})$ and obtain

$$\begin{aligned} N \cdot \mathbf{b} &= \alpha \cdot (\mathbf{r} \times \mathbf{l}) \\ &= \alpha \cdot r \cdot [(\mathbf{l} \times \mathbf{n}) \times \mathbf{l}] \\ &= \alpha \cdot r \cdot \mathbf{n} \end{aligned}$$

- With $\mathbf{b} = \mathbf{b} \cdot \mathbf{n}$ and the spacing **d** between the dislocations given by **d = r/N**, we obtain for the spacing $\mathbf{d_{tilt}}$ of dislocations in a pure tilt boundary with the boundary plane at right angles to the Burgers vector the relation used before:

$$d_{tilt} = \frac{b}{\alpha}$$

Similar considerations, which are straight forward but quite involved, can be made for the case of small angle grain boundaries with *two sets of dislocations* and the possible subsets (e.g. Burgers vectors in the plane of the boundary for pure twist boundaries).

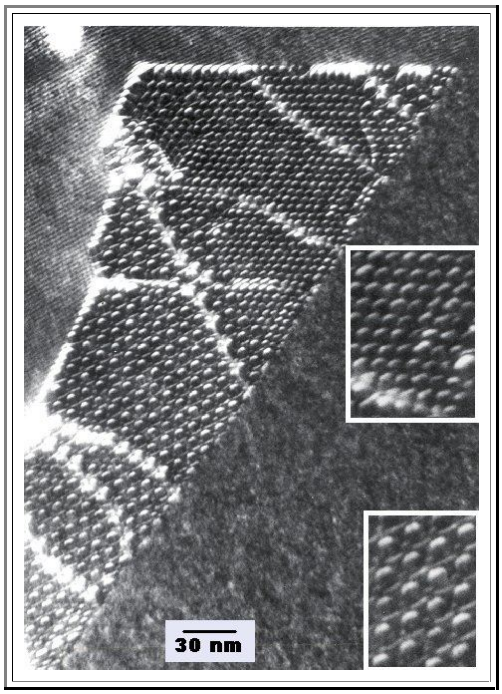
- For this and more, Hull and Bacons book can be consulted, which treats these cases in detail.

More important in the development of boundary structure theories is Bollmanns interpretation of Franks formula; which is the starting point of the **O**-lattice theory as will be discussed in the link.

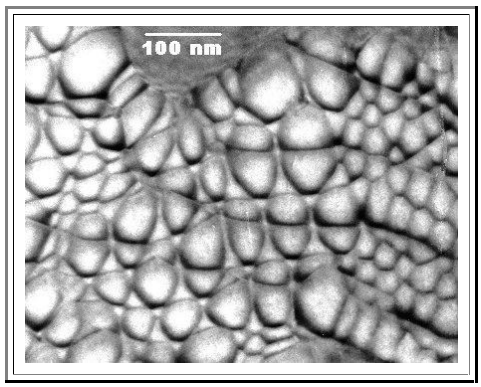
Complications in a Low Angle Twist Boundary on {111}

Advanced

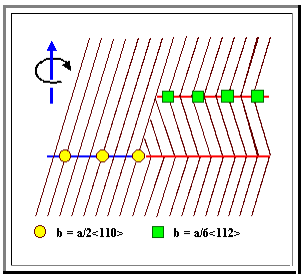
- Here is the picture of a small angle twist grain boundary once more at a larger size. The transition from one kind of network to the other kind is not prominent in the micrograph; the prominent white lines are extrinsic dislocation accommodated in the network.
- The topic in this module is not something you have to know for an examination, or for being considered knowledgable about defects. It just illustrates two general points about grain boundaries:
  - Even relatively simple situations can become very complicated (for us, not for the crystal).
  - Seeing a boundary in the electron microscope, and understanding what you see, is not the same thing.



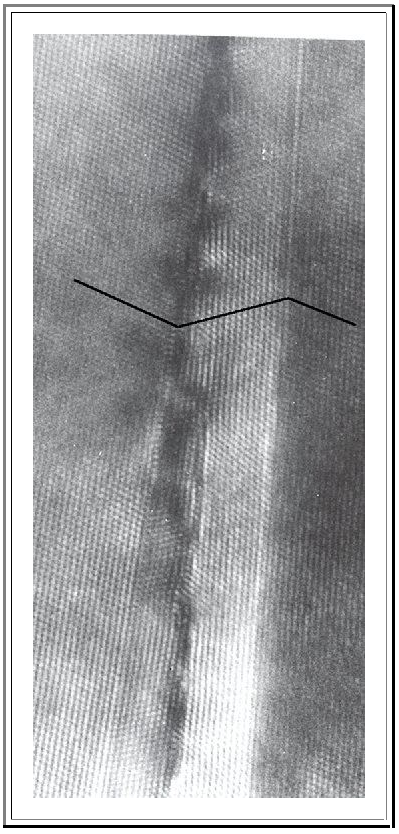
- Below is essentially the same situation showing the same kind of small angle twist boundary, but imaged with some special kind of bright field condition that gives good resolution and shows all dislocations at the same time. The changes in the network can be seen somewhat more clearly; the twinned region is on the right hand side.



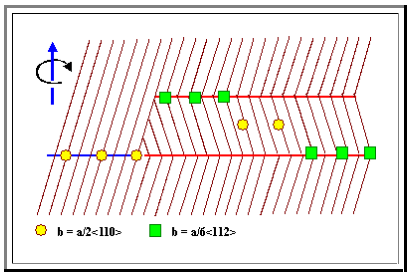
- In a cross-section, the whole structure looks like this:



- The traces of some **{111}** lattice plane across the twist boundary are shown; neglecting the (small) twist for sake of clarity.
- The boundary may split into *two* twin boundaries with a superimposed network of the proper **DSC** lattice dislocation (shown as green squares)
- This picture, however, gives rise to a question: Where is the network of **a/6<112>** dislocations? In the "upper" or "lower" twin boundary? Could it also be in between as a regular network of **a/2<110>** dislocations split into partials?



- This picture gives the answer. It shows a **HRTEM** cross-section of an artificially made small angle grain boundary in **Si**. It contains a large tilt component and probably some twist.
- The dislocations can be seen, partially as ending lattice planes, partially by a general localized disturbance of the picture (darkish areas).
- They obviously change form the perfect crystal (**Σ = 1** case) to one of the twin boundaries (**Σ = 3** case). In the **Σ = 1** case we would see the network with the extended and constricted stacking fault nodes, in the **Σ = 3** case we would see a regular hexagonal network of **a/6<112>** dislocations.
- We thus may have to expect all the possibilities shown below:



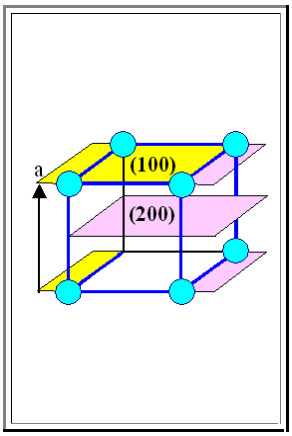
This means, among other things:

- If we see a dislocations network clearly belonging to a **Σ = 1** case, it does not mean that the grain boundary is not split into twins. Possibly, this may even be true for a **Σ = x** boundary, which may be split into two boundaries with **Σ = y** and **Σ = u** with a suitable relation between **y** and **u**.
- It is not easy to understand why the crystal does this. For our case we might surmise that the energy gain by having low energy **a/6<112>** dislocations in a **Σ = 3** interface is about equal to the energy needed to create the two twin boundaries since both possibilities occur without much preference between the two.

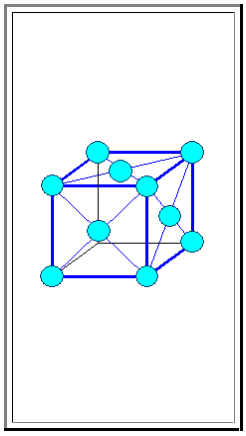
Stacking Faults in the DSC Lattice

Illustration

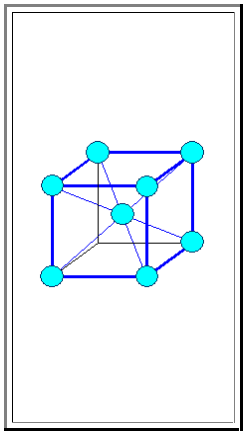
- In a *crystal* , a *lattice* point may be the seat of more than one atom, and the arrangement of atoms may have a higher degree of symmetry than the lattice.
  - In Bravais lattices, which are not necessarily the primitive lattices of a crystal, this feature expresses itself in the fact that lattice planes that *do not contain lattice points* of the elementary lattice, *may still contain atoms*.
  - Let's illustrate this somewhat abstract concept with the familiar **fcc** and **bcc** Bravais lattice with an atom on every lattice point



Cubic primitive lattice: Atoms are found on *all* **{100}** planes and on every *second* **{200}** plane. The *set* of **{100}** and **{200}** planes is shown on the left and right of the drawing, respectively



Cubic face-centered lattice: Atoms on *all* **{100}** planes and on all **{200}** plane with the same basic arrangement, just shifted by **a/2<010>** or **a/2<001>**.



Cubic body-centered lattice: Atoms on *all* **{100}** planes and on *all* **{200}** plane with the same basic arrangement, just shifted by **a/2<011>**.

- In terms of defects, this feature allowed for *stacking faults* in the *crystals*, which could not meaningfully exist just in the *lattice*.
- Well, the **CSL** lattice and the **DSC** lattice are *lattices*, after all. But physical reality still rests with the atoms. This may have somewhat exotic consequences.
  - The **DSC** lattice is a lattice that contains both lattices of the two crystals forming a **CSL** boundary as subsets. All atoms sitting on a *lattice point* of the crystal lattices therefore are also sitting on a *lattice point* of the **DSC** lattice
  - However, atoms *not* sitting on lattice points of the crystal lattices, may also *not* sit on lattice points of the **DSC** lattice. In analogy to the example above, there might be some additional symmetries hidden in the **DSC** lattice if we consider all atoms forming the crystals and their positions in the **DSC** lattice. In particular, the stacking of planes of the **DSC** populated with atoms may allow **stacking faults in the DSC lattice**, too, inextricably linked to **partial dislocations in the DSC lattice**.
- Since reality is (almost) always stranger than fiction, you *should expect* that this will happen - look out for stacking faults in the **DSC** lattice, and, as a corollary, **DSC** dislocation split into partial dislocations in the **DSC** lattice.
  - However, these defects in defects in defects may not be easy to find. Burgers vectors in the **DSC** lattice tend to be small which makes the contrast in **TEM** investigations (the only method with a chance at detecting this) rather weak. Partial **DSC** lattice dislocations would be even harder to see.
  - Moreover, the distance between secondary dislocations in the typical networks usually encountered, is mostly very small - there is not much room for splitting! Only in boundaries very close to a **CSL** orientation with roomy networks this effect may occur.
- So there is a good chance that you either never will see this or, if you see it, you may not recognize what you see.
  - You may even, and with good justification, be of the opinion that one shouldn't even look, because this topic is almost esoteric and is completely useless knowledge.
  - But some researchers *did* look and recognize - see below. Just for the hell of it, below the head of the article is reproduced as it appeared in "Phil. Mag."; i.e the **Philosophical Magazine**, which since its foundation in **1798** (which means it is one of the oldest science journals around) evolved into one of the major scientific journals covering **TEM** work in general and grain boundary stuff in particular.

PHILOSOPHICAL MAGAZINE A, 1981, VOL. 43, No. 1, 189–200

Partial secondary dislocations in germanium grain boundaries

I. Periodic network in a $\Sigma = 5$ coincidence boundary

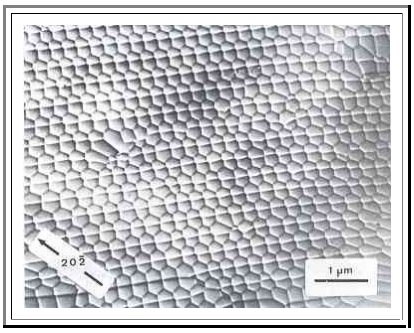
By J.-J. BACMANN, G. SILVESTRE and M. PETIT
Centre d'Etudes Nucléaires de Grenoble, Département de Métallurgie,
Service d'Etudes Radiométallurgiques, 85X-38041 Grenoble Cedex, France
and W. BOLLMANN
22 Chemin Vert, Pinchat, CH-1227 Carouge Geneva, Switzerland

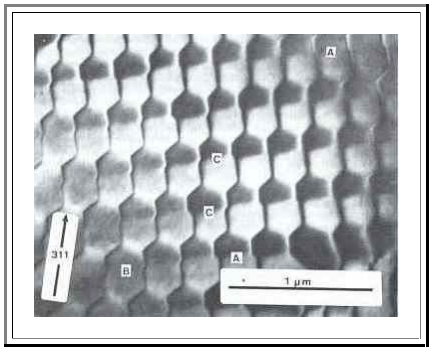
[Received 2 July 1979 and accepted 29 February 1980]

ABSTRACT

Interfacial dislocation networks have been studied in germanium bicrystals, using transmission electron microscopy. Periodic arrays of partial secondary dislocations, associated with a stacking-fault-like structure, have been observed in a near coincidence $\Sigma = 5$ grain boundary in which the deviation from exact coincidence orientation is a 0·05° rotation about [131], close to the boundary normal [130]. The dislocation grid is made of a honeycomb network of two partial secondary dislocations and a perfect secondary dislocation crossed by a set of parallel partial secondary dislocations. When some diffracting vectors common to the two crystals are used, areas of different contrast, limited by the partial dislocations, appear suggesting that the boundary is formed by two interfacial domains.

Here are two pictures of what they found.





- The left hand picture shows a dislocation network which is very unusual - nothing like it has ever been observed before with regular or **DSC** lattice dislocations. The right hand picture shows the same network, imaged under different diffraction conditions; the stacking faults in the **DSC** lattice are visible.

There is a second article directly following the first one with **Bollmann** as a first author. It analyzes the interaction of lattice dislocations in one of the crystals with the partial **DSC** lattice dislocations in the boundary.

- Not exactly easy stuff, it even taxed Bollmanns cunning. Suffice it to say that everything comes out as expected.

We may use this issue for a little test. Answer the question below for yourself and then click on the "Yes" or "No" according to where the majority of your answers are found.

Question	Yes	No
Do you consider knowledge about grain boundary dislocations apocryphal because it has no immediate technical uses?		
If in your work you run across pictures like the ones above, can you sleep well at night without knowing what they mean?		
Would you, as the referee, turn down a proposal to expand the **CSL/DSC** lattice theory to **6** dimensions in order to see if it can be used to describe *grain boundaries in quasicrystals* because the results - if any - are not only going to be completely useless for applications, but of interest to at most **100** researches in the world?		
In awarding a big Materials Science and Engineering price, would you prefer the person who stands behind a big new product (e.g. the blue **LED**) based on essentially known science, to the person who first explained some major, but useless material property that so far was not understood?		

Bollmanns Interpretation of Frank's Formula

Note: For ease of writing /reading in this module, variables are not in *italics*; instead vectors are underlined

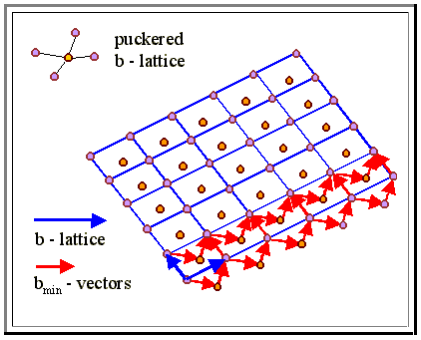
Frank's Formula Reconsidered

Advanced

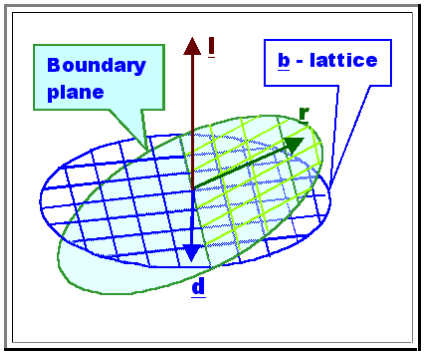
- Franks formula relates the sum **B** of all Burgers vectors cut by a vector **r** (which is required to be in the plane of the boundary) to the (small) rotation angle α around an arbitrary polar vector **l** that generates the second crystal from the first one. It states:

$$\mathbf{B} = (\mathbf{r} \times \mathbf{l}) \cdot \alpha$$

- Franks formula at this point is a *continuum equation*, it gives a value of **B** for every α and **r**
  - Burgers vectors, however, are *discrete*. This requires the vector **B** to be discrete, too.
  - Since Burgers vectors are translation vectors of the lattice, **B** can only be a sum of Burgers vectors.
  - If **l** is a lattice vector so that the "**b**-plane", the plane perpendicular to **l** that contains the possible Burgers vectors, is a lattice plane, too (i.e. it can be indexed with **{hkl}**, with **h, k, l** = integers). It contains lattice points that define the possible Burgers vectors in this plane.
  - Note that the Burgers vectors defined in this way must not necessarily be the shortest possible Burgers vectors $\mathbf{b}_{min}$, i.e. the Burgers vectors of real dislocations. It is, however, always possible to decompose the **b** vectors of the **b**-lattice into e.g., **a/2<110>** type Burgers vectors of the **fcc** lattice. This may require $\mathbf{b}_{min}$-vectors that are *not* contained in the **b**- plane - but all we have to do then is to imagine the net of **b**-vectors in this plane to be "puckered" as shown below.



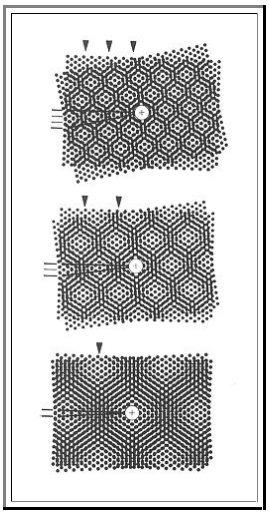
- In the plane of the boundary, an arbitrary **r** would intersect the projection of the **b**-lattice onto the boundary plane along the **l**-direction. In a schematic view we have the following situation:



- Franks formula can now be understood as a *discrete imaging of points in a two-dimensional "Burgers vector space" onto a plane in real space*.
  - The Burgers vector count along **r** (after translating it to smallest possible vectors $\mathbf{b}_{min}$) gives the number of dislocations that are found if going along **r** in the boundary plane. If even spacing is assumed, we also know the spacing in the particular direction given by **r**.
- Now comes something new. Remember that Franks formula did not make any statement about the arrangement of the dislocations, or - in other words - their line direction.
  - Bollmanns view is different: The line direction of the dislocation is obtained by probing the whole two-dimensional grain boundary space by sweeping **r** around. What happens then can be understood in purely geometrical terms, as we will see below.

Bollmanns view of Frank's Formula

First of all it is important to realize that the crossing of a dislocation by the "probing" vector **r** in the **b**-plane is directly imaged by the **Moirée pattern** of the superimposed two crystals obtained by rotating the **{hkl}** planes perpendicular to **l** on top of each other by α as shown below for three different α**s** with pictures from Bollmanns book.



- In the whitish (bright) areas, there is a high degree of coincidence of lattice points, whereas in the black areas the misfit is largest. These are of course the "**O**-points" in the full **O**-lattice theory.
- Whenever a vector from the origin crosses a black area to reach a whitish area again, the translation relative to an equivalent vector in the other lattice is just a lattice vector of the underlying plane, which is the **b**-plane in our definition. In other words, if you move to the same white area in crystal **I** and crystal **II**, the two vectors are on top of each other. But their tips would be separated by just a lattice vector if you now rotate the crystals back to a no-boundary situation.
- If the crystal now introduces a boundary, it will increase the whitish areas, the areas of best fit, and concentrate the misfit in the black areas - which correspond to the dislocations. A periodic structure results which we can describe as a lattice - the (**2**-dimensional) **b-lattice**.
- Since the Moirée pattern does not depend on the position along **l**, we can extend the **b**-lattice along the **l** direction and obtain a **3**-dimensional structure with lattice lines instead of points. If we enclose the lattice points (respectively lines) of the **b**-lattice in **Wigner-Seitz cells**, we obtain a kind of honeycomb structure.

The decisive point now is that the *boundary plane*, which can have *any* position relative to **l**, intersects this "honeycomb" **b**-lattice somehow, for an arbitrary case we obtain the following picture (again taken from Bollmanns book)

- The **b**-lattice consists of the yellow lattice points. It is turned into the three-dimensional "honeycomb" lattice by introducing Wigner-Seitz cells (blue lines) and continuing it along the **l** direction (magenta arrows). The lines would be the **O**-lattice in the full theory.
- An inclined boundary plane will have a dislocation wherever the boundary plane intersects the honeycombs. The resulting dislocation network is shown with red lines. The points of best fit (red points) are in the center of the network as it must be.
- The final network may still be different because the Burgers vectors of the dislocations now defined by the "red lines" might be too large and decompose, as pointed out above.

The final interpretation now is as follows:

- Wherever the boundary plane intersects a cell wall of the (three-dimensional) **b**-lattice, we have a dislocation with the Burgers vector as defined by the translation vectors in the **b**-lattice. The lines defined by the intersection of the boundary plane and the cell walls then directly define the dislocation lines - we get a direct rendering of the dislocation network in the boundary.
- Of course, the geometry of the dislocation network obtained in this way depends on the kind of unit cell we chose for the "honeycomb" **b**-lattice. Wigner-Seitz cells, while universal, may not be best choice possible. But it is always possible now to "develop" the network obtained to a *network with minimum energy* by using the rules of dislocation interaction as in the example with the small angle twist boundary on a **{111}** plane.

These and other complications need more considerations. However, remembering that Franks formula is an approximation and covers only small angle grain boundaries, it is not worth the effort to improve this limited theory. It is a better at this point to unleash the full power of **O**-lattice theory which contains Franks formula as a special case.

The concepts behind Bollmanns interpretation of Franks formula are not easy and lead into very deep water. Let's recapitulate the essential ideas:

- The orientation relationship between the two crystals (expressed here as *one* rotation) always leads to a kind of Moirée pattern that can be identified as a Burgers vector lattice (**b**-lattice) describing the localized displacements necessary to match the two crystals on some boundary plane.
- The **b**-lattice can be extended to three dimensions (the "honeycomb" lattice for the case treated here). The cell walls of this three-dimensional lattice define the dislocation content of the boundary and the Burgers vectors encountered in crossing a wall.
- The intersection lines obtained by cutting the three-dimensional **b**-lattice with the boundary plane defines directly the dislocation network.

Matrix Algebra

Matrices, Tensors, and Coordinate Transformations

Basics

- For sake of clarity we do not write variables in *italics* in this module

- This is not a course in matrix algebra (including vector and tensor calculus), but a quick reminder, assuming you know the basic facts of life here.
  - We also cut a lot of corners, not distinguishing much between matrices (a mathematical object) and tensors (a physical object), "true" vectors and "polar" vectors, Cartesian and non-Cartesian coordinate systems and the like.
  - We will deal with some topics of matrix algebra roughly in the sequence they come up in the backbone chapters

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

  - A matrix then is an assembly of nine numbers arranged as shown on the left.

- In a simplified way of speaking, a **matrix** (or better **tensor**) allows to correlate **vectors** in a simple linear way.
  - Every component of the vector $\underline{r} = (r_1, r_2, r_3)$ can be expressed as a linear function of all components of a second vector $\underline{t} = (t_1, t_2, t_3)$ by the equations

$$\begin{aligned} r_1 &= a_{11} \cdot t_1 + a_{12} \cdot t_2 + a_{13} \cdot t_3 \\ r_2 &= a_{21} \cdot t_1 + a_{22} \cdot t_2 + a_{23} \cdot t_3 \\ r_3 &= a_{31} \cdot t_1 + a_{32} \cdot t_2 + a_{33} \cdot t_3 \end{aligned}$$

  - In matrix notation we simple write

$$\underline{r} = \mathbf{A} \cdot \underline{t}$$

  - With **A** being the symbol for the matrix defined above.
  - We then have already defined how a matrix is multiplied with a vector and that a new vector is the result of the multiplication.

- The matrix **A**, if interpreted as an entity that relates two vectors with each other, must have *certain properties* that are not required for a general matrix (that might express, e.g., the coefficients of a linear system of equations with several unknowns).
  - If we change the coordinate system in which we express the vectors, the components of the vectors will be different numbers, but the vectors themselves (the arrows) stay unchanged. This imposes some conditions on the set of nine numbers - the matrix - connecting the components of the vectors and *any matrix meeting these conditions we call a tensor*.
  - A *tensor* thus is a set of nine numbers, and the numerical value of these numbers depends on the coordinate system in which the tensor is expressed. If we do a coordinate transformation, the numerical values of the nine components must then transform in a specific way.

- Transforming a coordinate system into another one is done by matrices as follows:
  - If the first vector $\underline{r}$ is chosen to be one of the *unit vectors* defining some Cartesian coordinate system, the second vector $\underline{r}'$ obtained by multiplying $\underline{r}$ with the transformation matrix ***T***, can be interpreted as the *unit vector* of some new coordinate system
  - The set of unit vectors $\underline{r}_i$ with **i = x,y,z** will be changed to a new set $\underline{r}'_i$ by

$$\underline{r}'_i = \mathbf{T} \cdot \underline{r}_i$$

- and **T** is called the **transformation matrix**. It is clear that **T** must have certain properties if the $\underline{r'}_i$ are also supposed to be unit vectors
- While this is clear, it is not so clear what we have to do if we want to reverse the transformation. The simple thing is to write

$$\underline{r}_i = (T^{-1}) \cdot \underline{r'}_i$$

- and defining (T^{-1}) to be the **inverse matrix** to **T** so that the operation can be reversed.

But how do we calculate the numerical values of the components of T^{-1} if we know the numerical values of the components of **T** ??

- In order to be able to give a simple formula, we first have to introduce something else, the **determinant of a matrix**

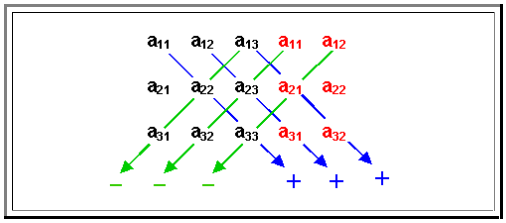
The Determinant of a Matrix

The determinant **|A|** of a matrix **A** is a *single number* calculated by summing up the diagonal products in a special fashion.

- For a **3 × 3** matrix we have

$$|A| = \begin{array}{l} a_{11} \cdot a_{22} \cdot a_{33} + a_{12} \cdot a_{23} \cdot a_{31} + a_{13} \cdot a_{21} \cdot a_{32} \\ - a_{13} \cdot a_{22} \cdot a_{31} - a_{11} \cdot a_{23}a \cdot {}_{32} - a_{12} \cdot a_{21}a \cdot {}_{33} \end{array}$$

- Look at the written matrix **A** above and you see that you start by doing the products by going down diagonally from left to right, adding the products of the three possible diagonals - always completing a diagonal by repeating the matrix if necessary. Then you subtract the product you obtain by going down the diagonal from right to left.
- This sounds more complicated as it is; graphically it looks like this:



The determinant of a matrix obtained in this way is a number that comes up a lot in all kinds of matrix operation; the same is true for a related quantity, the **subdeterminant** A_{ik} of the matrix **A**

- There are as many subdeterminants as there are elements in the matrix. A_{ik} is obtained by

  1. Erasing the line and the row that contains the element a_{ik} and calculating the determinant of the matrix that remains, and
  2. multiplying the number obtained by $(-1)^{i \cdot k}$

With the concept of a subdeterminant, we can also define the **rank of a matrix**:

- The rank of a matrix is the number of row (or columns, resp.) of the determinant or largest subdeterminant with non-zero value. In other words, the rank of a **3 × 3** matrix **A** is **rank(A) = 3** if $|A| \neq 0$; if **|A| = 0**, you look for the largest subdeterminant

With determinant and subdeterminant, the *inverse matrix is easy to formulate* :

- The inverse matrix A^{-1} to **A** has the elements $(a_{ik})^{-1}$ given by

$$(a_{ik})^{-1} = \frac{A_{ki}}{|A|}$$

- i.e. the value of the respective subdeterminant divided by the value of the determinant. Note that the indexes are interchanged ("**ik**" ⇒ "**ki** "); and that the " $^{-1}$" must be read as " *inverse*", it is not an exponent!!!
- We will not prove it here; but it is not too difficult - just solve the system of equations given above for the t_i.

Two more important points follow directly:

- An inverse matrix **(A $^{-1}$)** to **A** only exists if the determinant of **A** is not zero!
- The product of A^{-1} and **A** results in the **identity matrix I**

$$A^{-1} \cdot A = I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The last claim is unproved, we first need the multiplication rule for matrices to prove it.

- Multiplication of the matrix **A** with the matrix **B** gives a new matrix **C**; and the element c_{ik} of **C** is obtained by taking the scalar product of the "line" or "row" vector in row **i** of matrix **A** times the column vector of column **k** of matrix **B**. This is best seen in a kind of graph:

$$\begin{pmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & c_{32} & \times \end{pmatrix} = \begin{pmatrix} \times & \times & \times \\ \times & \times & \times \\ a_{13} & a_{12} & a_{32} \end{pmatrix} \cdot \begin{pmatrix} \times & b_{12} & \times \\ \times & b_{22} & \times \\ \times & b_{32} & \times \end{pmatrix}$$

- Now it is still fairly messy, but straightforward to prove the claim from above - you may want to try it.

A useful relation is that the multiplication of any matrix with the identity matrix **I** doesn't change anything.

$$I \cdot A = A$$

- And this is also true for multiplying a vector with **I** :

$$I \cdot \underline{r} = \underline{r}$$

From the various definitions you may get the feeling, that signs are important and possibly tricky. Well, that's true.

- Matrix multiplication, in general is not commutative, i.e. $A \cdot B \neq B \cdot A$ - you must watch out if you multiply from the left or from the right.
- Still, we now can solve "mixed" vector - matrix equations. Take, for example

$$\underline{r}_0 = A^{-1} \cdot \underline{r} + \underline{T}(I)$$

- Multiplying from the right with **I** yields.

$$I \cdot \underline{r}_0 = I \cdot A^{-1} \cdot \underline{r}_0 + I \cdot \underline{T}(I)$$

$$I \cdot \underline{r}_0 = A^{-1} \cdot \underline{r}_0 + I \cdot \underline{T}(I)$$

- That looks a bit stupid, but with this cheap trick we now we have only tensors in connection with $\underline{r}_0$, which means we now can combine the "factors" of $\underline{r}_0$, giving

$$(I - A^{-1}) \cdot \underline{r}_0 = \underline{T}(I)$$

our **O**-lattice theory master equation.

One last important property of transformation matrices is that their determinant gives directly the volume ratio of the unit cells:

$$|A| = \frac{\text{V(after the transformation)}}{\text{V(before the transformation)}}$$

This is not particularly easy to see, but simply consider two points:

1. The base vector **a(I)** is transformed to the base vector **a(II)** via

$$\underline{a}_i(II) = A \cdot \underline{a}_i(I)$$

2. The volumes **V** of elementary cells is given by

$$V = (\underline{a}_1 \times \underline{a}_2) \cdot \underline{a}_3$$

Since we produce the **O**-lattice from a crystal lattice with the matrix $I - A^{-1}$, the volume V_O of an **O**-lattice cell (in units of the volume of a crystal unit cell) is

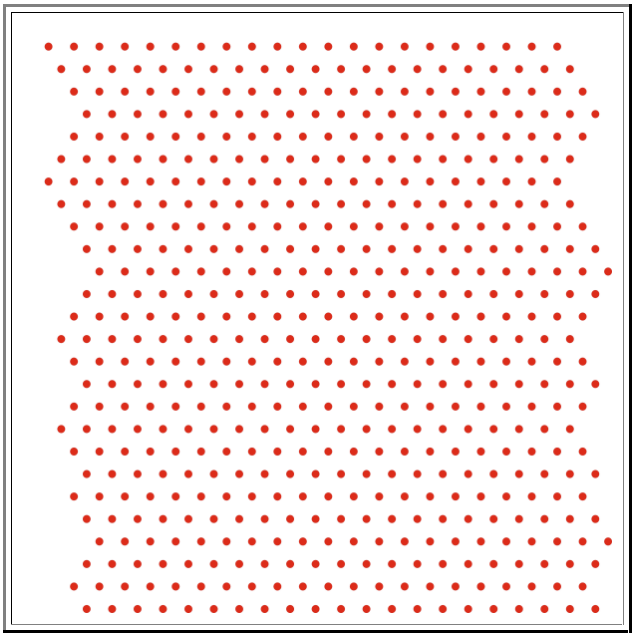
$$\frac{1}{V_O} = |I - A^{-1}|$$

Again, as remarked above; watch out for signs.

Animation: Two-Dimensional Coincidence Lattices Obtained by Rotating Two Hexagonal Lattices

Illustration

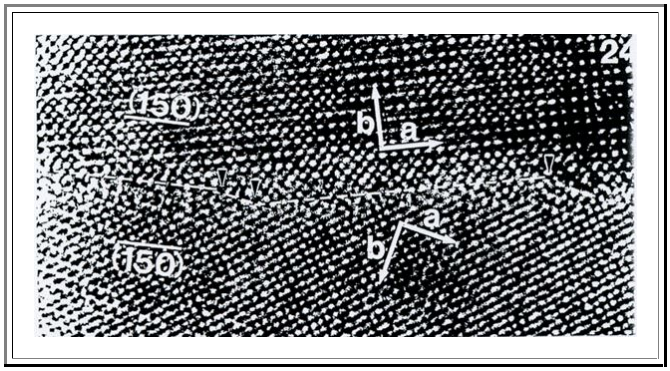
- The animation rotates a hexagonal lattice on top of an identical one. For certain angles some lattice points coincide. This is emphasized by stopping the rotation for a few seconds.
  - Due to small aberrations in the drawings, the coincidence is not perfect, but close enough to show the development of a "**Coincidence Site Lattice**"



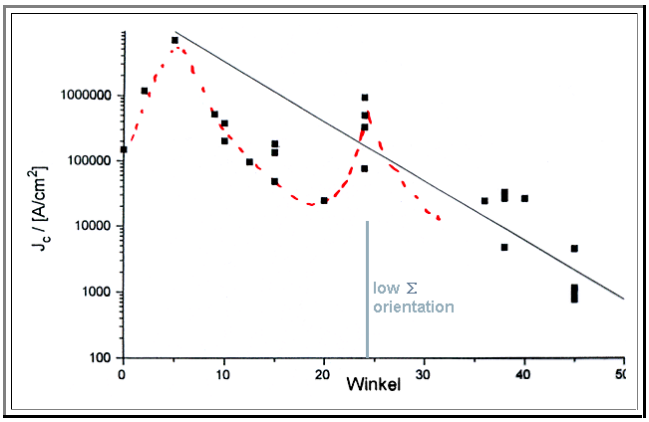
Grain Boundaray in Superconducting $YBa_2Cu_3O_7$ and Critical Current Density

Illustration

The **HRTEM** picture shows a grain boundary in the famous high-temperature superconductor **$YBa_2Cu_3O_7$**. This boundary was intentionally made with a specific orientation. Facetting can be clearly seen. The picture and the following results are from the **TEM** group of Prof. Urban from the Jülich Research Center.



A whole collection of boundaries was made and the critical current densities measured. The critical current density is the current density at which the superconducting state of the material is destroyed. The relation obtained is shown below.

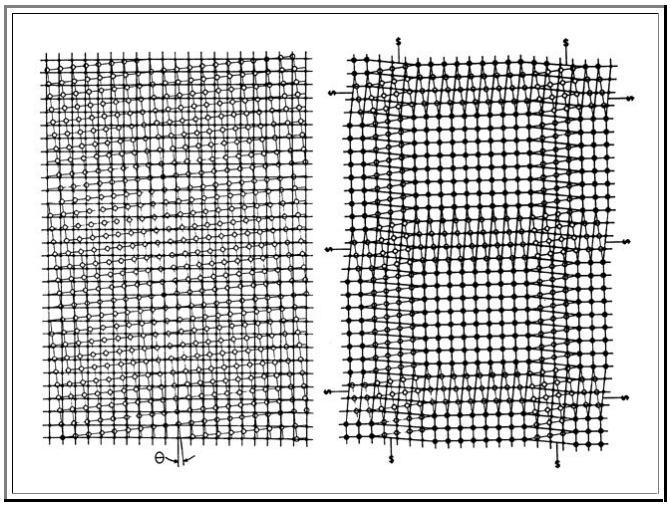


In general, the critical current density decreases wit increasing misorientation. But around a specific misorientation it has a remarkable maximum. It comes as no surprise that this specific misorientation corresponds to a low Σ orientation.

Generation of a Screw Dislocation Network in a Small Angle Twist Boundary

Illustration

- This is a famous drawing (probably coming from **Read**) that can be found in just about any book on dislocations.
  - It shows the direct superposition of two twisted cubic lattices on the left, and the formation of a screw dislocation network with perfect (**= Σ1**) areas in between.

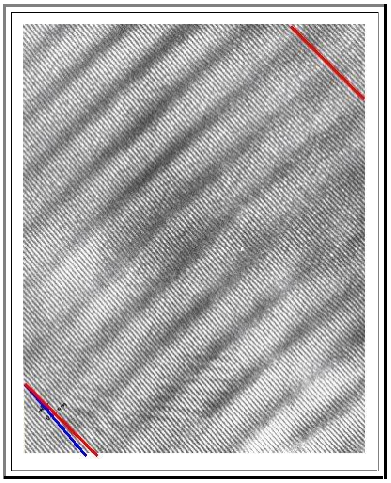


HRTEM of Screw Dislocations in a Small Angle Grain Boundary

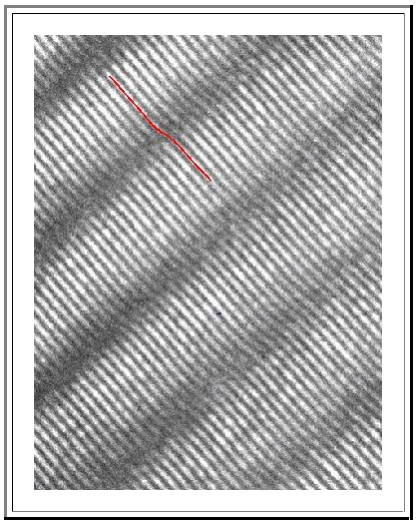
Illustration

The **HRTEM** picture shows sets of **{111}** lattice planes for a tilt boundary on **{111}**. There are three sets of screw dislocations, but only one set can be seen.

- The left picture is an overview. The small angle grain boundary is inclined (it is essentially contained between the two red lines); to the far left or right just one of the two grain is seen.
- Depending on the local thickness of the grains in the area where they overlap, grain one or grain two dominates the picture.
- Traces of the two grains are marked in red and blue; the red trace has been shifted to show the twist angle (and the extension of the boundary).



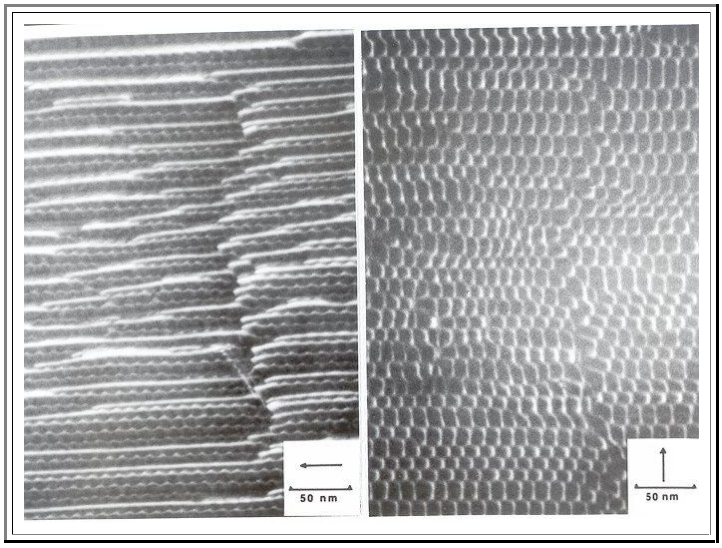
- The enlargement with a tracing of a **{111}** plane shows the effect of the screw dislocation.



Small Angle Grain Boundary with Twist and Tilt

Illustration

- On the right-hand side is the (full-size) picture from the backbone, on the left-hand side *the same area* is shown, but imaged with different weak-beam conditions.
  - It can be seen that a whole new set of dislocations lights up. They are edge dislocations accounting for a fairly large degree of (unintended) tilt in this grain boundary. They interact with the screw dislocations visible on the right hand side to form a fairly complicated network of grain-boundary dislocations.
  - The big distortion in the edge dislocation structure running from top to bottom in the right-hand side of the image is probably due to a change of the grain boundary plane: All dislocations must move "up" or "down": the structure changes.

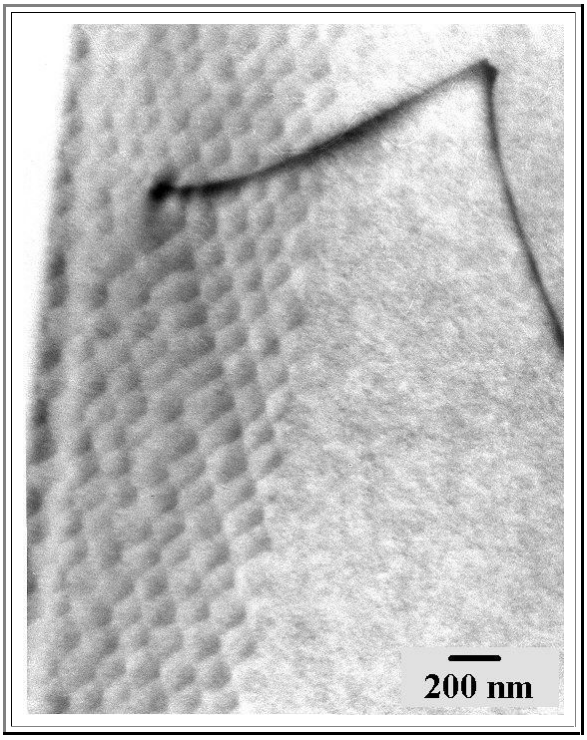


  - This is another good example of the power of contrast analysis with **TEM** and the difficulties of extracting the information contained in the picture (just try to draw the network with all the Burgers vectors indicated).

Extrinsic Dislocation Reacting with Grain Boundary Dislocations

Illustration

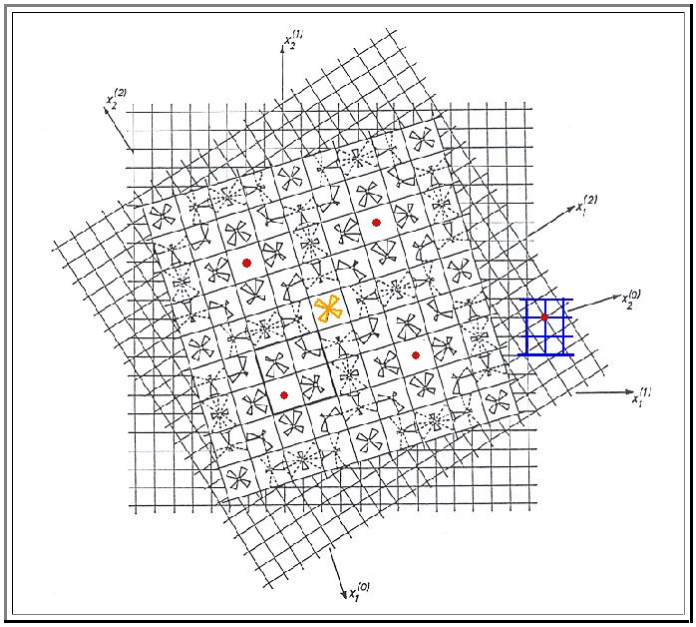
- Seen is some grain boundary with a well developed network of **DSC** lattice dislocations. Their Burgers vector must be rather small, because their contrast is weak (in comparison to the contrast of the lattice dislocation coming from the right).
  - The lattice dislocation clearly ends in the boundary and decomposes into grain boundary dislocations. The resolution is not good enough to show details, but the general disturbance of the network is evident.



Original "Pattern Drawing" from Bollmanns Book

Illustration

- This is a drawing from Bollmanns book (with some color added). It shows the pattern in **O**-lattice cells as defined in the backbone text.
  - This drawing probably had to be done by hand with a high precision - some work! Most of the other pictures in Bollmanns book are rather complicated too, and this explains why they are so poorly described in the captions and the text:
  - After conceiving and finishing a picture like this, you are so intimately familiar with everything it contains - you can't imagine that others are not immediately aware of everything it projects!
  - What exactly are the dashed lines in some of the pattern elements? It is not explained anywhere - can you figure it out?

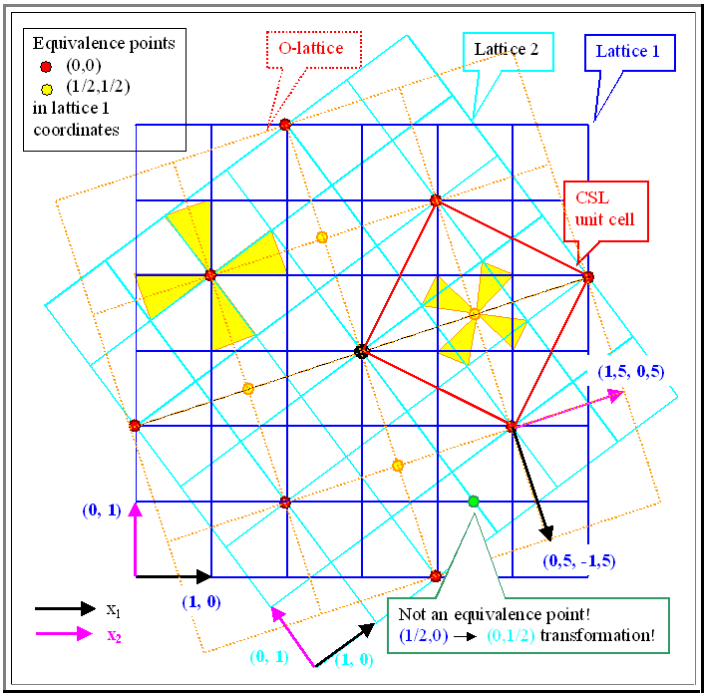


- This is an **O**-lattice belonging to a **CSL** lattice, i.e. a special **O**-lattice with a periodic structure; **Σ = 53** in this case. One oft the pattern elements was marked yellow.
  - The red points are *coincidence points* oft the lattices (i.e. the coinciding equivalence points are lattice points), they constitue the **CSL lattice**.
  - This is shown by drawing in one more **O**-point on the outside of the picture with one lattice marked in blue. It is evident that there is indeed a coincidence of the lattice points.
  - Note that all points in the center of the cells containing patterns are **O**-points, too. The **O**-lattice thus has a much smaller lattice constant than the **CSL** lattice.

Σ = 5 CSL and O-Lattice

Illustration

- Here is a large scale illustration of a **Σ = 5** relationship in the **O**-lattice.
  - There are more **O**-lattice points than lattice coincidence points - we have one extra equivalence point (equivalence coordinates **(1/2, 1/2)** in addition to the **O**-point with equivalence coordinates **(0, 0)**.
  - The lattice constant (and therefore the unit vectors) of the **O**-lattice are smaller by a factor of $\mathbf{2^{1/2}}$.



- In just looking at the picture, it is tempting to identify more **O**-points, the green point, e.g., looks very much like an **O**-point. Well, *it is not*, because:
  - The green point, while marking middle positions on both lattices, is *not* an **O**-point, because its internal coordinates are **(1/2, 0)** in lattice **1** and **(0, 1/2)** in lattice **2**. And while, yes, this marks a point midway between to lattice points in both lattices, it is still *not* an equivalence point!
- The picture contains a new, very important feature:
  - The yellow triangles denote **pattern elements**. While they indicate the rotation, they may simply be taken as *symbols representing the specific arrangement of atoms at specific equivalence points*. Equal symbols indicate equal arrangements, and identical equivalence points have identical pattern elements.
  - The unit vectors of the three lattices are also shown; this will be important in some future context.

If You Chose "No"

- You are a *scientist* and not an engineer.
  - There is nothing wrong with that. You will receive satisfaction from finding out things and from understanding how things work.
  - You may never experience the joy of having lived through a project that actually resulted in a *product*, which other people buy and use.
  - In compensation, you may on occasion experience the joy of being the only human who for a brief moment of glory found out something about nature that nobody else on this side of Pluto knows about [1)].
  - You probably will not earn (or at least not receive) a great deal of money and become the leader of men (and women), but probably you won't miss it very much.

[1)] An old joke along this line is: Hans **Bethe** (or take **Weizsäcker** if you like) was taking a walk with some girl friend at a nice summer night. The girl remarks on how *nice* the stars are shining. "Yes", says Bethe, "and right now I am the only one who knows *why* they are shining" - having just discovered the principles of the hydrogen fusion chain that fires the stars!

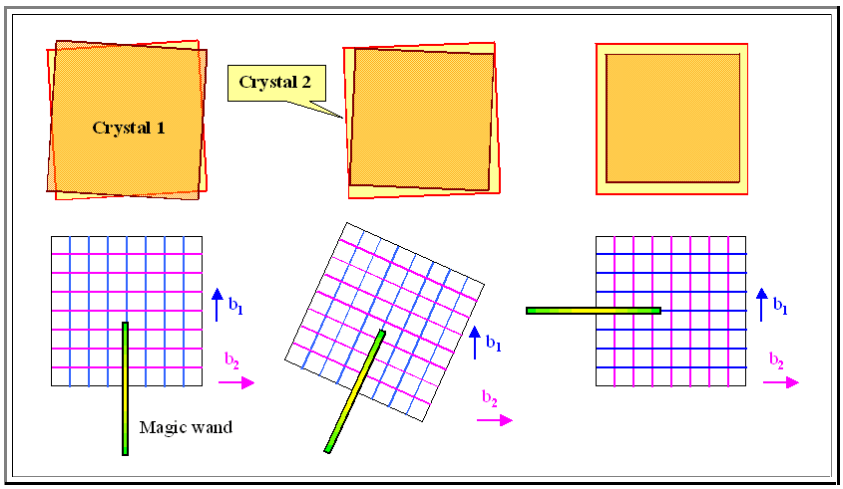
If You Chose "Yes"

- You are an *engineer* and not a scientist.
  - There is nothing wrong with that. You will receive satisfaction from making things work and from seeing mankind (or at least some chosen people, possibly including yourself) being better off because of your products.
  - You may never experience the joy of being the only one who for a brief moment understands something about nature that nobody else (on this side of Pluto) comprehends.
  - In compensation, you may have to accept a great deal of money and become a leader of men (and women).

Compliant Substrates

Advanced

- In **1997**, the idea came up to accomodate the stress in a phase boundary arising from the misfit by using existing defects some distance away from the interface which then may not be harmful to the device. In particular, a small-angle grain boundary some **100 nm** away from the phase boundary was found to do the job.
  - The concept is easy to understand on the background of the case studies for small angle twist boundaries discussed before.
  - Lets discuss how you can make a phase boundary *free of misfit dislocations even for misfits* ***> 10 %*** and layer thicknesses of many **nm**. We will do this in the form of a recipe, giving the ingredients with a brief discussion of what they do.
- Lets assume we want to produce a **GaAs** layer on top of a **Si** substrate (this is something a lot of people would love to do! [1)]). The misfit - roughly - is **10 %** so there is no chance whatsoever to produce a misfit dislocation free interface by just depositing **GaAs** on top of **Si**. We do it as follows:
  - Bond two **Si** wafers together with a (twist) misorientation of about $\mathbf{10^o}$. A small angle grain boundary will form that is identical to the one shown before - except that the spacing of the dislocations will be considerably smaller.
  - Polish off one of the wafers until only a layer with a thickness of a few **100 nm** remains. This is not exactly easy, but state of the art in wafer processing.
- Now you have a **compliant substrate**. Deposit your **GaAs** on top of it and be confident that you have no misfit dislocations in the phase boundary.
- How does this work? On the one hand, the details are none to clear, one the other hand, it is simple. We look at the other hand.
  - Imagine a ***magic wand*** that you can glue to the screw dislocation network in the small angle grain boundary. Now hold your substrate crystal firmly in place, and *rotate the complete dislocation network by* $\mathbf{90^o}$. What then happens is shown below.



  - If you rotate the dislocation network by $\mathbf{90^o}$, you produce an *edge dislocation network*. Remember that the *Burgers vector* is fixed; it does not depend on the direction of the *line vector* - which is the only vector you change by the rotation.
  - The spacing ***d*** of the dislocation network remained unchanged and it is now exactly the kind of network you need to accommodate differences in lattice constants. Compare the networks in the small angle twist boundary in **{111} Si** with the network in the phase boundary **{111}Si - (hex)$NiSi_2$**. While the networks are identical in geometry, one consists of *screw dislocations*, the other one of *edge dislocations*.
  - In the twist boundary, the misorientation angle was given by (approximating $\mathbf{\sin(\alpha)} \approx \alpha$):

$$\alpha = \frac{b}{d}$$

  - For an edge dislocation network, the misfit in lattice constants is simply

$$d = \frac{b \cdot a}{\Delta a}$$

- We thus can now accommodate a misfit of

$$\frac{\Delta a}{a} = \frac{b}{d} = \alpha$$

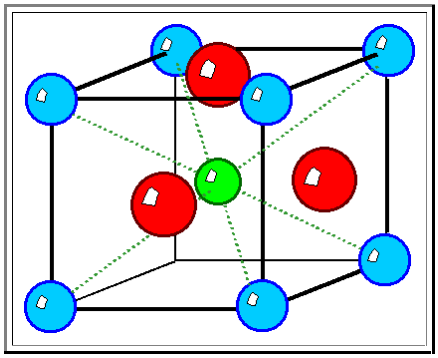
- **Wow!** An angle of **10^o**, easily within the range of small angle grain boundaries, will have a value of about **0,175** in angular radians and thus corresponds to a misfit of **17.5 %** !!!!
  - If this works, we could accommodate huge misfits with no dislocations in the phase boundary. The prize to pay is that we have a dense area of edge dislocations some **100 nm** below the phase boundary. But that may not be detrimental to the electronic or optoelectronic uses you had in mind for your phase boundary.
- The question, of course, is: *Does it work?* Especially if your magical wand is at the repair shop? The answers are:
  - **1.** Yes - it works, at least in principle. But much research and optimization needs most certainly to be done before compliant substrates can be used for products.
  - **2.** Your magic wand is supplied by the forces acting on the dislocations as soon as you start depositing the strained layer. These forces will try to rotate the dislocations from screw to edge orientation. So not having a wand is not the real problem.
  - However, there is no way to rotate a complete network as a whole. But *patches* of network, separated by a third set of dislocations accommodating steps or some small tilt component as seen in the example, can possibly rotate independent of each other.
- Finding out exactly how this can happen (and thus how to optimize it by creating an optimized boundary structure) will be one of the keys for success with this technique.
  - This shows to demonstrate that knowing a few things about dislocations may come in handy one day.
  - So try it. See if you can figure out how the screw dislocation network can rotate patch by patch by suitable dislocation interactions, involving, maybe, a bunch of additional dislocations as needed, e.g. to accommodate a small tilt component.

1) They actually did it, consult the link!

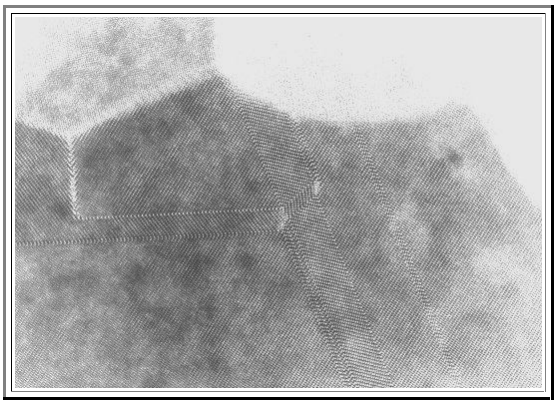
Interpreting HRTEM Images

Advanced

Here is a **HTEM** picture of grain boundaries in **$BaTiO_3$**, one of the most important electroceramic materials. Its lattice is of the Perovskite type and looks like this:



The **Ti** atom is in the center and the **O** atoms on the face centers positions. The lattice is not exactly cubic but slightly elongated. A **HRTEM** image from Prof. Urbans group (Res. Center Jülich) shows many grain boundaries:

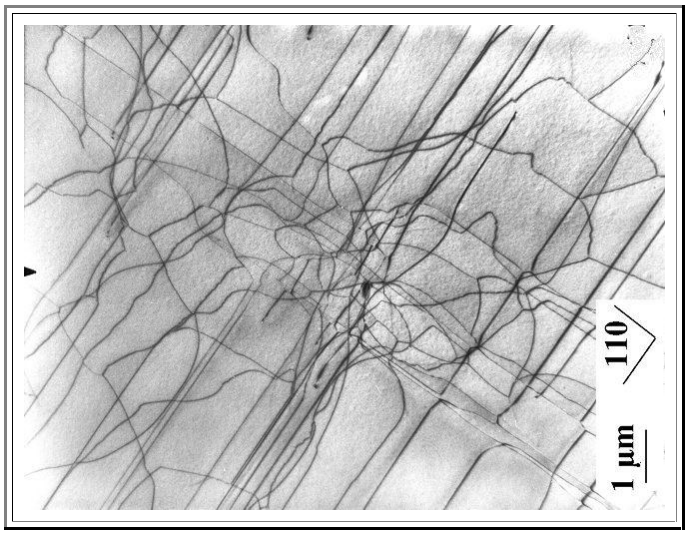


The link provides a large-size copy. Can you figure out the structure of the boundaries? Are there any dislocations present? This is not so much an exercise but a demonstration that even images with atomic resolution do not solve all problems.

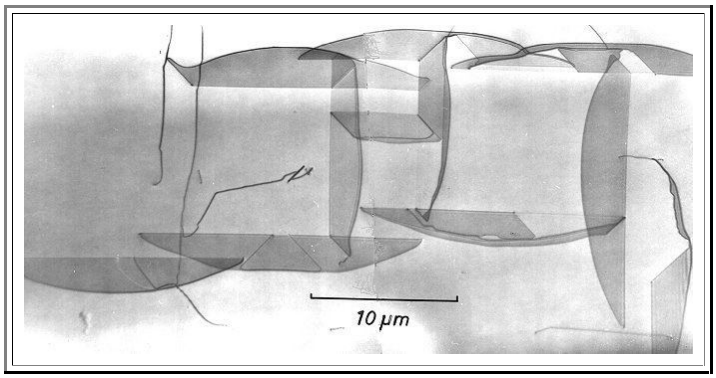
Misfit Dislocations in the Interface between Heavily and Normally Doped Silicon

Illustration

- The **TEM** micrograph shows a loose network of dislocations between "regular" and heavily **B**-doped Silicon. The expected square network has not yet fully developed. Many dislocations are "on their way" from the surface to their proper place in the interface.
  - The geometry is also not too well defined, because there is no abrupt change of lattice constants as in the case of phase boundaries between chemically different phases. The lattice constant changes continuously following the **B**-concentration which obeys some diffusion profile.



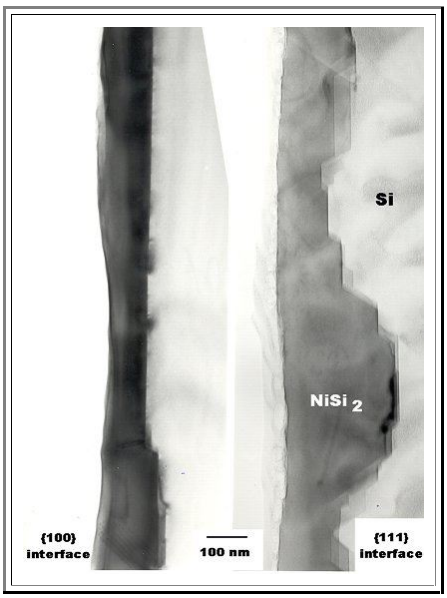
- On occasions, a stacking fault network instead of a dislocation network is observed as shown below. The reasons for this unclear. Stacking faults of this gigantic size should be totally unstable and would be expected to unfault.



Cross-Sectional TEM of Si - $NiSi_2$ Interfaces

Illustration

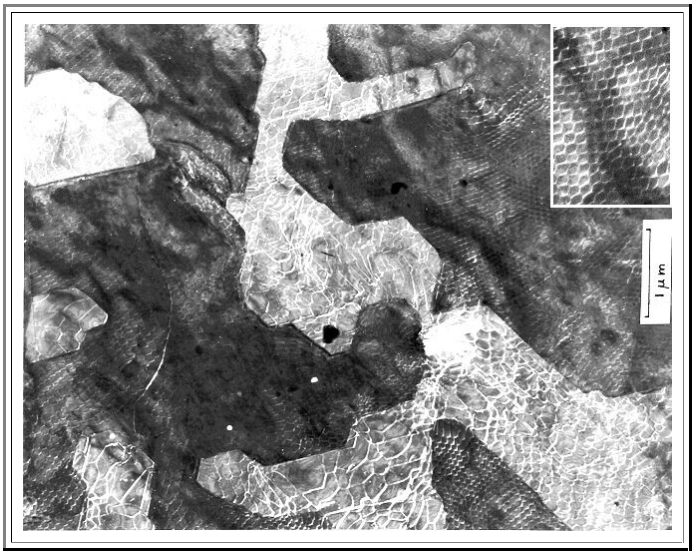
- Shown are two cross-section of the **Si - $NiSi_2$** interface (and the **$NiSi_2$** surface).
  - **Facetting** of the interface on **{111}** is very prominent, but the **$NiSi_2$** surface shows facets, too..



Dislocation Network in a Si - $NiSi_2$ Interface

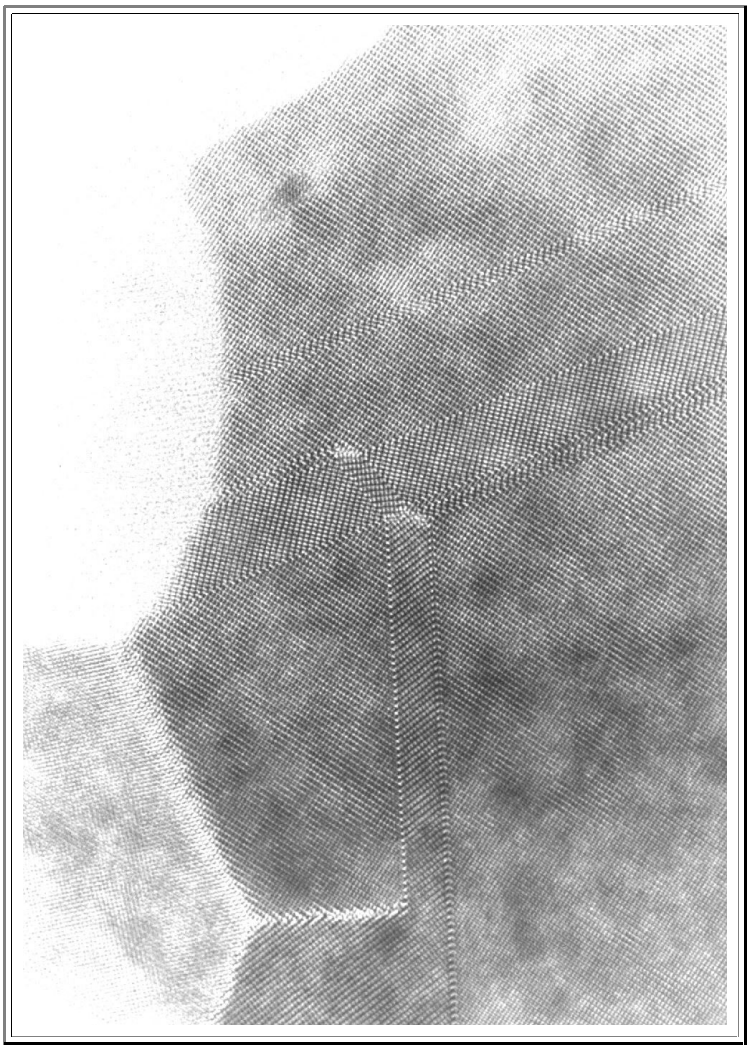
Illustration

- This overview picture shows the "*interphase*" structure of the **$NiSi_2$ - Si** interface.
  - In the bright patches we have a direct epitaxial relationship, the network consists of ***a*/2<110>** dislocations. split into partial dislocations, with extended and constricted dislocation nodes. This is exactly as we have seen it before in the small angle grain boundary in Si, except that the dislocations now are *edge* dislocations and not *screw* dislocations!
  - In the darker areas the **$NiSi_2$** layer is *twinned* with respect to the substrate. The dislocation network is composed of the ***a*/6<112>** dislocation of the **DSC** lattice belonging to a **$\Sigma = 3$** relation. The inset shows this network at higher magnification. This, again, is quite similar to the splitting of the small angle grain boundary on **Si** into a microtwin plus dislocation network.



Grain Boundaries in $BaTiO_3$ - Large Size

Here is the **HRTEM** image of grain boundaries in **$BaTiO_3$** in full glory.



Illustration