Nobody is Perfect

1. Introduction

1.1 Organization

1.1.1 Use of the Hyperscript

- There are a number of special modules that you should use for navigating through the Hyperscript:
  - Detailed table of contents of the main part (called "backbone")
  - Matrix of Modules; showing all modules in context. *This is your most important "Metafile"!!!*
  - Indexlist; with direct links to the words as they appear in the modules. All words contained in the indexlist are marked **black and bold** in the text.
  - List of names; with direct links to the words as they appear in the modules. All names contained in the name list are marked **red and bold** in the text.
  - List of abbreviations; with direct links to the symbols and abbreviations as they appear in the modules
  - Dictionary; giving the German translation of not-so-common English words; again with direct links to the words as they appear in the modules. All words found in the dictionary are marked ***italic, black, and bold***. *The German translation appears directly on the page if you move the cursor on it*
- All lists are automatically generated, so errors will occur.
  - Note: *Italics and red* emphasizes something directly, without any cross reference to some list.
  - All numbers, chemical symbols etc. are written with **bold** character. There is no particular reason for this except that it looks better to me.
  - *Variables* in formulas etc. are written in italics as it should be - except when it gets confusing. Is ***v*** a **v** as in velocity in italics, or the greek ν? You get the point.

1.1.2 What it is All About

- The lecture course "Defects in Crystals" attempts to teach all important structural aspects (as opposed to electronic aspects) of defects in crystals. It covers all types of defects (from simple ***vacancies*** to ***phase boundaries***; including more complicated ***point defects***, ***dislocations***, ***stacking faults***, ***grain boundaries***), their role for properties of materials, and the analytical tools for detecting defects and measuring their properties
  - If you are not too sure about the role of defects in materials science, turn to the preface.
  - If you want to get an idea of what you should know and what will be offered, turn to chapter 2
- A few more general remarks
  - The course is far to short to really cover the topic appropriately, but still overlaps somewhat with other courses. The reasons for this is that defects play a role almost everywhere in materials science so many courses make references to defects.
  - The course has a *special format for the exercise part* similar to "Electronic Materials", but a bit less formalized. Conventional exercises are partially abandoned in favor of "professional" presentations including a paper to topics that are within the scope of the course, but will not be covered in regular class. A list of topics is given in chapter 1.2.1
- The intention with this particular format of exercises is:
  - Learn how to research an unfamiliar subject by yourself.
  - Learn how to work in a team.
  - Learn how to make a scientific presentation in a limited time (Some hints can be found in the link)
  - Learn how to write a coherent paper on a well defined subject.

- Learn about a new (and hopefully exciting) topic concerning "defects".

Accordingly, the contents and the style of the presentation will also be discussed to some extent. The emphasize, however, somewhat deviating from "Electronic Materials", is on content. For details use the link.

1.1.3 Relation to Other Courses

The graduate course "Defects in Crystals" interacts with and draws on several other courses in the materials science curriculum. A certain amount of overlap is unavoidable. Other courses of interest are

Introduction to Materials Science I + II ("MaWi I + II"; Prof. Föll)

- Required for all "Dipl.-Ing." students; **3rd** and **4th** semester
- Undergraduate course, where the essentials of crystals, defects in crystals, band structures, semiconductors, and properties of semiconductors up to semi-quantitave ***I-V-***characteristics of p-n-junctions are taught.
- For details of contents refer to the Hyperscripts (in german)
  MaWi I
  MaWi II

Physical Metallurgy I ("Metals I", Prof. Faupel)

- Includes properties of dislocations and hardening mechanisms

Sensors I

- Will, among other topics, treat point defects equilibria and reactions in the context of sensor applications

Materials Analytics I + II ("Analytics I + II", Prof. Jäger)

- Covers in detail some (but not all) of the experimental techniques, e.g. Electron Microscopy

Solid State Physics I + II ("Solid State I + II" Prof. Faupel)

- Covers the essentials of solid state physics, but does not cover structural aspects of defects.

Semiconductors (Prof. Föll)

- Covers "everything" about semiconductors except **Si** technology (but other uses of **Si**, some semiconductor physics, and especially optoelectronics). Optpelectronics needs heterojunctions and heterojunctions are plagued by defects.

1.1.4 Books

Consult the **list of books**

1.1.2 Required Background Knowledge

- **Mathematics**
  - Not much. Familiarity with with basic undergraduate math will suffice.
- **General Physics and Chemistry**
  - Familiarity with thermodynamics (including statistical thermodynamics), basic solid state physics, and general chemistry is sufficient.
- **Materials Science**
  - You should know about basic crystallography and thermodynamics. The idea is that you emerge from this course *really understanding* structural aspects of defects in some detail. Since experience teaches that abstract subjects are only understood after the second hearing, you should have heard a little bit about point defects, dislocations, stacking faults, etc. before.

1.1.3 Organization

- Everything of interest can be found in the "**Running Term**" files
  - Index to running term

1.2 Exercises and Seminar

1.2.1 General Topics

- *This module contains brief general information about exercises and the seminar.*
  - Whatever is really happening in the running term, will be found in the links
    - Running Term
    - Seminar topics
- As far as *exercise classes* take place, the questions will be either from the Hyperscript or will be constructed along similar lines.
- As far as the *seminar* part is concerned: Which group will deal with which topic will be decided in the first week of the class.
  - You may choose your subject from the list of topics, or *suggest* a subject of your interest which is not on the list.
  - Presentation will be clustered at the second half of the term (or, if so demanded, in the semester break); the beginning date depends on the number of participants
- For most topics, you can sign out some materials to get you started; there is also always helpavailable from the teaching assistants.

1.2.2 Rules for Seminar

General Rules

- Teams:
  - Two (or, as an exception), three students form a team.
  - The team decides on the the detailed outline of the presentation, collects the material and writes the paper.
  - The delivery of the presentation can be done in any way that divides the time about equally between the members of the team.
  - Every team has an advisor who is always available (but do call ahead).
- *Selection of Topics and Schedules*
  - The relevant list of topics available for the current term will be presented and discussed at the first few weeks of the lecture class.
  - You may suggest your own topic.
- *Preparation of the Presentation*
  - Starting material will be issued in the second week of the course, but it is the teams responsibility to find the relevant literature.
  - The teams should consult their advisor several weeks prior to the presentation and discuss the outline and the contents of the presentation.

Presentation and Paper

- *Language*
  - The presentation and the paper should be given in *English language*. Exceptions are possible upon demand; but ***vuegraphs*** *must be* in English without exception. Language and writing skills will *not* influence the grading.
  - Papers must be handed in at the latest one day before the presentation in an *electronic format* (preferably html), and as a copy-ready paper. Very good papers written in **HTML** will be included in the hyperscript.
  - Copies for the other students will be made and issued by the lecture staff
  - Papers that are handed in *at least* one week before the presentation will be corrected with respect to language (this might improve the copies you hand out!)

Presentation

- The presentations *must not* exceed **45 min.** (For exceptions, ask your advisor).
- Presentations will be filmed if so desired (tell your advisor well ahead of time). The video is only available to the speakers.
- The presentation is followed by a discussion (**10 - 15 min.**) The discussion leader (usually the advisor) may ask questions to the speaker and the audience.

1.3 Defects, Materials and Products

1.3.1 General Classification of Defects

- Crystal lattice defects (defects in short) are usually classified according to their dimensions. Defects as dealt with in this course may then be classified as follows:
- **0-dimensional defects**
  - We have "**point defects**" (on occasion abbreviated ***PD***), or, to use a better but unpopular name, "**atomic size defects**" .
  - Most prominent are **vacancies** (***V***) and **interstitials** (***i***). If we mean **self-interstitials** (and you should be careful with using the name interstitials indiscriminately), these two point defects (and if you like, small agglomerates of these defects) are the only possible **intrinsic** point defects in element crystals.
  - If we invoke *extrinsic* atoms, i.e. **impurity atoms** on lattice sites or interstitial sites, we have a second class of point defects subdivided into interstitial or substitutional impurity atoms or **extrinsic point defects**.
  - In slightly more complicated crystals we also may have mixed-up atoms (e.g. a **Ga** atom on an **As** site in a **GaAs** crystal) or **antisite defects**
- **1-dimensional Defects**
  - This includes all kinds of **dislocations**; for example:
  - Perfect ***dislocations***, partial dislocations (always in connection with a ***stacking fault***), dislocation loops, grain boundary and phase boundary dislocations, and even
  - Dislocations in **quasicrystals**.
- **2-dimensional Defects**
  - Here we have **stacking faults** (***SF***) and **grain boundaries** in crystals of *one* material or phase, and
  - **Phase boundaries** and a few special defects as e.g. boundaries between ordered domains.
- **3-dimensional Defects**
  - This includes: **Precipitates**, usually involving impurity atoms.
  - **Voids** (little holes, i.e. agglomerates of vacancies in three-dimensional form) which may or may not be filled with a gas, and
  - Special defects, e.g. stacking fault tetrahedra and tight clusters of dislocations.

- If you understand German, you will find an elementary introduction to all these topics in chapter 4 of the "Materialwissenschaft I" Hyperscript

1.3.2 Materials Properties and Defects

Material Properties and Defects

- Defects determine many properties of materials (those properties that we call "**structure sensitive properties**"). Even properties like the specific resistance of semiconductors, conductance in ionic crystals or diffusion properties in general which may appear as intrinsic properties of a material are defect dominated - in case of doubt by the intrinsic defects. Few properties - e.g. the melting point or the elastic modulus - are not, or only weakly influenced by defects.
- To give some flavor of the impact of defects on properties, a few totally subjective, if not speculative points will follow:
  - Generally known are: ***Residual resistivity***, conductivity in semiconductors, diffusion of impurity atoms, most mechanical properties around plastic deformation, optical and optoelectronic properties, *but we also have* :
  - Crystal growth, recrystallization, phase changes.
  - Corrosion - a particularly badly understood part of defect science.
  - Reliability of products, lifetimes of minority carriers in semiconductors, and lifetime of products (e.g. chips). Think of electromigration, cracks in steel, ***hydrogen embrittlement***.
  - Properties of quantum systems (superconductors, quantum Hall effect)
  - Evolution of life (defects in **DNA** "crystals")
- A large part of the worlds technology depends on the manipulation of defects: All of the "metal bending industry"; including car manufacture, but also all of the semiconductor industry and many others.

Properties of Defects

- Defects have many properties in themselves. We may ask for:
  - *Structural properties:* Where are the atoms relative to the perfect reference crystal?
  - *Electronic properties:* Where are the defect states in a band structure?
  - *Chemical properties:* What is the chemical potential of a defect? How does it participate in chemical reactions, e.g. in corrosion?
  - *Scattering properties:* How does a defect interact with particles (phonons, photons of any energy, electrons, positrons, ...); what is the scattering cross section?
  - *Thermodynamic properties:* The question for formation enthalpies and -entropies, interaction energies, migration energies and entropies, ...
- Despite intensive research, many questions are still open. There is a certain irony in the fact that point defects are least understood in the material where they matter most: In Silicon!

Goals of the course

- This course emphasizes structural and thermodynamic properties. You should acquire:
  - A good understanding of defects and defect reactions.
  - A rough overview of important experimental tools.
  - Some appreciation of the elegance of mother nature to make much (you, crystals, and everything else) out of little (**92** elements and a bunch of photons).

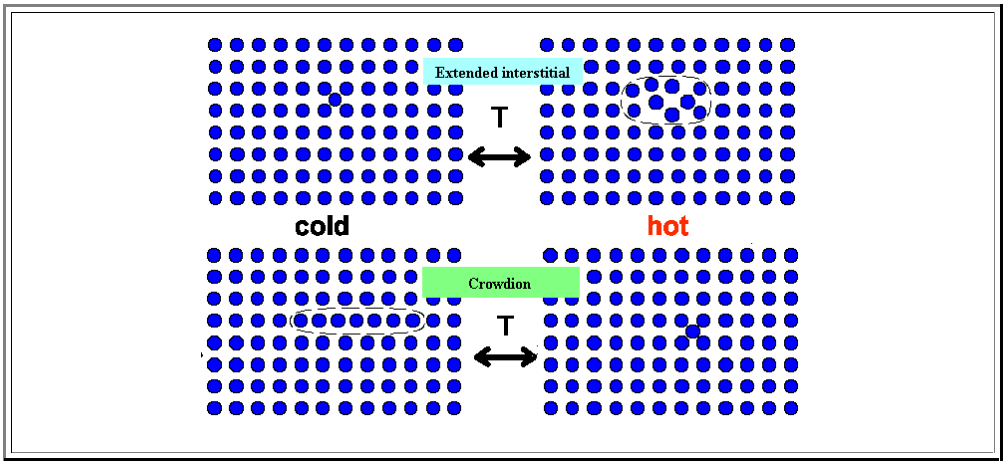
1.3.3 The larger View and Complications

Looking More Closely at Point Defects

- This subchapter means to show that even the seemingly most simple defects - vacancies and interstitials - can get pretty complex in real crystals. This is already true for the most simple real crystal, the **fcc** lattice with one atom as a base, and very true for fcc lattices with two identical atoms as a base, i.e. **Si** or diamond. In *really* complicated crystals we have at least as many types of vacancies and interstitials as there are different atoms - it's easy to lose perspective.
  - To give just two examples of real life with point defects: In the seventies and eighties a bitter war was fought concerning the precise nature of the self-interstitial in elemental **fcc** crystals. The main opponents where two large German research institutes - the dispute was never really settled.
  - Since about **1975** we have a world-wide dispute still going on concerning the nature of the intrinsic point defects in **Si** (and pretty much all other important semiconductors). We learn from this that even point defects are not easy to understand.
- You may consider this sub-chapter as an overture to the point defect part of course: Some themes touched upon here will be be taken up in full splendor there. Now lets look at some phenomena related to point defects
- We start with a simple vacancy or interstitial in (**fcc**) crystals which exists in thermal equilibrium and ask a few questions (which are mostly easily extended to other types of crystals):

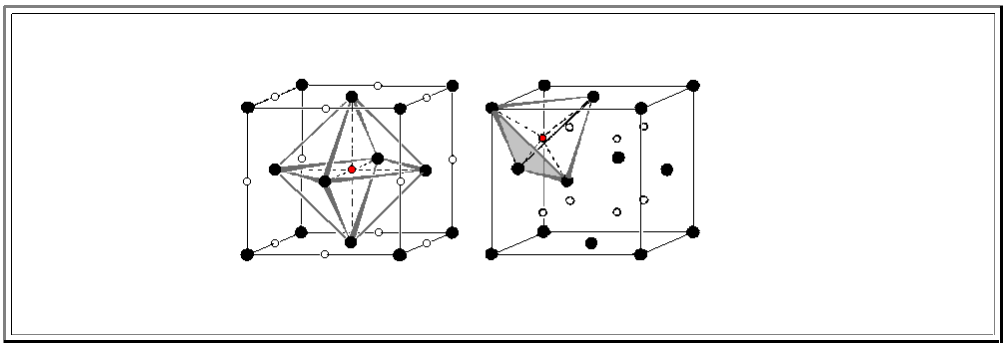
The atomic structure

- What is the *atomic structure* of point defects? This seems to be an easy question for vacancies - just remove an atom!
  - But how "big", how extended is the vacancy? After all, the neighboring atoms may be involved too. Nothing requires you to have only simple thoughts - lets think in a complicated way and make a vacancy by removing **11** atoms and filling the void with **10** atoms - somehow. You have a vacancy. What is the structure now?
  - How about interstitials? Lets not be unsophisticated either. Here we could fill our **11**-atom-hole with **12** atoms. We now have some kind of "**extended**" interstitial? *Does this happen*? (Who knows, its possibly true in **Si**). How can we discriminate between "localized" and "extended" point defects?
  - With interstitials you have several possibilities to put them in a lattice. You may choose the dumbbell configuration, i.e. you put two atoms in the space of one with some symmetry conserved, or you may put it in the octahedra or tetrahedra interstitial position. Perhaps surprisingly, there is still one more possibility:
  - The ***"crowdion"***, which is supposed to exist as a metastable form of interstitials at low temperatures and which was the subject of the "war" mentioned above.
  - Then we have the **extended interstitial** made following the general recipe given above, and which is believed by some (including me) to exist at high temperatures in **Si**. Lets see what this looks like:



- Next, we may have to consider *the charge state* of the point defects (important in semiconductors and ionic crystals).
  - Point defects in ionic crystals, in general, must be charged for reasons of charge neutrality. You cannot, e.g. form **Na**-vacancies by removing $\mathbf{Na^+}$ ions without either giving the resulting vacancy a positive charge or depositing some positive charges somewhere else.
  - In semiconductors the charge state is coupled to the energy levels introduced by a point defects, its position in the bandgap and the prevalent Fermi energy. If the Fermi energy changes, so does, perhaps, the charge state.
- Now we might have a *coupling between charge state and structure*. And this may lead to an athermal diffusion mechanisms; something really strange (after **Bourgoin**).

- Just an arbitrary example to illustrate this: The *neutral* interstitial sits in the octahedra site, the *positively charged* one in the tetrahedra site (see below). Whenever the charge states changes (e.g. because its energy level is close to the Fermi energy or because you irradiate the specimen with electrons), it will jump to one of the nearest equivalent positions - in other word it diffuses *independently of the temperature*.



These examples should convince you that even the most simplest of defects - point defects - are not so simple after all. And, so far, we have (implicitly) only considered the simple case of *thermal equilibrium*! This leads us to the next paragraph:

But is there thermal equilibrium?

The list above gives an idea what *could* happen. But what, actually, *does* happen in an ideal crystal in *thermal equilibrium*?

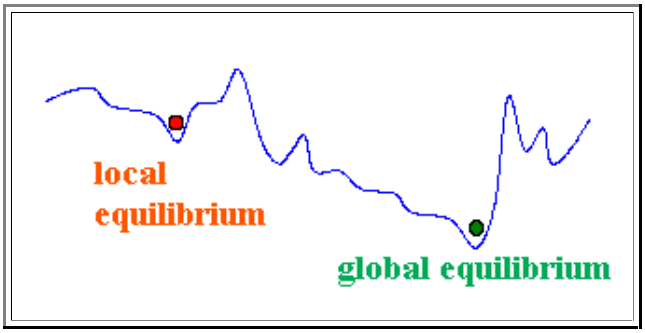
- While we believe that for common fcc metal this question can be answered, it is still open for many important materials, including Silicon. You may even ask: *Is there thermal equilibrium at all*?
- Consider: Right after a new portion of a growing crystal crystallized from the melt, the concentration of point defects may have been controlled by the growth kinetics and not by equilibrium. If the system now tries to reach equilibrium, it needs sources and sinks for point defects to generate or dump what is required. Extremely perfect **Si** crystals, however, do not have the common sources and sinks, i.e. dislocations and grain boundaries. So what happens? Not totally clear yet. There are more open questions concerning Si; activate the link for a sample.

Well, while there may be some doubt as to the existence of thermal equilibrium now and then, there is no doubt that there are many occasions where we definitely do not have thermal equilibrium. What does that mean with respect to point defects?

Non-equilibrium

Global equilibrium, defined by the *absolute* minimum of the free enthalpy of the system is often unattainable; the second best solution, *local* equilibrium where some local minimum of the free enthalpy must suffice. You always get non-equilibrium, or just a local equilibrium, if, starting from some equilibrium, you change the temperature.

- Reaching a new local equilibrium of any kind needs kinetic processes where point defects must move, are generated, or annihilated. A typical picture illustrating this shows a potential curve with various minima and maxima. A state caught in a local minima can only change to a better minima by overcoming an energy barrier. If the temperature ***T*** does not supply sufficient thermal energy **k*T***, global equilibrium (the deepest minimum) will be reached slowly or - for all practical purposes - never.



- One reaction helpful for reaching a minima in cases where both vacancies and interstitials exist in non-equilibrium concentrations (e.g. after lowering the temperature or during irradiation experiments) could be the mutual annihilation of vacancies and interstitials by recombination. The potential barrier that must be overcome seems to be only the migration enthalpy (at least one species must be mobile so that the defects can meet).

- There might be unexpected new effects, however, with extended defects. If an localized interstitial meets an extended vacancy, how is it supposed to recombine? There is no local empty space, just a thinned out part of the lattice. Recombination is not easy then. The barrier to recombination, however, in a kinetic description, is now an **entropy barrier** and not the common energy barrier.

- Things get really messy if the generation if point defects, too, is a non-equilibrium process - if you produce them by crude force. There are many ways to do this:
  - **Crystal Growth** As mentioned above, the incorporation of point defects in a growing interface does not have to produce the equilibrium concentration of point defects. An "easy to read" paper to this subject (in German) is available in the Link
  - **Quenching**, i.e. ***rapid cooling***. The point defects become immobile very quickly - a lot of sinks are needed if they are to disappear under these conditions - a rather unrealistic situation.
  - **Plastic deformation**, especially by dislocation climb, is a non-equilibrium source (or sink) for point defects. It was (and to some extent still is) the main reason for the degradation of Laser diodes.
  - **Irradiation** with electrons (mainly for scientific reasons), ions (as in ion implantation; a key process for microelectronics), neutrons (in any reactor, but also used for **neutron transmutation doping** of **Si**), α-particles (in reactors, but also in satellites) produces copious quantities of point defects under "perfect" non-equilibrium conditions.
  - **Oxidation** of **Si** injects **Si** interstitials into the crystal.
  - **Nitridation** of **Si** injects vacancies into the crystal.
  - **Reactive Interfaces** (as in the two examples above), quite generally, may inject point defects into the participating crystals.
  - **Precipitation** phenomena (always requiring a moving interface) thus may produce point defects as is indeed the case: (**SiO_2-*precipitation*** generates, **SiC**-precipitation uses up Si-interstitials.
  - **Diffusion** of impurity atoms may produce or consume point defects beyond needing them as diffusion vehicles.

- And all of this may critically influence your product. The **Si** crystal growth industry, grossing some **8 billion $** a year, continuously runs into severe problems caused by point defects that are not in equilibrium.
  - So-called swirl-defects, sub-distinguished into **A**-defects and **B**-defects caused quite some excitement around **1980** and led the way to the acceptance of the existence of interstitials in **Si**.
  - Presently, **D**-defects are the hot topics, and it is pretty safe to predict that we will hear of **E**-defects yet.

- Now, most of the examples of possible complications mentioned here are from pretty recent research and will not be covered in detail in what follows.
  - And implicitely, we only discussed defects in monoatomic crystals - metals, simple semiconductors. In more complicated crystals with two or more different atoms in the base, things can get really messy - look at chapters 2.4 to get an idea.
  - Anyway, you should have the feeling now that acquiring some knowledge about defects is not wasted time. Materials Scientists and Engineers will have to understand, use, and battle defects for many more years to come. Not only will they not go away - they are needed for many products and one of the major "buttons" to fiddle with when designing new materials

2. Properties of Point Defects

2.1 Intrinsic Point Defects and Equilibrium

2.1.1 Simple Vacancies and Interstitials

Basic Equilibrium Considerations

- We start with the most simple point defects imaginable and consider an *uncharged* ***vacancy*** in a simple crystal with a base consisting of only one atomic species - that means mostly metals and semiconductors.
  - Some call this kind of defect "Schottky Defect, although the original **Schottky** defects were introduced for *ionic* crystals containing at least *two* different atoms in the base.
  - We call vacancies and their "opposites", the self-intersitals, **intrinsic point defects** for starters. *Intrinsic* simple means that these point defects can be generated in the ideal world of the ideal crystal. No external or *extrinsic* help or stuff is needed.
- To form *one* vacancy at constant pressure (the usual situation), we have to add some free enthalpy G_F to the crystal, or, to use the name commonly employed by the chemical community, **Gibbs energy**.
  - G_F, the free enthalpy of vacancy formation, is defined as

$$G_F = H_F - T \cdot S_F$$

  - The index F always means "**formation**"; H_F thus is the **formation enthalpy** of *one* vacancy, S_F the **formation entropy** of *one* vacancy, and T is always the absolute temperature.
- The formation *enthalpy* H_F in solids is practically indistinguishable from the formation *energy* E_F (sometimes written U_F) which has to be used if the volume and not the pressure is kept constant.
  - The *formation entropy*, which in elementary considerations of point defects usually is omitted, must not be confused with the *entropy of mixing* or configurational entropy; the entropy originating from the many possibilities of arranging *many* vacancies, but is a property of a *single* vacancy resulting from the disorder introduced into the crystal by changing the vibrational properties of the neighboring atoms (see ahead).
- The next step consists of minimizing the free enthalpy G of the complete crystal with respect to the number n_V of the vacancies, or the concentration $c_V = n_V / N$, if the number of vacancies is referred to the number of atoms N comprising the crystal. We will drop the index "**V**" from now now on because this consideration is valid for all kinds of point defects, not just vacancies.
  - The number or concentration of vacancies in **thermal equilibrium** (which is not necessarily identical to **chemical equilibrium!**) then follows from finding the minimum of G with respect to n (or c), i.e.

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

  - with G_0 = Gibbs energy of the perfect crystal, G_1 = Work (or energy) needed to generate n vacancies = $n \cdot G_F$, and $G_2 = -T \cdot S_{conf}$ with S_{conf} = **configurational entropy** of n vacancies, or, to use another expression for the same quantity, the **entropy of mixing** n vacancies.
- We note that the partial derivative of G with respect to n, which should be written as $[\partial G / \partial n]_{everything\ else\ =\ const.}$ is, *by definition*, the **chemical potential** μ of the defects under consideration. This will become important if we consider chemical equilibrium of defects in, e.g., ionic crystals.
- The partial derivatives are easily done, we obtain

$$\frac{\partial G_0}{\partial n} = 0$$

$$\frac{\partial G_1}{\partial n} = G_F$$

  - which finally leads to

$$\frac{\partial G}{\partial n} = G_F - T \cdot \frac{\partial S_{conf}}{\partial n} = 0$$

$$= \text{chemical potential } \mu_V \text{ in equilibrium}$$

We now need to calculate the *entropy of mixing* or configurational entropy S_{conf} by using **Boltzmann's** famous formula

$$S = k_B \cdot \ln P$$

- With k_B = **k** = Boltzmanns constant and P = number of different configurations (= microstates) for the same **macrostate**.
- The exact meaning of P is sometimes a bit confusing; activate the link to see why.

A *macrostate* for our case is any possible combination of the number n of vacancies and the number N of atoms of the crystal. We obtain $P(n)$ thus by looking at the number of possibilities to arrange n vacancies on N sites.

- This is a standard situation in combinatorics; the number we need is given by the **binomial coefficient**; we have

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

- If you have problems with that, look at exercise 2.1-1 below.

The calculation of $\partial S / \partial n$ now is straight forward in principle, but analytically only possible with two approximations:

- **1.** *Mathematical Approximation:* Use the **Stirling formula** in its simplest version for the ***factorials***, i.e.

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

- **2.** *Physical Approximation:* There are always far fewer vacancies than atoms; this means

$$N - n \approx N$$

As a first result we obtain "approximately"

$$T \cdot \frac{\partial S}{\partial n} \approx kT \cdot \ln \frac{N}{n}$$

If you have any doubts about this point, you should do the following exercise.

Exercise 2.1-1

Derive the Formula for c_V

With $n/N = c_V$ = concentration of vacancies as defined before, we obtain the familiar formula

$$c_V = \exp\left(-\frac{G_F}{kT}\right)$$

- or, using $G_F = H_F - T\,S_F$

$$c_V = \exp\left(\frac{S_F}{k}\right) \cdot \exp\left(-\frac{H_F}{kT}\right)$$

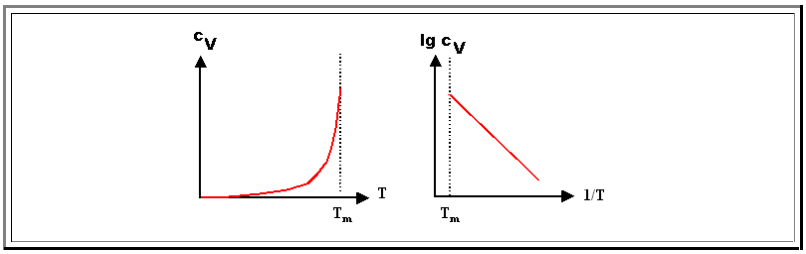
For **self-interstitials**, exactly the same formula applies if we take the formation energy to be now the formation energy of a self-interstitial.

- However, the formation enthalpy of self-interstitials is usually (but not necessarily) considerably larger than that of a vacancy. This means that their equilibrium concentration is usually substantially smaller than that of vacancies and is mostly simply neglected.
- Some numbers are given in this link; far more details are found here. The one number to remember is:

$$H_F(\text{vacancy}) \text{ in simple metals} \approx 1\ \text{eV}$$

It goes without saying (I hope) that the way you look at equations like this is via an **Arrhenius plot**. In the link you can play with that and refresh your memory

- Instead of plotting $c_V(T)$ vs. T directly as in the left part of the illustration below, you plot the logarithm $\lg[c_V(T)]$ vs. $1/T$ as shown on the right.
- In the resulting "Arrhenius plot" or "Arrhenius diagram" you will get a straight line. The (negative) slope of this straight line is then "**activation" energy** of the process you are looking at (in our case the formation energy of the vacancy), the y-axis intercept gives directly the pre-exponential factor.



Compared to simple formulas in elementary courses, the factor $\exp(S_F/k)$ might be new. It will be justified below.

Obtaining this formula by shuffling all the factorials and so on is is not quite as easy as it looks - lets do a little fun exercise

Exercise 2.1-2

Find the mistake!

Like always, one can second-guess the assumptions and approximations: Are they really justified? When do they break down?

- The reference enthalpy G_0 of the perfect crystal may not be constant, but dependent on the chemical environment of the crystal since it is in fact a sum over chemical potentials including all constituents that may undergo reactions (including defects) of the system under consideration. The concentration of oxygen vacancies in oxide crystals may, e.g., depend on the partial pressure of O_2 in the atmosphere the crystal experiences. This is one of the working principles of **Ionics** as used for **sensors**. Chapter 2.4 has more to say to that.
- The simple equilibrium consideration does not concern itself with the kinetics of the generation and annihilation of vacancies and thus makes no statement about the *time required to reach equilibrium*. We also must keep in mind that the addition of the surplus atoms to external or internal surfaces, dislocations, or other defects while generating vacancies, may introduce additional energy terms.

- There may be more than one possibility for a vacancy to occupy a lattice site (for interstitials this is more obvious). This can be seen as a degeneracy of the energy state, or as additional degrees of freedom for the combinatorics needed to calculate the entropy. In general, an additional entropy term has to be introduced. Most generally we obtain

$$c = \frac{Z_d}{Z_0} \cdot \exp - \frac{G_F}{kT}$$

- with Z_d or Z_0 = ***partition functions*** of the system with and without defects, respectively. The link (in German) gets you to a short review of statistical thermodynamics including the partition function.

- Lets look at two examples where this may be important:
  - The energy state of a vacancy might be "degenerate", because it is charged and has trapped an electron that has a spin which could be either up or down - we have two, energetically identical "versions" of the vacancy and $Z_d/Z_0 = 2$ in this case.
  - A double vacancy in a **bcc** crystals has more than one way of sitting at one lattice position. There is a preferred orientation along **<111>**, and $Z_d/Z_0 = 4$ in this case.

Calculation and Physical Meaning of the Formation Entropy

- The formation entropy is associated with a single defect, *it must not be mixed up with the entropy of mixing* resulting from many defects.
  - It can be seen as the *additional* entropy or disorder added to the crystal with every additional vacancy. There is disorder associated with every single vacancy because the *vibration modes* of the atoms are disturbed by defects.
  - Atoms with a vacancy as a neighbour tend to vibrate with lower frequencies because some bonds, acting as "springs", are missing. These atoms are therefore less well localized than the others and thus more "unorderly" than regular atoms.
- Entropy residing in lattice vibrations is nothing new, but quite important outside of defect considerations, too:
  - Several **bcc** element crystals are stable *only* because of the entropy inherent in their lattice vibrations. The $-TS$ term in the free enthalpy then tends to overcompensate the higher enthalpy associated with non close-packed lattice structures. At high temperatures we therefore find a tendency for a phase change converting **fcc** lattices to **bcc** lattices which have "softer springs", lower vibration frequencies and higher entropies. For details compare Chapter 6 of **Haasens** book.
- The *calculation* of the formation entropy, however, is a bit complicated. But the *result* of this calculation is quite simple. Here we give only the essential steps and approximations.
  - First we describe the crystal as a sum of **harmonic oscillators** - i.e. we use the well-known harmonic approximation. From quantum mechanics we know the energy E of an harmonic oscillator; for an oscillator number **i** and the necessary quantum number **n** we have

$$E_{i,n} = \frac{h\,\omega_i}{2\pi} \cdot (n + 1/2)$$

- We are going to derive the entropy from the all-encompassing **partition function** of the system and thus have to find the correct expression.
  - The partition function Z_i of *one* harmonic oscillator as defined in statistical mechanics is given by

$$Z_i = \sum_n \exp - \frac{h\,\omega_i \cdot (n + ½)}{2\pi \cdot kT}$$

  - The partition function of the *crystal* then is given by the *product* of all individual partition function of the $p = 3N$ oscillators forming a crystal with N atoms, each of which has three degrees of freedom for oscillations. We have

$$Z = \prod_{i=1}^{p} Z_i$$

From statistical thermodynamics we know that the free energy F (or, for solids, in a very good approximation also the free enthalpy G) of our oscillator ensemble which we take for the crystal is given by

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

Likewise, the entropy of the ensemble (for const. volume) is

$$S = -\frac{\partial F}{\partial T}$$

Differentiating with respect to T yields for the entropy of our - so far - ideal crystal without defects:

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

Now we consider a crystal with just *one* vacancy. All ***eigenfrequencies*** of all oscillators change from ω_i to a new as yet undefined value ω'_i. The entropy of vibration now is S'.

- The formation entropy S_F of our single vacancy now can be defined, it is

$$S_F = S' - S$$

- i.e. the difference in entropy between the perfect crystal and a crystal with *one* vacancy.

It is now time to get more precise about the ω_i, the frequencies of vibrations. Fortunately, we know some good approximaitons:

- At temperatures higher then the **Debye temperature**, which is the interesting temperature region if one wants to consider vacancies in reasonable concentrations, we have

$$\frac{h\omega_i}{2\pi} << kT$$

$$\frac{h\omega'_i}{2\pi} << kT$$

- which means that we can expand $h\omega_i/2\pi$ into a series of which we (as usual) consider only the first term.

Running through the arithmetic, we obtain as final result, summing over all eigenfrequencies of the crystal

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

- This now calls for a little exercise:

Exercise 2.1-3

Do the Math for the formula for the formation entropy

- For analytical calculations we only consider next neighbors of a vacancy as contributors to the sum; i.e. we assume $\omega = \omega'$ everywhere else. In a linear approximation, we consider bonds as linear springs; missing bonds change the frequency in an easily calculated way. As a result we obtain (for all cases where our approximations are sound):
  - S_F **(single vacancy)** ≈ **0.5 k (Cu) to 1.3 k (Au)**.
  - S_F **(double vacancy)** ≈ **1.8 k (Cu) to 2.2 k (Au)**.
- These values, obtained by assuming that only nearest neighbors of a vacancy contribute to the formation entropy, are quite close to the measured ones. (How formation entropies are measured, will be covered in **chapter 4**). *Reversing the argumentation*, we come to a major conclusion:
- *The formation entropy measures the spatial extension of a vacancy, or, more generally, of a zero-dimensional defect.* The larger S_F, the more extended the defect will be because than more atoms must have changed their vibrations frequencies.
  - As a rule of thumb (that we justify with a little exercise below) we have:
  - S_F ≈ **1k** corresponds to a truly atomic defect, S_F ≈ **10k** correponds to *extended defects* disturbing a volume of about **5 - 10** atoms.
  - This is more easily visualized for interstitials than for vacancies. An "atomic" interstitials can be "constructed" by taking out *one* atom and filling in *two* atoms without changing all the other atoms appreciably. An interstitial extended over the volume of e.g. **10** atoms is formed by taking out ***10*** atoms and filling in ***11*** atoms without giving preference in any way to one of the **11** atoms - you cannot identify a given atom with the interstitial.
- Vacancies or interstitials in elemental crystal mostly have formation entropies around **1k**, i.e. they are "point like". There is a big exception, however: ***Si*** *does not fit this picture*.
  - While the precise values of formation enthalpies and entropies of vacancies and interstitials in **Si** are still not known with any precision, the formation entropies are definitely large and probably temperature dependent; values around **6k - 15k** at high temperatures are considered. Historically, this led **Seeger** and **Chik** in **1968** to propose that in **Si** the self-interstitial is the dominating point defect and not the vacancy as in all other (known) elemental crystals. This proposal kicked of a major scientific storm; the dust has not yet settled.

Exercise 2.1-4

Calculate formation entropies

Multi Vacancies (and Multi - Interstitials by Analogy)

- So far, we assumed that there is no interaction between point defects, or that their density is so low that they "never" meet. But interactions are the rule, for vacancies they are usually attractive. This is relatively easy to see from basic considerations.
- Let's first look at **metals**:
  - A vacancy introduces a disturbance in the otherwise perfectly periodic potential which will be screened by the free electrons, i.e. by a rearrangement of the electron density around a vacancy. The formation enthalpy of a vacancy is mostly the energy needed for this rearrangement; the elastic energy contained in the somewhat changed atom positions is comparatively small.
  - If you now introduce a second vacancy next to to the first one, part of the screening is already in place; the free enthalpy needed to remove the second atom is smaller.
  - In other word: There is a certain **binding enthalpy** (but from now on we will call it energy, like everybody else) between vacancies in metals (order of magnitude: **(0,1 - 0,2) eV**).

Covalently bonded crystals

- The formation energy of a vacancy is mostly determined by the energy needed to "break" the bonds. Taking away a second atom means that fewer bonds need to be broken - again there is a positive binding energy.

Ionic crystals

- Vacancies are charged, this leads to Coulomb attraction between vacancies in the cation or anion sublattice, resp., and to repulsion between vacancies of the same nature. We may have positive and negative binding energies, and in contrast to the other cases *the interaction can be long-range*.

The decisive new parameter is the **binding energy E_{2V}** between two vacancies. It can be defined as above, but we also can write down a kind of "chemical" **reaction equation** involving the binding energy E_{2V} (the sign is positive for attraction):

$$1V + 1V \Leftrightarrow V_2 + E_{2V}$$

- **V** in this case is *more* than an abbreviation, it is the "*chemical symbol*" for a vacancy.
- If you have some doubts about writing down chemical reaction equation for "things" that are not atoms, you are quite right - this needs some special considerations. But rest assured, the above equation is correct, and you can work with it exactly as with any reaction equation, i.e. apply reaction kinetics, the mass action law, etc.

Now we can do a calculation of the equilibrium concentration of **Divacancies**. We will do this in two ways.

First Approach: Minimize the total free enthalpy (as before):

- First we define a few convenient quantities

$$G_{F(2V)} = H_{F(2V)} - TS_{F(2V)}$$

$$H_{F(2V)} = 2H_{F(1V)} - E_{2V}$$

$$S_{F(2V)} = 2S_{F(1V)} + \Delta S_{2V}$$

- With ΔS_{2V} = **entropy of association** (it is in the order of **1k - 2k** in metals), and E_{2V} = binding energy between two vacancies.
  We obtain in complete analogy to single vacancies

$$c_{2V} = \frac{z}{2} \cdot \exp\frac{S_{2V}}{k} \cdot \exp -\frac{H_{F(2V)}}{kT}$$

$$c_{2V} = c_{1V}^2 \cdot \frac{z}{2} \cdot \exp\frac{\Delta S_{2V}}{k} \cdot \exp\frac{E_{2V}}{kT}$$

- The factor **z/2** (**z =** coordination number = number of (symmetrically identical) next neighbors) takes into account the different ways of aligning a divacancy on one point in the lattice as already noticed above. We have **z = 12** for **fcc**, **8** for **bcc** and **4** for diamond lattices.

The formula tells us that the concentration of divacancies in *thermal equilibrium* is always *much* smaller than the concentration of single vacancies since $c_V << 1$. "Thermal equilibrium" has been emphasized, because in *non-equilibrium things are totally different*!

- Some typical values for metals close to their melting point are

$$c_{1V} = 10^{-4} - 10^{-3}$$

$$c_{2V} = 10^{-6} - 10^{-5}$$

In the *second approach*, we use the **mass action law**.

- With the reversible reaction **1V + 1V** $\Leftrightarrow$ $V_{2V} + E_{2V}$ and by using the mass action law we obtain

$$\frac{(c_{1V})^2}{c_{2V}} = K(T) = \text{const} \cdot \exp -\frac{\Delta E}{kT}$$

- With ΔE = energy of the *forward reaction* (you have to be extremely careful with sign conventions whenever invoking mass action laws!). This leads to

$$c_{2V} = (c_{1V})^2 \cdot \text{const}^{-1} \cdot \exp \frac{\Delta E}{kT}$$

- In other words: Besides the "**const.$^{-1}$**" we get the same result, but in an "easier" way.
  - The only (small) problem is: You have to know something additional for the determination of reaction constants if you just use the mass action law. And that it is not necessarily easy - it involves the concept of the chemical potential and does not easily account for factors coming from additional freedoms of orientation. e.g. the factor **z/2** in the equation above.
- The important point in this context is that the reaction equation formalism also holds for **non-equilibrium**, e.g. during the cooling of a crystal when there are too many vacancies compared to equilibrium conditions. In this case we must consider **local** instead of **global** equilibrium, see chapter 2.2.3.

- There would be much more to discuss for single vacancies in simple mono-atomic crystals, e.g. how one could calulate the formation enthalpy, but we will now progress to the more complicated case of point defects in crystals with *two* different kinds of atoms in the base.
  - That is not only in keeping with the historical context (where this case came first), but will provide much food for thought.

Questionaire

Multiple Choice questions to 2.1.1

Exercise 2.1-7

Quick Questions to 2.1.1

2.1.2 Frenkel Defects

- Frenkel defects are, like Schottky defects, a speciality of **ionic crystals**. Consult this illustration modul for pictures and more details.
  - In fact, the discussion of this defect in **AgCl** in **1926** by **Frenkel** more or less introduced the concepts of point defects in crystals to science.
- In ionic crystals, **charge neutrality** requires (as we will see) that defects come in **pairs** with opposite charge, or at least the sum over the net charge of all charged point defects must be zero.
  - "**Designer defects**" (defects carrying name tags) are *special cases* of the general point defect situation in non-elemental crystals. Since any ionic crystal consists of at least two different kinds of atoms, at least two kinds of vacancies and interstitials are possible in principle.
  - Thermodynamic equilibrium always allows *all* possible kinds of point defects simultaneously (including charged defects) with arbitrary concentrations, but always requiring a minimal free enthalpy *including the electrostatic energy components* in this case.
  - However, if there is a charge inbalance, electrostatic energy will quickly override everything else, as we will see. As a consequence we need charge neutrality in total *and* in any small volume element of the crystal - we have a kind of independent boundary condition for equilibrium.
  - Charge neutrality calls for at least *two* kinds of differently charged point defects. We could have more than just two kinds, of course, but again as we will see, in real crystals usually two kinds will suffice.
- One of two simple ways of maintaining charge neutrality with two different point defects is to always have a vacancy - interstitial pair, a combination we will call a **Frenkel pair**.
  - The generation of a Frenkel defect is easy to visualize: A lattice ion moves to an interstitial site, leaving a vacancy behind. The ion will always be the positively charged one, i.e. a cation interstitial, because it is pretty much always smaller than the negatively charged one and thus fits better into the interstitial sites. In other words; its formation enthalpy will be smaller than that of a negatively charged interstitial ion. Look at the pictures to see this very clearly.
  - It may appear that electrostatic forces keep the interstitial and the vacancy in close proximity. While there is an attractive interaction, and close Frenkel pairs do exist (in analogy to excitons, i.e. close electron-hole pairs in semiconductors), they will not be stable at high temperatures. If the defects can diffuse, the interstitial and the vacancy of a Frenkel pair will go on independent random walks and thus can be anywhere, they do not have to be close to each other after their generation.
- Having vacancies and interstitials is called **Frenkel disorder**, it consists of Frenkel pairs or the **Frenkel defects**.
  - Frenkel disorder is an extreme case of general disorder; it is prevalent in e.g. **Ag - halogen** crystals like **AgCl**. We thus have

$$n_i = n_V = n_{FP}$$

- This implies, of course, that **vacancies carry a charge**; and that is a bit of a conceptual problems. For ions as interstitials, however, their charge is obvious. How can we understand a charge "nothing"?
  - Well, vacancies can be seen as charge carriers in analogy to **holes** in semiconductors. There a missing electron - a hole - is carrying the *opposite* charge of the electron.
  - For a vacancy, the same reasoning applies. If a Na^+ lattice ion is missing, a *positive* charge is *missing* in the volume element that contains the corresponding vacancy. Since "missing" charges are non-entities, we have to assign a *negative* charge to the *vacancy* in the volume element to get the charge balance right.
  - Of course, any monoatomic crystal could (and will) have arbitrary numbers of vacancies *and* interstitials at the same time as intrinsic point defects; but only if charge consideration are important $n_i = n_V$ holds exactly; otherwise the two concentrations are uncorrelated and simply given by the formula for the equilibrium concentrations.
  - Indeed, since the equilibrium concentrations are never exactly zero, *all* crystals will have vacancies *and* interstitials present at the same time, but since the formation energy of interstitials is usually much larger than that of vacancies, they may be safely neglected for most considerations (with the big exception of Silicon!).
- Of course, in biatomic ionic crystals, there could (and will) be *two* kinds of Frenkel defects: cation vacancy and cation interstitial; anion vacancy and anion interstitial; but in any given crystal one kind will always be *prevalent*.
  - We will take up all these finer points in modules to come, but now let's just look at the simple limiting case of pure Frenkel disorder.

Calculation of the Equilibrium Concentration of Frenkel Defects

- Lets consider a simple ionic crystal, e.g. **AgCl** (being the paradigmatic crystal for Frenkel defects). With N = number of positive ions in the lattice and N' = number of interstitial sites, we obtain
  - $N' = 2N$ for interstitials in the tetrahedral position
    $N' = 6N$ for the dumbbell configuration
    $N' = ...$ etc.
  - The change of the free enthalpy upon forming n_{FP} Frenkel pairs is

$$\Delta G = n_{FP} \cdot H_{FP} - n_{FP} \cdot TS_{FP} - kT \cdot \left(\ln \frac{N!}{(N - n_V)! \cdot n_V!} + \ln \frac{N'!}{(N' - n_i)! \cdot n_i!} \right)$$

  - With H_{FP} and S_{FP} being the formation energy and entropy, resp., of a Frenkel pair. The configuration entropy is simply the sum of the entropy for the vacancy and the interstitial; we wrote n_V and n_i to make that clear (even so we already know that $n_V = n_i = n_{FP}$).
- With the equilibrium condition $\partial G / \partial n = 0$ we obtain for the concentration c_{FP} of Frenkel pairs

$$c_{FP} = \frac{n_{FP}}{N} = \left(\frac{N'}{N} \right)^{1/2} \cdot \exp \frac{S_{FP}}{2k} \cdot \exp - \frac{H_{FP}}{2kT}$$

  - The factor **1/2** in the exponent comes from equating the formation energy H_{FP} or entropy, resp., with a *pair* of point defects and not with an individual defect.
- What is the reality, i.e. what kind of formation enthalpies are encountered? Surprisingly, it is not particularly easy to find measured values; the link, however, will give some numbers.
- That was rather straight forward, and we will not discuss Frenkel defects much more at this point. We will, however, show in the next subchapter from first principles that, indeed, charge neutrality has to be maintained.

Questionaire
Multiple Choice questions to 2.1.2

Exercise 2.1-8
Quick Questions to 2.1.2 - 2.1.4

2.1.3 Schottky Defects

- **Schottky**- **defects** use the second *simple* possibility to maintain charge equilibrium in ionic crystals with two atoms in the base; they consist of the *two possible types of vacancies* which automatically carry opposite charge.
  - Look at the illustrations in the link for visualizations of Schottky defects as a well as other defects in ionic crystals.
  - You may call the *single* vacancy in metals a Schottky defect, if you like (some do), but that somehow misses the point.
  - As pointed out in the context of Frenkel defects, the vacancy in ionic crystal carries a net charge the same way a hole - a missing electron - carries a charge.
- We have postulated, that charge neutrality must be maintained, but we have not proved it. Moreover, even for equal numbers of oppositely charged vacancies (or Frenkel defects for that matter), charge neutrality can only be maintained *on average*; on a scale comparable to the average distance between the point defects, we must have electrical fields which only (on average) cancel each other for larger volumes.
  - The total formation energy therefore *must* contain some electrostatic energy part because we do have electrical fields around single point defects. The same is true for the interstitials in Frenkel defects or in the general case of mixed defects, and the consideration we are going to make for vacancies applies in an analogous way to interstitials, too.
  - Moreover, the electrical field of one vacancy will be felt by other charged point defects, which means that there is also some electrostatic interaction between vacancies, or vacancies and other charged point defects. This interaction is stronger and has a much larger range than the elastic interactions caused by the lattice deformation around a defect.
- Let's look at a relatively simple example. Here we are only going through the physical argumentation, the details of the calculations are contained in the link.
- Calling the formation energy of the **anion vacancy** (= missing anion = *positively* charged vacancy) H^+, the formation energy of the **cation vacancy** H^-, and the binding energy between close pairs H^B, we obtain for ΔG, the change in free enthalpy, upon introducing n^+, n^-, and n^B anion vacancies, cation vacancies, and vacancy pairs, resp., is

$$\Delta G = \int_V \int \int \left(H^+ \cdot n^+(\underline{r}) + H^- \cdot n^-(\underline{r}) + [H^+ + H^- - H^B] \cdot n^B(\underline{r}) + 1/2\, \rho(\underline{r})\, V(\underline{r}) - T S_n \right) dx\, dy\, dz$$

  - With $\rho(\underline{r})$ = local charge concentration, $V(\underline{r})$ = electrical potential, S_n = entropy of mixing; $\underline{r}$ is the space vector. The usual sum is replaced by an integral because the electrical potential is a smooth function and not strongly localized.
- The number of point defects is now dependent on $\underline{r}$. The (non-compensated) electrical charge stems from the charged vacancies, the net electrical charge density at any point $\underline{r}$ is thus given by

$$\rho(\underline{r}) = e \cdot \left(n^+(\underline{r}) - n^-(\underline{r}) \right)$$

  - With **e** being the elementary charge.
- The electrostatic potential follows from the charge density via the **Poisson equation**, we have

$$\Delta V(\underline{r}) = -\frac{1}{\epsilon \epsilon_0} \cdot \rho(\underline{r})$$

  - With ϵ = dielectric constant of the material and ϵ_0 = vacuum constant
- For equilibrium condition, ΔG must be minimal, i.e. we have to solve the variation problem

$$\Delta G = 0$$

  - for variations of the ***n*'s**. Using the conventional approximations one obtains solutions being rather obvious on hindsight:

$$n^{+}(\underline{r}) = N \cdot \exp \frac{H^{+} - e \cdot V(\underline{r})}{kT}$$

$$n^{-}(\underline{r}) = N \cdot \exp - \frac{H^{-} + e \cdot V(\underline{r})}{kT}$$

$$n^{B} = N \cdot z \cdot \exp \frac{H^{+} + H^{-} - H^{B}}{kT}$$

- For uncharged vacancies (where the **– eV** term would not apply), this is the old result, except that now the divacancy is included.

We still must find the electrostatic potential as a function of space. It may be obtained by expressing the charge density now with the formulas for the charged vacancy densities given above, and then solving Poissons equation. We obtain

$$\Delta V(\underline{r}) = - \frac{e \cdot N}{\epsilon\epsilon_0} \cdot \left(\exp - \frac{H^{-} + e \cdot V(\underline{r})}{k\,T} - \exp - \frac{H^{+} - e \cdot V(\underline{r})}{k\,T} \right)$$

This is a differential equation for the electrostatic potential; the problem now is a purely mathematical one: Solving a tricky differential equation.

- It is now useful to introduce a "normalized" potential **v** by shifting the zero point in a convenient manner, and by utilizing a useful abbreviation. We define

$$v := \frac{e \cdot V(r) - 0{,}5(H^{+} + H^{-})}{kT}$$

$$\chi^{-2} := \frac{2 \cdot N \cdot e^{2}}{\epsilon\epsilon_0 \cdot kT} \cdot \exp - \frac{H^{+} + H^{-}}{kT}$$

- This gives us a simple looking differential equation for our new quantities

$$\Delta v(\underline{r}) = \chi^{2} \cdot \sinh \{v(\underline{r})\}$$

χ^{-1} has the dimension of a length, it is nothing (as it will turn out) but the (hopefully) well known Debye length for our case.

The "simple" differential equation obtained above, however, still cannot be solved easily. We must resort to the usual approximations and linearize, i.e. use the approximate relation **sinh v** $\approx$ **v** for small **v**´s.

- We also need some boundary conditions as always woth differential equations. They must come from the physics of the problem.
- The first guess is always to assume an infinite crystal with **v = 0** for $x = \pm \infty$. The solutions for an infinite crystal are trivial, however, we therefore assume a crystal infinite in ***y***- and ***z***-direction, but with a surface in ***x***-direction at $x = 0$; from there the crystal extends to infinity.
- Now we have a one-dimensional problem with $\underline{r} = x$. One general solution is (please appreciate that I didn't state "obvious solution")

$$v(x) = \frac{v_0^2 - \left(\frac{v_0'}{\chi}\right)^2 + \left(v_0 + \frac{v_0'}{\chi}\right)^2 \cdot \exp\left[2\chi \cdot (x - x_0)\right]}{\left(v_0 + \frac{v_0'}{\chi}\right) \cdot \exp\left[\chi \cdot (x - x_0)\right]}$$

- with $v_0 = v(x = 0)$ and v_0' = integration constant which needs to be determined.

Cool, but we can do better yet:

- If we do not want infinities, the divergent term $\exp\left[2\chi(x - x_0)\right]$ must disappear for x approaching infinity, this means

$$\lim_{x \to \infty} v(x) = 0 \quad \Rightarrow \quad v_0 = -\frac{v_0'}{\chi}$$

$$\Rightarrow \quad v(x) = 2v_0 \cdot \exp -\chi \cdot x$$

χ^{-1} obviously determines at which distance from a charged surface, or more generally, from *any* charge, the (normalized) potential (and therefore also the real potential) decreases to **1/e** - this is akin to the definition of the Debye length.

- For a charged surface and $x >> \chi$ (i.e. the bulk of the crystal) we obtain $v(x >> \chi) = 0$, and therefore

$$V(r >> \chi) = \frac{H^+ + H^-}{2e}$$

If we substitute **V(*r*)** into the equations for the equilibrium concentrations above, we obtain the *final equations* for the vacany concentrations in the case of Schottky defects

$$n^+ = n^- = N \cdot \exp -\frac{H^+ + H^-}{2kT}$$

- i.e. both concentrations are identical, and *charge neutrality is maintained*!
- The energy costs of not doing it would be very large! That is exactly what we expected all along, *except* that now we proved it.

More important, however, we now can calculate what happens if there are electrical fields that *do not have their origin in the vacancies themselves*, but may originate from fixed charges on e.g. surfaces and interfaces (including grain boundaries or precipitates), or from the outside world.

- In this case the concentrations of the defects may be quite different from the equilibrium concentrations in a neighborhood "Debye Length" (= χ) from the fixed charges.
- We also see that we have (average) electrical neutrality in the bulk and a statistical distribution of vacancies there, *but this is not necessarily true in regions within one Debye length* χ next to an external or internal surface.

Charged surfaces thus may change point defect concentrations within about one Debye length. And charged point defects, if they are mobile, may carry an electrical current or redistribute (and then changing potentials), if surface or interface charges change.

This is the basic principle of using ionic conductors (and, to some extent, semiconductors) for *sensor applications*!

The interaction between point defects, the electrical potential and the Debye length may be demonstrated nicely by plotting the relevant curves for different sets of parameters; this can be done with the following JAVA module.

We see that the Debye length as expressed in the formulas is strongly dependent on temperature. Only for high temperatures do we have enough charged vacancies so that their redistribution can screen a external potential on short distances. For **NaCl** we have, as an example, the following values

T [K]	χ^{-1} [cm]
1100	$1.45 \cdot 10^{-7}$
900	$4.55 \cdot 10^{-7}$
700	$2.83 \cdot 10^{-6}$
500	$8.21 \cdot 10^{-5}$
300	$2.20 \cdot 10^{-1}$

- The large values, however, are unrealistic in real crystals, because grain boundaries, other charged defects, and especially impurities must also be considered in this region; they will always decrease the Debye length.

It's time for exercises!

Questionaire
Multiple Choice questions to 2.1.3

Exercise 2.1-8
Quick Questions to 2.1.2 - 2.1.4

2.1.4 Mixed Point Defects

- Here we treat the most general and still sensible case of point defects in simple ionic crystals of the type **AB**.
  - You may want to look up the illustration of the various possibilities before you proceed.
- We consider the *simultaneous* occurrence of Frenkel and Schottky defects, or the simultaneous occurrence of *two* kinds of vacancies and *one* kind of interstitial. This is the realistic general case because you may always safely neglect one of the two possible kinds of interstitials in equilibrium.
  - This case is most elegantly treated invoking a "chemical" reaction equation and the mass action law. As mentioned before, you must be aware of the ***snares*** of this approach. Lets see this by writing down a reaction equation describing first the formation of a Frenkel pair in **NaCl**. This reaction must be able to go both ways, ***i.e.*** it must describe the generation *and* the annihilation of Frenkel pairs.
  - Essentially, a *positively* charged cation, here a Na^+ ion "jumps" in an interstitial site, leaving behind a (*negatively* charged) vacancy on a **Na** lattice site, denoted V_{Na} or **cation vacancy** and generating a (of course *positively* charged) Na^+ interstitial, denoted Na_i or **cation interstitial**.
- Now you may be tempted to write this down in a first reaction equation as follows

$$0 \Leftrightarrow Na_i + V_{Na}$$

  - While this is not necessarily wrong, it is at least strange: You create, in a kind of chemical reaction, something from nothing - what keeps you from applying this equation to vacuum, which is surely not sensible? Maybe you should somehow get the crystal involved as the reference system within which things happen?
- So let's device a more elaborate system by looking at our crystal *before* and *after* a Frenkel pair was formed
  - *Before* a Frenkel pair is formed, the site occupied by the vacancy *after* the formation process is a **Na** site, we denote it by Na_{Na}. This simply means that a **Na** atom occupies a **Na** site before a vacancy is formed there.
  - At the interstitial site, where the **Na** interstitial *after* the formation process is going to be, you have nothing *before* the process. However, all those possible interstitial sites also form a lattice (e.g. the lattice of the octahedral sites); in a perfect crystal all those sites are *occupied by vacancies*, we consequently denote an empty interstitial site by V_i = vacancy on an interstitial site.
  - A **Na** ion on an interstitial site then is Na_i, and a **Na** vacancy becomes V_{Na} . Now we can write down a reaction equation that reads

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

- This looks like a cool reaction equation, we now create a Frenkel pair within a crystal and not out of thin air.
  - Indeed, the reaction equation *does* look so much better this way! Small wonder, we just invented part of the so-called **Kröger**-**Vink** **notation**, in use since the fifties of the **20th** century - not all that long ago, actually.
  - This notation is also called notation by **structure elements** and it is very useful for formulating all kinds of reactions involving point defects. However, the first law of economics applies ("There is no such thing as a free lunch"):

    Don't use the mass action law uncritically with these kinds of reaction equations!. The reason for this is simple, but usually never mentioned in the context of chemical reaction formulation:

    A *proper* reaction equation contains *only* reaction partners that are **independent**

  - This means that you can, in principle, change the concentration of *every* reaction partner *without* changing the others.
- Consider for example the following purely chemical simple reaction equation:

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

  - You can put arbitrary amounts of all three reaction partners in a container and change any individual amount at will without changing the others.
- In our reaction equation for point defects, however, you *cannot* do this. If you consider, e.g., to change the Na_{Na} concentration a little, you *automatically* change V_i, too - those quantities are *not independent*!

- This was the bad news about using Kröger-Vink notation. The good news are: In most practical cases it doesn't matter! Chapter 2.4 - often alluded to - will contain details about all of this.

It is not easy to grasp the reaction equation concept for point defects in all its complexity, but it is worthwhile if you want to dig deeper into point defects. For the purpose of this paragraph let's just postulate that the two *sums* left and right of the reaction equation would constitute the proper reactants (those sums, by the way, are called **building elements** in the **Schottky notation**).

- Be that as it may, we now apply the mass action law, keeping in mind that the reaction equation from above in full splendor actually contains a reaction enthalpy $G_{Reaction}$, i.e.: $Na_{Na} + V_i + G_{Reaction} \Leftrightarrow Na_i + V_{Na}$)

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

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

- $G_{Reaction}$, of course, is the free enthalpy change of the crystal upon the formation of one **mol** of Frenkel pairs. If we relate it to **1** Frenkel pair, it becomes H_{FP}.
- The **[...]** are the *molar* concentrations of the respective quantities if we use *molar* reaction enthalpies.
- OK, now let's spell it out. If we have a crystal with N *mols* of **NaCl**, we have the *molar* concentration of $[Na_{Na}] = N$ for really obvious reasons.
- $[V_i] = N$, most likely, will hold, too, but here we may have to dig deeper. How many different places for interstitials do we have in the given unit cell? We can figure it out, but for the sake of generality their molar concentration could be larger or N' - so it can be different in principle from N as we have seen before.

Since we usually go for atomic concentrations, we note that $c_V(C)$ = *atomic* concentration of the cation-vacancy = $[V_{Na}] / N$ and $c_i(C)$ = *atomic* concentration of the cation-interstitial = $[Na_i] / N$ we now obtain *one* equation for the *two* unknowns $c_V(C)$ · and $c_i(C)$, which will be the *first* of the equations we will need for what follows.

$$c_V(C) \cdot c_i(C) = \frac{N'}{N} \cdot \exp - \frac{H_{FP}}{kT} \qquad (1)$$

Note that this is *not* our old result, because it does *not* imply that $c_V = c_i$. All the mass action law can do is to supply *one* equation for whatever number of unknowns.

We need a second independent equation. This is - of course (?) - always **electroneutrality**. Looking just at Frenkel pairs, we have directly

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

for Frenkel Pairs only

- Now we have two equations for two unknown concentrations that we could easily solve.

However, we are interested in *mixed defects* here, so we must also consider *Schottky defects* and then mix them with Frenkel defects, always maintaining electroneutrality.

We might now go through the same procedure as before by using a similar reaction equation for Schottky defects - with a few more complications in finding the proper reaction equation. We will not do this here (do it yourself or use the link), just note the rather simple result:

- With $c_V(A)$ and $c_V(C)$ denoting the vacancies on the anion or cation sublattice, resp., and with H_S = formation enthalpy of a Schottky pair, we obtain for a *second* equation

$$c_V(A) \cdot c_V(C) = \exp - \frac{H_S}{kT} \qquad (2)$$

Again, this is *not* the old equation for Schottky defects - the concentrations are *not* necessarily equal once more

- Note that the vacancies on the anion or cation sublattice are *positively* or *negatively* charged - opposite to the charge of the (*negatively* charged) anion or (*positively* charged) cation that was removed! A **cation vacancy** thus carries a *negative* charge and so on, whereas a **cation interstitial** carries a *positive* charge. Look at the illustrations if you are not clear about this!

- Knowing that electroneutrality has to be maintained (look at the direct calculation for Schottky defects), we introduce electroneutrality now for the more general case of our *three* charged defects: The sum of all charges on the point defects must be zero; we obtain the third equation

$$c_V(C) = c_V(A) + c_i(C) \qquad (3)$$

- Or: *Sum over all negative charges = Sum over all positive charges*.

- Now we have **3** equations for **3** unknown concentrations, which can be solved with ease (haha). We obtain for the general situation of mixed defects

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

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

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

- These equations contain the "pure" Frenkel and Schottky case as limiting cases.

- Was that worth the effort? Probably not - as long as you just look at simple ionic crystals (where one defect type will prevail anyway). being in simple equilibrium without considering surfaces and the environment.
  - *However!* In real life, where point defects in ionic (and oxide) crystal are used for **sensor** applications, this kind of approach is the *only* way to go! It will be far more complicated, there will be approximations and "short-cuts", but the basic kind of reasoning will be the same.

- Now it is time for an exercise

Exercise 2.1.5

Do the math and solve equations (1) - (3)

- We now have the general equations and thus can answer the essential question for this case: How likely is a mixed case?
  - In a more quantitative fashion we ask: How different must the formation enthalpies be if just one defect type should dominate?
  - That is, of course, an excellent exercise question.

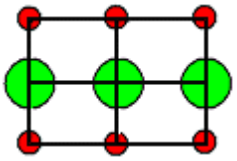
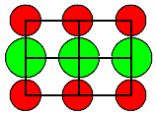
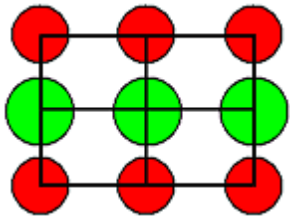
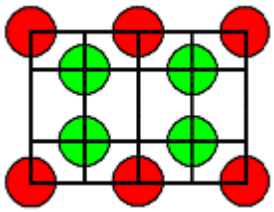
Exercise 2.1.6

Enthalpy difference for the limiting cases

- If you do the exercise, you will find out that relatively small differences in the order of **0.1 eV** in the formation energies of Schottky or Frenkel defects will already lead to the preponderance of one defect type.
  - This is important! Since it is quite likely that formation enthalpies for Schottky or Frenkel defects in a given crystal will differ by more than just **0.1 eV**, we are quite justified to look at the defect situation in an "either - or" mode: "Either we will have Frenkel defects, or Schottky defects. One kind will practically always "win".

- The final questions are then: When do we have Schottky disorder, when do we have Frenkel disorder? What are typical formation enthalpies? Are there general criteria for what kind is more likely to occur in a given crystal? Can we tell or at least make an educated guess?

Well let's look at a few experimental findings (all numbers are from "Hayes and Stoneham" ; more numbers can be found in the link):

Crystal	Type of lattice	Type of disorder	Formation enthalpies $H_{S/F}$ [eV]		<110>projection Ion size to scale Cations = red Anions = green
NaCl and all other alkali halides	NaCl-type	**Schottky**	LiF	2,5	 projection" TITLE="click to enlarge size">
			NaCl	2,3	
			KCl	2,3	
			KBr	2,4	
			CsI	1,9	 projection" TITLE="click to enlarge size">
Some Oxides			MgO	5,7	
AgCl and AgBr,	NaCl-type	**Frenkel** with cation interstials	AgCl	1,45	 projection" TITLE="click to enlarge size">
$NaNO_3$, KNO_3, ..	complex				
CaF_2 and SrF_2, BaF_2	CaF_2 type	Frenkel with anion interstials (**"Anti-Frenkel"**)	CaF_2	2,7	
			SrF_2	2,3	
			BaF_2	1,9	 projection" TITLE="click to enlarge size">
			$SrCl_2$	1,7	

What we see is:

1. It is hard to make a prediction of what kind of defect you will find. Even looking at a lattice projection with the ions drawn to scale (using tabulated ionic radii) does not offer immediate clues (besides the obvious one that you only would expect *Anti-Frenkel defects* with anion interstitials, if the anion is not much larger than the cation).

2. The formation enthalpies of Frenkel defects tends to be a bit lower than those of Schottky defects.

3. The formation enthalpies in the more simple ionic crystals (no oxides) tend to be rather small - around **1 eV** per single defect; quite comparable to the vacancy formation enthalpy of simple metals. This might be taken as a hint that charge matters less than bonding and lattice distortion.

Exercise 2.1-8

Quick Questions to 2.1.2 - 2.1.4

One more word of warning: Be careful in applying mass-action laws!

- If you would, e.g., attempt to carry over the results obtained here to **Si** (because its the only known elemental crystal where interstitials occur in sizeable quantities), you would be making a *mistake*, because the reaction equation describing the formation of Frenkel pairs given above is *not* the right equation for **Si**.
- The reason for this is that your independent quantities are different: You may add vacancies *without changing the interstitial count* because electroneutrality in a semiconductor may be achieved by changing *electron or hole concentrations* and you must not only balance charged point defects against each other.
- In addition, never forget that chemical reaction equations always demand that there are *no other reactions* (e.g. parasitic reactions) that consume non-negligible quantities of the reactants. What about other crystal lattice defects? A vacancy, after all, may disappear at a dislocation; causing the dislocation to climb; or it may react with precipitates. Some precipitates even need vacancies in order to be to be able to grow. Others emit interstitials while growing - you must include those reactions in your description of the system if it is to be correct.

2.1.5 Essentials to Chapter 2.1: Point Defect Equilibrium

In global equilibrium all crystals contain point defects with a concentration c_{PD} given by an Arrhenius expression of the form:

$$c_{PD} = A \cdot \exp -\frac{G_F}{kT} = A \cdot \exp \frac{S_F}{k} \cdot \exp -\frac{H_F}{kT}$$

- ***A*** is a constant around (**1 ....10**), reflecting the geometric possibilities to introduce **1 PD** in the crystal (***A* = 1** for a simple vacancy).
- G_F, H_F, S_F are the free energy of formation, enthalpy (or colloquial "energy") of formation, and entropy of formation, respectively, of **1 PD**
- The entropy of formation reflects the disorder introduced by **1 PD**; it is tied to the change in lattice vibrations (circle frequency ω) around a **PD** and is a measure of the extension of the **PD**. It must not be confused with the entropy of mixing for many **PDs**!

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

- Formation enthalpies are roughly around **1 eV** for common crystals ("normal" metals"); formation entropies around **1 k**.

Small **PD** clusters (e.g. di-vacancies) are still seen as **PDs**, their concentration follows from the same considerations as for single **PDs** to:

$$c_{2V} = \frac{z}{2} \cdot \exp \frac{S_{2V}}{k} \cdot \exp -\frac{H_{F(2V)}}{kT}$$

$$c_{2V} = \frac{z}{2} \cdot c_{1V}^2 \cdot \exp \frac{\Delta S_{2V}}{k} \cdot \exp -\frac{E_{2V}}{kT}$$

- The constant ***A*** for di-vacancies is half the coordination number ***z*** (= number of possibilities to arrange the axis of a di-vacancy ***dumbbell)***
- The formation enthalpy and entropy of a **PD** cluster can always be expressed as the sum of these parameters for single **PDs** minus a binding enthalpy ***E*** and a binding entropy ΔS
- The term c_{1V}^2 or c_{1V}^n for a cluster of ***n*** vacancies makes sure that the concentration of clusters is always far smaller than the concentration of single **PDs**.

The same relations can be obtained by "making" di-vacancies (or any cluster) by a "chemical" reaction between the **PDs** and employing the mass action law:

$$1V + 1V \Leftrightarrow V_{2V} + E_{2V}$$

$$\frac{(c_{1V})^2}{c_{2V}} = K(T) = \text{const} \cdot \exp -\frac{\Delta E}{kT}$$

- There are, however, some pitfalls in using the mass action law; we also loose any information about the factor ***A***
- Most important in doing "defect chemistry" with mass action, is a proper definition of the "ingredients" to chemical reaction equations. A vacancy, after all, is not an entity like an atom that can exist on its own. More to that in chapter 2.4.

- Note: All of the above is generally valid for *all independent* **PD**s: "**A**" and "**B**" vacancies, interstitials, antisite defects,. ... .
  - *However:* If there are *additional restraints* (like charge neutrality), we may have to consider pairs of (atomic) **PDs** as one point defect; e.g. Frenkel or Schottky defects.
  - First principle" calculation show that charge neutrality can only be locally violated on length scales given by the *Debye length* of the crystal.

$$c_V = A_V \cdot \exp\frac{S_F^V}{k} \cdot \exp -\frac{H_F^V}{kT}$$

$$c_i = A_i \cdot \exp\frac{S_F^i}{k} \cdot \exp -\frac{H_F^i}{kT}$$

- Frenkel and Schottky defects are vacancy-interstitial or vacancy$^-$- vacancy$^+$ pairs in ionic crystals.
  - They are extreme cases of the general "mixed defect case" containing all possible **PDs** (e.g. **V^-, V^-, i^+, i^+**) while maintaining charge neutrality.
  - Usually, one finds either Frenkel defects or Schottky defects - if the respective formation enthalpies H_{Fre} or H_{Scho} differ by some **0.1 eV**, one defect type will dominate.
  - It is, however, hard to predict the dominating defect type from "scratch".

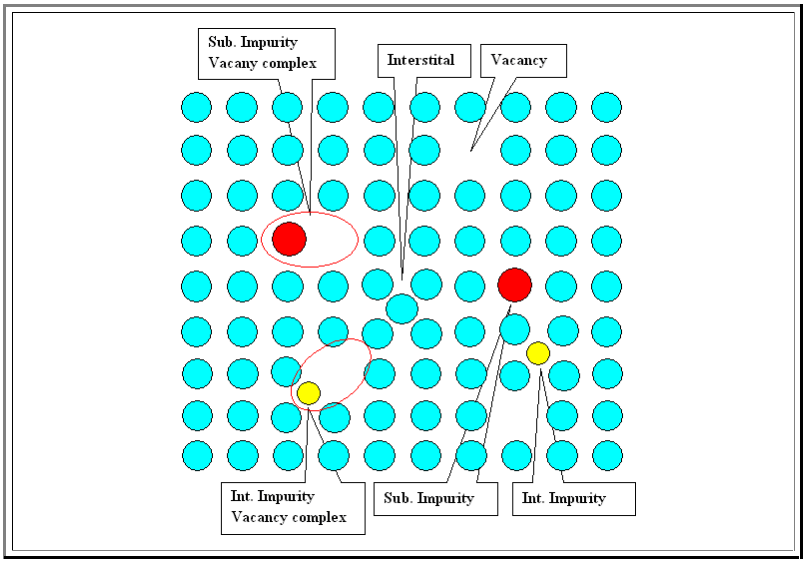
Frenkel defect:	**$V^- + i^+$**
Anti-Frenkel defect:	**$V^+ + i^-$**
Schottky defect:	**$V^- + V^+$**

Frenkel disorder in:	**AgCl, AgBr, CaF_2, BaF_2, PbF_2, ZrO_2, UO_2, ...**
Schottky disorder in:	**LiF, LiCl, LiBr, NaCl, KCL, KBr, CsI, MgO, CaO, ...**

2.2 Extrinsic Point Defects and Point Defect Agglomerates

2.2.1 Impurity Atoms and Point Defects

- Consider a *real* crystal - take even a hyperpure single crystalline **Si** crystal if you like. It's not perfect! It just is not. It will always contain some impurities. If the impurity concentration is below the **ppm** level, then you will have **ppb**, or **ppt** or **ppqt** (figure that out!), or... - it's just never going to be zero.
  - The highest vacancy concentration your are going to have in simple metals close to the melting point is around $\mathbf{10^{-4}}$ **= 100 ppm**; in **Si** it will be far lower. On the other hand, even in the best **Si** you will have some **ppm** of $\mathbf{O_i}$ (oxygen interstitials) and $\mathbf{C_i}$(Carbon substitutionals).
  - In other words - it is quite likely that besides your **intrinsic** equilibrium point defects (usually vacancies) squirming around in equilibrium concentration, you also have comparable concentrations of various **extrinsic** non-equilibrium point defects. So the question obviously is: what is going to happen between the vacancies and the "dirt"? How do intrinsic and extrinsic point defects interact?
- Let's look at the impurities first. Essentially, we are talking **phase diagrams** here. If you know the phase diagram, you know what happens if you put increasing amounts of an **impurity atom** in your crystal. Turned around: If you know what your impurity "does", you actually can construct a phase diagram.
  - However, using the word "*impurity*" instead of "*alloy*" implies that we are talking about *small* amounts of **B** in crystal **A**.
- The decisive parameter is the **solubility** of the impurity atom as a function of temperature.
  - In a first approximation, the equilibrium concentration of impurity atoms is given by the usual **Arrhenius representation**, akin to the case of vacancies or self-interstitials. This is often only a good approximation below the eutectic temperature (if there is one). Instead of the formation energies and entropies, you resort to **solubility energies** and **entropies**.
  - There is a *big difference with intrinsic point defects*, however. The concentration of impurity atoms in a given crystal is pretty much constant and not a quantity that can find its equilibrium value. After all, you can neither easily form nor destroy impurity atoms contained in a crystal.
  - That means that thermal equilibrium is only obtained at *one* specific temperature, if at all. For all other temperatures, impurity atoms are either **undersaturated** or **oversaturated**.
- Now the obvious: Vacancies, divacancies, interstitials etc. may interact with impurity atoms to form **complexes** - provided that there is some attractive interaction. Interactions may be elastic (e.g. the lattice deformation of a big impurity interstitial will attract vacancies) or electrostatic if the point defects are charged. Schematically it may look like this:



  - An impurity - vacancy complex (also known as **Johnson** complex) is similar to a divacancy, just one of the partners is now an impurity atom. The calculation of the **equilibrium concentration** of impurity - vacancy complexes thus proceeds in analogy to the calculations for double vacancies, but it is somewhat more involved. We obtain (for details use the link).

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

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

- With c_C = concentration of vacancy-impurity atom complexes, c_F = concentration of impurity atoms, c_V = equilibrium concentration of (single) vacancies, and ΔS_C or H_C = binding entropy or enthalpy, resp., of the pair. z, again, is the coordination number.
- That the coordination number z appears in the equation above is not surprising - after all there are always z possibilities to form one complex. Note that the term $1 - z \cdot c_F$ must be some correction factor, obviously accounting for the possible case of rather large impurity concentrations c_F. Why? - Well, for small c_F, this term is just about **1** and we get the approximation from above.
- Note also that as far as equilibrium goes, we have a kind of mixed case here. The impurity atoms have some concentration c_F that is *not* an equilibrium concentration. But if we redefine equilibrium as the state of crystal plus impurities (essentially we simply change the G_0 = Gibbs energy of the "perfect" crystal in one of our first equations), than the concentration c_C of vacancy-impurity atom complexes is an *equilibrium* concentration.

The equation above for c_C is quite similar to what we had for the divacancy concentration.

- If you forget the "correction factor" for a moment, we have identical exponential terms describing the binding enthalpy, and pre-exponential factors of $z \cdot c_V \cdot c_F$ for divacancies and $z \cdot c_V \cdot c_V$ for the vacancy - impurity complexes.
- In both cases the concentrations decreases exponentially with temperature. However, assuming identical binding enthalpies for the sake of the argument, in an Arrhenius plot the slope for divacancies would be twice that of vacancy-impurity complexes - I sincerely hope that you can see why!

The *total* vacancy concentration c_{1V}**(total)** (= concentration of isolated vacancies + concentration of vacancies in the complexes) as opposed to *the equilibrium concentration without impurities* c_{1V}**(eq)** is given by

$$c_V(\text{total}) = c_V(\text{eq}) + c_C$$

- That's what equilbrium means! If impurity atoms snatch away some vacancies that the crystal "made" in order to be in equilibrium, it just will make some more until equilibrium is restored.
- c_C thus can be seen as a correction term to the case of the perfect (impurity free) crystal which describes the perturbation by impurities. This implies that $c_V >> c_C$ under normal circumstances.

We will find out if this is true and more about vacancy - impurity complexes in an exercise.

- You don't have to do it all yourself; but at least look at it - it's worth it.

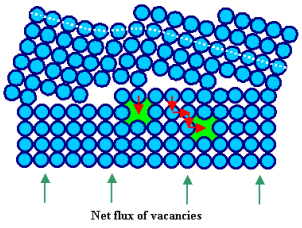
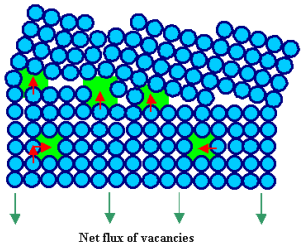
Exercise 2.2-1
Properties of Johnson complexes

Questionaire
Multiple Choice questions to 2.2.1

Exercise 2.2-2
Quick Questions to 2.2

2.2.2 Local and Global Equilibrium

- **Global thermal equilibrium** at arbitrary temperatures, i.e. the *absolute* minimum of the free enthalpy, can only be achieved if there are mechanisms for the **generation** and total **annihilation** of point defects.
  - This means there must be **sources** and **sinks** for vacancies and (intrinsic) interstitials that operate with small activation energies - otherwise it will take a long time before global equilibrium will be achieved.
- At this point it is essential to appreciate that an *ideal perfect* (= infinitely large) crystal has *no* sources and sinks - it can *never* be in thermal equilibrium.
  - An atom, to be sure, cannot simply disappear leaving a vacancy behind. Even if the crystal is finite, it cannot simply disappear leaving a vacancy behind and then miraculously appear at the surface, as we assumed in equilibrium thermodynamics, where it does not matter *how* a state is reached.
  - On the other hand, infinitely large perfect crystals do not exist - but semiconductor-grade dislocation-free single **Si** crystals with diameters of **300 mm** and beyond, and lengths of up to **1 m** are coming reasonably close. These crystals form a special case as far as point defects are concerned but nevertheless incorporate point defects in equilibirum.
- In real life we need *other defects* - surfaces, crystal-melt interfaces, grain boundaries, dislocations, precipitates, and so on, as sources and sinks for point defects. In regular metals or ceramics and so on, we have almost always plenty of those defects.
- How a **grain boundary** can act as source or sink for vacancies is schematically shown in the pictures below.
  - It is clear from these drawings that the activation energy (which is *not* the formation energy of a vacancy!!) needed to *emit* (not to form from scratch!) a vacancy from a grain boundary is small.

<table>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Grain boundary absorbs 1 vacancy, i.e. acts as
sink after one more jump of the proper atom.</td><td>Grain boundary emits 3 vacancies, i.e. acts as
source after one more jump of the 3 proper atoms.</td></tr>
<tr><td colspan="2">The red arrows indicate the jumps of individual atoms. The flux of the vacancies is always opposite to the flux of diffusing atoms.</td></tr>
</table>

- We thus may expect that at sufficiently high temperatures (meaning temperatures large enough to allow diffusion), we will be able to establish global *point defect equilibrium* in a real (= non-ideal) crystal, but not really global *crystal* equilibrium, because a crystal with dislocations and grain boundaries is never at global equilibrium.
- Sources and sinks are a thus a *necessary*, but not a *sufficient* ingredient for point defect equilibrium. We also must require that the point defects are able to move, there must be some diffusion - or you must resign yourself to waiting for a long time. In other words, we must look at the *temperature* now.
  - At *low temperatures*, when all diffusion effectively stops, nothing goes anymore. Equilibrium is unreachable. For many practical cases however, this is of no consequence. At temperatures where diffusion gets ***sluggish***, the equilibrium concentration c_{eq} is so low, that you cannot measure it. For all practical purposes it surely doesn't matter if you really achieve, for example, $c_{eq} = 10^{-14}$, or if you have non-equilibrium with the actual concentration c a thousand times larger than c_{eq} (i.e. $c = 10^{-11}$). For all practical purposes we have simply $c = 0$.
  - At *high temperatures*, when diffusion is fast, point defect equilibrium will be established very quickly in all real crystals with enough sources and sinks.
- The *intermediate temperatures* thus are of interest. The mobility is not high enough to allow many point defects to reach convenient sinks, but not yet too small to find other point defects.
  - In other words, the average **diffusion length** or mean distance covered by a randomly diffusing point defect in the time interval considered, is smaller than the average distance between sinks, but larger than the average distance between point defects.
  - This is important, so let's say it once more in yet other words: In the intermediate temperature range we are considering here, a given vacany will still be able to move around sufficiently to encounter another vacancy, but not a dislocation, precipitate or grain boundary.

- Global point defect equilibrium as the best state of being is thus unattainable at *medium temperatures*. **Local equilibrium** is now the second best choice and far preferable to a huge supersaturation of single point defects slowly moving through the crystal in search of sinks.
  - Local equilibrium then simply refers to the state with the smallest free enthalpy *taking into account the restraints of the system*. The most simple restraint is that the total number of vacancies in vacancy clusters of all sizes (from a single vacancy to large "voids") is constant. This acknowledges that vacancies cannot be annihilated at sinks under these conditions, but still are able to cluster.
- Let us illustrate this with a relevant example. Consider vacancies in a metal crystal that is cooled down after it has been formed by casting.
  - As the temperature decreases, global equilibrium demands that the vacancy concentration decreases exponentially. As long as the vacancies are very mobile, this is possible by annihilation at internal sinks.
  - However, at somewhat lower temperatures, the vacancies are less mobile and have not enough time to reach sinks like grain boundaries, but can still cover distances much larger than their average separation. This means that divacancies, trivacancies and so on can still form - up to large clusters of vacancies, either in the shape of a small hole or void, or, in a two-dimensional form, as small dislocation loops. Until they become completely immobile, the vacancies will be able to cover a distance given by the diffusion length L (which depends, of course, on how quickly we cool down).
  - In other words, at intermediate temperatures small vacancy clusters or agglomerates can be formed. Their maximum size is given by the number of vacancies within a volume that is more or less given by L^3 - more vacancies are simply not available for any one cluster.
  - Obviously, what we will get depends very much on the cooling rate and the mobility or diffusivity of the vacancies. We will encounter that again; here is a link looking a bit ahead to the situation where we cool down as fast as we can.
- It remains to find out which mix of single vacancies and vacancy clusters will have the smallest free enthalpy, assuming that the total number of vacancies - either single or in clusters - stays constant. This minimum enthalpy for the specific restraint (number of vacancies = const.) and a given temperature then would be the **local equilibrium configuration** of the system.
- How do we calculate this? The simplest answer, once more, comes from using the the mass-action law. We already used it for deriving the equilibrium concentration of the divacancies. And we did *not* assume that the vacancy concentration was in global thermal equilibrium! The mass action law is valid for *any* starting concentrations of the ingredients - it simply describes the equilibrium concentrations for the set of reacting particles present. This corresponds to what we called local equilibrium here.
  - The reaction equation from sub-chapter 2.2.1 was $\mathbf{1V + 1V \Leftrightarrow V_2}$ and in this case this is a valid equation for using the mass action law. The result obtained for the concentration of divacancies with the single vacancy concentration in *global* thermal equilibrium was

$$c_{2V} = (c_{1V})^2 \cdot \frac{z}{2} \cdot \exp\frac{\Delta S_{2V}}{k} \cdot \exp\frac{B_{2V}}{kT}$$

  - Don't forget that concentrations here are defined as n/N, i.e. in *relative* units (e.g. $c = 3{,}5 \cdot 10^{-5}$) and not in absolute units, e.g. $c = 3{,}5 \cdot 10^{15}\ \mathrm{cm}^{-3}$.
  - For arbitrary clusters with n vacancies $\mathbf{(1V + 1\ V + ... + 1V \Leftrightarrow V_n)}$ we obtain in an analogous way for the concentration c_{nV} of clusters with n vacancies

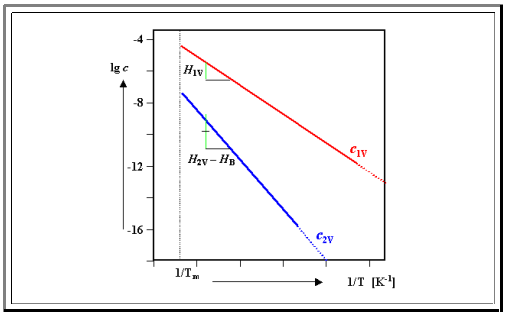
$$c_{nV} = (c_{1V})^n \cdot \alpha \cdot \exp\frac{\Delta S_{nV}}{k} \cdot \exp\frac{B_{nV}}{kT}$$

  - with B_{nV} = average binding energy between vacancies in an n-cluster, $c_{1V} = \mathrm{const.}$ concentration of the vacancies (*and no longer the thermal equilibrium concentration !*), and α = number of possible "orientations" of the n-cluster divided by the indistinguishable permutations. The value of α will depend to some extent how we arrange n vacanies: in a row, on a plane or three-dimensionally - but we won't worry about that because the other factors are far more important.

The essential point now is to realize that these equations still work for local equilibrium! They now describe the *local equilibrium of vacancy clusters* if a *fixed* concentration of vacancies is given. The situation now is *totally different* from global equilibrium. If we consider divacancies for example, we have:

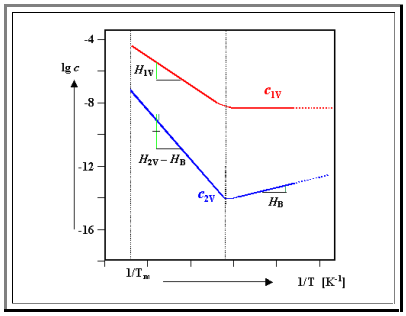
Global equilibrium

- $c_{2V}(eq) << c_{1V}(eq)$; and $c_{2V}(eq)$ rapidly *decreases* with decreasing temperatures since $c_{1V}(eq)$ decreases.



Local equilibrium

c_{2V} is *increasing* with decreasing T since c_{1V} *stays about constant*, but we still have the $\exp + B_{nV}/kT$ term that increases with T



Whereas the concentration of clusters may still be small, they now contain most of the vacancies.

- Generally speaking, it is always energetically favorable, to put the surplus vacancies in clusters instead of keeping them in solid solution if there is no possibility to annihilate them completely. It thus comes as no surprise that in rapidly cooled down crystals with not to many defects that can act as sinks, we will find some vacancy clusters at room temperature

It also should come as no surprise that the same is true for impurity atoms - vacancy clusters. The equations governing this kind of point defect agglomeration are, after all, quite similar to the equations discussed here.

- If you now take the extreme case of a rather perfect **Si** single crystal (no sinks for point defects), where we do not just have vacancies at thermal equilibrium, but also some relevant concentration of interstitials, interstitial oxygen and substitutional carbon, you might well wonder what one will find at room temperature.
- Well - don't wonder! Get to work! It is not all that clear. And even if that puzzle has been solved before you reach productive scientisthood, there is always **GaAs**, or **InP**, or **SiC**, or - well, you will find something left to do, don't worry.

It's time for exercises!

Questionaire

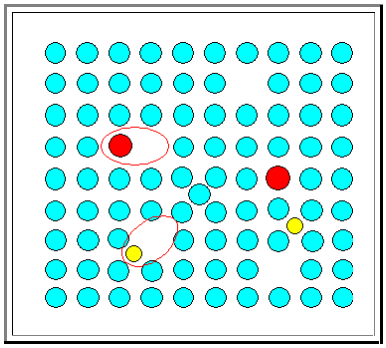
Multiple Choice questions to 2.1.1

Exercise 2.1-8

Quick Questions to 2.1.2 - 2.1.4

2.2.3 Essentials to Chapter 2.2: Extrinsic Point Defects and Point Defect Agglomerates

- Besides intrinsic point defects, crystals always contain extrinsic point defects - impurity atoms on substitutional or interstitial sites.
  - The concentration c_F of extrinsic point defects is pretty much constant - it cannot be in equilibrium. What is going on in a macroscopic way is given by the phase diagram of crystal plus impurity atoms
  - If we discuss extrinsic point defects (in contrast to alloys), we discuss small c_F values. Nevertheless, c_F might be much larger than c_C, the equilibrium concentration of vacancies; especially at low temperatures.



- The concentration c_C of impurity atom - vacancy complexes is easy to calculate, the decisive parameters are the concentrations of the partners and their binding energy H_C
  - The situation is quite similar to the case of divacancies or multi-vacancies, except that c_F is constant. The concentration of extrinsic-intrinsic complexes *in equilibrium* decreases with temperature

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

- With extrinsic point defects the crystal can no longer be in *global* equilibrium; what we are looking now is local equilibrium - the minimum of the free enthalpy obtainable under some fixed circumstances
- Fixed circumstances may include that the concentration of the intrinsic point defects is not in (global) equilibrium. There are many reasons for that:
  - Changing the equilibrium concentration of intrinsic point defects needs two ingredients:
    1. sources and sinks for intrinsic point defects - external or internal surfaces (= grain boundaries), dislocations - other defects.
    2. Sufficient mobility of the point defects to get away from or to the sources and sinks, respectively.
- The latter condition always fails at low temperatures, as a result the formation of point defect clusters is favored.
  - The decisive quantity is the total diffusion length L of a point defect during the thermal history of its crystal - how far can it "go" before it is frozen into immobility
  - If $L >$ average distance to a sink, we will find mostly equilibrium conditions; if $L <$ average distance to a sink (**Si** case!), we will have to expect point defects complexes of n point defects at a concentration c_{nV}.
  - The upper limit for n is the concentration of point defects contained in the volume L^3.

$$c_{nV} = (c_{1V})^n \cdot \frac{z}{2} \cdot \exp\frac{\Delta S_{nV}}{k} \cdot \exp\frac{B_{nV}}{kT}$$

$c_{nV}(T)$ *increases* with $\exp\{B_{nV} / kT\}$ as soon as c_V stays constant.

2.3. Point Defects in Semiconductors like Silicon

2.3.1 General Remarks

- In all semiconductors, lattice defects change the electronic properties of the material locally, and this may result in electronic energy states in the band gap of the semiconductor and this is true for all kinds of lattice defects
  - Semiconductor technology actually depends completely on this fact. **Doping** a semiconductor, after all, mostly means the incorporation of (usually) substitutional **extrinsic point defects** in defined concentrations in defined regions of the crystal - we have **B**, **As** and **P** for **Si**.
  - Our extrinsic point defects now exist in two states: we have some concentration $[P]^0$ of e.g. a neutral donor like **P** and some concentration $[P]^+$ of ionized donors; and $[P]^0 + [P]^+ = [P_0]$, the total concentration of **P** holds at all conditions.
  - The concentration $[P]^+$ is simply given by the the total concentration times the probability that the electronic state associated with the **P** impurity atom is *not* occupied by an electron.
  - If this electrons state is at an energy E_D in the band gap, basic semiconductor physics tells us that for a given E_F and temperature T the concentration of ionized impurity atoms is given by

$$[P^+] = [P^0] \cdot \{1 - f(E_D, E_F; T)\}$$

$$\approx [P^0] \cdot \exp \frac{E_F - E_D}{kT}$$

- There is no reason whatsoever that a vacancy (or any other point defect you care to come up with) should not have a energy level (or even more than one) in the band gap of its host semiconductor. This level then will be occupied or not occupied by electrons exactly like the extrinsic point defect.
  - If the vacancy is mobile at the temperature considered, it will diffuse around - exactly like an extrinsic mobile defect.
  - If the temperature changes, the intrinsic point defects concentration changes to the extent that it can establish equilibrium - *in pronounced contrast to the extrinsic point defects*.
- It should be clear form this, that intrinsic point defects in semiconductors are not all that simple. Charge states must be considered that depend on primary doping with extrinsic point defects and temperature. If things get really messy, the intrinsic point defects change the actual doping and their mobility (or diffusion coefficient) depends on their charge state.
- Looking at jus at few topics in the case of **Si**, we obtain a bunch of complex relations, which shall only be touched upon:
  - Once again, the equilibrium concentration of charged point defects depends on the **Fermi energy** E_F (which is the chemical potential of the electrons). As an example, for a negatively charged vacancy we obtain

$$c(V^-) = c(V) \cdot \exp \frac{E_F - E_A}{kT}$$

  - With E_F = Fermi energy, and E_A = *acceptor level of the vacancy* in the band gap.
  - This tells us that besides the formation energies and entropies, we now also must know the **energy levels** of the defects in the band gap!
  - The dependence of the concentration of arbitrarily charged point defects on the carrier concentration (i.e. on doping) is given by

$$\frac{c_{Vx}(n)}{c_{Vx}(n_i)} = \left(\frac{n}{n_i}\right)^{-x}$$

  - With n_i, n = (intrinsic) carrier density, x = charge state of point defect.
- As a ***Si** special*, we also must consider **self-interstitials** (which, if you remember, we always can safely neglect for just about any other elemental crystal)
  - Local equilibrium between vacancies and interstitials follows this relation:

$$c_V(\text{loc}) \cdot c_i(\text{loc}) \approx c_V(\text{equ}) \cdot c_i(\text{equ})$$

- Considering that carrier densities and the Fermi energy depend on the temperature, too, things obviously get complicated!
- It thus should not be a big surprise that the scientific community still has not come up with reliable, or least undisputed numbers for the basic properties of intrinsic point defects in **Si**, not to mention the more complicated semiconductors.
  - But do not let yourself be deceived by this: While *you* might have problems coming up with numbers for e.g. the vacancy concentration in **Si** at some temperature and so on, the ***Si*** *crystal* has no problems whatsoever to "produce" the concentration that is just right for this condition.

- Here are two relevant articles that can be read as a *pdf file*:
  - **The Engineering of Intrinsic Point Defects in Silicon Wafers and Crystals**
    R. **Falster** and V.V. **Voronkov**
  - **Defects in Monocrystalline Silicon**
    Wilfried von Ammon, Andreas Sattler, Gudrun Kissinger; in: Springer Handbook of Electronic and Photonic Materials, Safa Kasap, PeterCapper (Eds.), 2017

2.4 Point Defects in Ionic Crystals

2.4.1 Motivation and Basics

- Point defects in ionic crystals (e.g. **NaCl** or **$AgCl_2$**) and oxides (e.g. **SnO_2** or **ZrO_2**) are quite important and put to technical uses.
  - Unfortunately (from the metal oriented persons point of view) the scientific community working with those materials has its own way for dealing with point defects, which differs in some respects from the viewpoint of the metal and semiconductor community.
  - There are historical and "cultural" reasons for this, but there are also good reasons. Essentially, in dealing with more complicated crystals - and ionic materials or oxides are always more complicated than metals or simple semiconductors - a more chemical point of view is traditional and useful. Let us look at some important points that have to be considered in this context.
- First, we look at the **stoichiometry** of these crystals.
  - Ionic crystals must consist of at least two different kinds of ions. They may then contain point defects in concentrations far above thermal equilibrium (as defined relative to a perfect crystal), if the real material is non-stoichiometric. If you imagine a single crystal of, lets say, **NaCl** with the composition **$Na_{1-\delta}Cl$** and **$\delta << 1$**, i.e close to, but not exactly at stoichiometry (which is what you always would expect in reality), your only way of forming a crystal seems to be to use some point defects as integral part of the crystal.
  - You *might* consider, e.g., to introduce a concentration of δ vacancies on the Na lattice sites, or to put a concentration of δ **Cl^-** ions in interstitial positions, or to mix both defect types in a ratio where the sum of the concentrations somehow equals δ.
  - But now *lets think again*. If you consider a crystal of **$Na_{1-\delta}Cl$**, you are really talking about a crystal with ***N*** atoms of negatively charged **Cl^-**- ions and **$N \cdot (1-\delta)$** positively charged **Na^+** ions, which means that the crystal would carry a net negative charge of **$e \cdot \delta \cdot N$** and thus a dramatically high energy. No such crystal can exist - there must always be equal numbers of **Na^+** and **Cl^-** ions - *as long as there are no impurity atoms*.
- This leads us to the second point, the necessity for **charge equilibrium** or "*zero net charge condition*" considered before.
  - If we stay with the above example of **NaCl**, we are forced to conclude that a **NaCl** crystal would be necessarily perfectly stoichiometric - it cannot grow in any other way. However, no crystal exists without some impurities. If, for example, some **Ca** atoms are to be included into an otherwise perfectly stoichiometric **NaCl** crystal, they will always be doubly charged **Ca^{++}** ions, and we now must remove twice the number of **Na^+** ions to preserve charge neutrality (or introduce twice the number of additional **Cl^-** ions). Obviously we now *must* introduce a **Na** vacancy for every **Ca^{++}** ion included in the crystal (or **Cl^-** interstitials and so on).
  - The concentration of vacancies now could be much higher than the *thermal equilibrium* concentration. But we still may have equilibrium; namely **chemical equilibrium**, or, if the defects are charged, **electrochemical equilibrium**!
- We see with this simple example, that there is a linkage between stoichiometry, charge neutrality, impurities and defects, with the added complication that it is not necessarily clear which kinds of point defects must be present in what concentration.
  - We also see that point defects in concentrations that have nothing to do with the thermal equilibrium concentration in *perfect* crystals may be an integral part of a real ionic crystal.
- The simple example, however, makes also clear that stoichiometry, impurity, and charge neutrality considerations still do not tell us exactly what kinds of point defects are needed in which concentration, but at best will give some integral numbers.
- Let us look at a third point. It concerns the surface and its interaction with the surroundings - this is where many applications come in.
  - Consider a **ZrO_2** crystal in thermal equilibrium with a gas containing a certain **O_2** concentration, at a temperature where the oxygen in the crystal is mobile to some extent (maybe because there are **O**-vacancies?). We must expect some "chemical" reaction to take place. Some additional oxygen may be incorporated into the crystal, or some oxygen may diffuse out of the crystal into the gas. The tendency of whatever is going to happen in this case will be determined by the conditions for *chemical* equilibrium, or, in other word, by the chemical potentials of the participating species.
  - But we must expect that point defects are involved in whatever happens across the interface. For the particular example given (which happens to describe the principle of an **oxygen sensor**) we must expect that some *electrical* effects take place as well because introducing excess oxygen (always negatively charged) into the crystal or taking some out, will influence the charge distributions and thus *electrical potentials* in the crystal.
  - Some *electrochemical* equilibrium will be reached that contains electrical potential differences - a voltage develops across the interface.

- The common denominator in all considerations made so far was: We always had some kind of linkage between "*chemistry*" as expressed in reactions between atoms or in stoichiometric considerations, and (usually charged) *point defects*..
- We now get the idea of what needs to be done for a general treatment of point defects and ionic crystals:

We want to define point defects in a way
where they can be included
into the familiar concept of chemical reaction equations
We then treat them the same way we treat chemical reactions

- In other words, we want to write equations analogous to

$$n \cdot A + m \cdot B \Leftrightarrow A_nB_m$$

$$\frac{[A]^n \cdot [B]^n}{[A_nB_m]} = \text{const.}$$

  - With the option that "**A**" and "**B**" may refer not only to atoms, *but also to point defects*.
- We want to do this in a way where we can use the full box of tools developed for chemical reactions, e.g. the **mass action law** (shown above), **chemical potentials** and **activities** instead of concentrations, the concept of **kinetics**, of **chemical equilibrium**, etc.
- As is turns out, this is possible, but it is not *obvious* how to do this right. There are several approaches and compromises to achieve the best description. We will look at this in the next subchapter.

2.4.2 Kröger-Vink Notation

- How do we treat point defects in perfect analogy to atoms and molecules in chemical reaction equations? A very clear way was suggested by **Kröger** and **Vink**, it is therefore called "**Kröger-Vink notation**" or notation by **"structure elements"** - we already had a glimpse of this.
  - We define vacancies and interstitials as particles which occupy a defined site in a crystal and which may have a charge.
  - Sites in a crystal are the points where the atoms, the interstitials, or the vacancies can be. For a crystal composed of two kinds of atoms we have, e.g., the "**A**-sites" and the "**B**-sites". An **A**-atom on an **A**-site we denote by $\mathbf{A_A}$, a vacancy on a **B**-site is a $\mathbf{V_B}$
  - This leaves the interstitials out of the picture. We therefore simply name all possible interstitials sites with their own place symbol and write $\mathbf{A_i}$ or $\mathbf{B_i}$ for an **A**-atom or a **B**-atom, resp., on its appropriate interstitial site.
  - An interstitial site *not* occupied by an interstitial atom then, by definition, is occupied by a vacancy and symbolized by $\mathbf{V_i}$. A perfect crystal in the Kröger-Vink notation thus is full of vacancies on interstitial sites!
- In order to facilitate book keeping with respect to the electrical charge, we only note the **excess charge relative to the neutral lattice**. Positive excess charge is marked by a point (e.g. $\mathbf{A^{\bullet}}$), negative charge by a hyphen or dash or whatever you like to call it (e.g. $\mathbf{A^{/}}$) to distinguish this relative charge from the absolute charge. If we consider a positively charged $\mathbf{Na^+}$ ion in the **NaCl** lattice, we write $\mathbf{Na_{Na}}$ as long as it is sitting on its regular lattice position, i.e. without a charge symbol. If we now consider a vacancy on the **Na**-site, the **Na**-ion as interstitial, or a $\mathbf{Ca^{++}}$ ion on the **Na**-site, we write
  - $\mathbf{V^{/}_{Na}}$, $\mathbf{Na^{\bullet}_i}$, and $\mathbf{Ca^{\bullet}_{Na}}$ because this defines the charge relative to the neutral lattice.
  - Running through all the possible combinations for our **NaCl** crystal with some **Ca**, we obtain the matrix

	A atom (Na^+)	**B atom (Cl^-)**	**Vacancy**	**C atom (Ca^{++})**
A-site	Na_{Na}	$Cl^{/}{}_{Na}$	$V^{/}_{Na}$	$Ca^{\bullet}_{Na}$
B-site	$Na^{\bullet\bullet}_{Cl}$	Cl_{Cl}	$V^{\bullet}_{Cl}$	$Ca^{\bullet\bullet\bullet}_{Cl}$
i-site	$Na^{\bullet}_{i}$	$Cl^{/}_{i}$	V_i	$Ca^{\bullet\bullet}_{i}$

- What have we gained by this? We now can describe all kinds of structure elements - atoms, molecules and defects - and their reactions in a clear and unambiguous way *relative to the empty space*. Lets look at some examples
  - Formation of Frenkel defects in, e.g., **AgCl**:

$$\mathbf{Ag_{Ag} + V_i = V^{/}_{Ag} + Ag^{\bullet}_i}$$

  - We see why we need the slightly strange construction of a vacancy on an interstitial site.
  - Formation of Schottky defects for an **AB** crystal

$$\mathbf{A_A + B_B = V^{/}_A + V^{\bullet}_B + A_A + B_B}$$

$$\mathbf{A_A + B_B = V^{/}_A + V_B^{\bullet} + AB}$$

  - The second equation simply considers the two dislodged atoms as a molecule that must be put somewhere.
- This looks *good*. The question is, if we now can use the **mass action law** to determine equilibrium concentrations. If the Frenkel defect example could be seen as analogous to the chemical reaction **A + B = AB**, we could write a mass action law as follows:

$$\mathbf{\frac{[Ag_{Ag}] \cdot [V_i]}{[V^{/}_{Ag}] \cdot [Ag_i^{\bullet}]} = const}$$

  - with **[A]** meaning "concentration of **A**". The reaction constant is a more or less involved function of pressure ***p*** and temperature ***T***, and especially the *chemical potentials* of the particles involved.

Unfortunately, this is wrong!

Why? Well, the notion of **chemical equilibrium** and thus the mass action law, at the normal conditions of constant temperature ***T*** and pressure ***p***, stems from finding the minimum of the free enthalpy ***G*** (also called *Gibbs energy*) which in our case implies the equality of all chemical potentials. You may want to read up a bit on the concept of chemical potentials, this can be done in the link.

- In other words, we are searching for the equilibrium concentration of the particles n_i involved in the reaction, which, at a given temperature and pressure, lead to $dG = 0$.
- The equation $dG = 0$ can always be written as a total differential with respect to the variables dn_i:

$$dG = \frac{\partial G}{\partial n_1} \cdot dn_1 + \frac{\partial G}{\partial n_2} \cdot dn_2 + \ldots$$

- The partial derivatives are defined as the chemical potentials of the particles in question and we always have to keep in mind that the *long version* of the above equation has a subscript at every partial derivative, which we, like many others, conveniently "forgot". If written correctly the partial derivative for the particle n_i reads (in ***HTML*** somewhat awkwardly),

$$\frac{\partial G}{\partial n_i}\Big|_{p,\, T,\, n_{j \neq i} = \text{const}}$$

- Meaning that ***T***, ***p***, and all *other* particle concentrations must be kept constant.

Only if that condition is fulfilled, a mass action equation can be formulated that involves all particles present in the reaction equation! And fulfilling the condition means that you can - at least in principle - change the concentration of *any* kind of particle (e.g. the vacancy concentration) *without* changing the concentration of all the other particles.

- This "*independence condition*" is automatically *not fulfilled* if we have additional constraints which link some of our particles. And such constraints *we do have* in the Kröger-Vink notation, as alluded to before!
- There is no way within the system to produce a vacancy, e.g. V_A without removing an **A**-particle, e.g. generating an A_i or adding another **B**-particle, B_B.

S... ! We now have a very useful way of describing chemical reactions, including all kinds of charged defects, but we cannot use simple thermodynamics! That is the point where other notations come in.

You now may ask: Why not introduce a notation that has it all and be done with it?

- The answer is: It could be done, but only by losing simplicity in describing reactions. And simplicity is what you need in real (research) life, when, in sharp contrast to text books, you do *not* know what is going on, and you try to get an answer by mulling over various possibility in your mind, or on a sheet of paper.

So "defects-in-ceramics" people live with several kinds of notation, all having pro and cons, and, after finding a good formulation in one notation, translate it to some other notation to get the answers required. We will provide a glimpse of this in the next subchapter.

2.4.3 Schottky Notation and Working with Notations

Schottky Notation

- The Kröger-Vink notation defined *structure elements* - atoms, molecules, point defects and even electrons and holes relative to *empty space*. Despite the problem with the inapplicability of the mass action law, this notation is in use throughout the scientific community dealing with point defects.
- The other important notation - the "**Schottky notation**" or "**building element notation**" is defined as follows:
  - Defects are defined relative to the *perfect crystal*.
  - Charges are notated as in the Kröger-Vink way, i.e. *relative to the perfect crystal*. We again use the "$\cdot$" for positive (relative) charge and the "**/**" for (relative) negative charge.
- To make things a bit more complicated, there are *two* ways of writing the required symbols, the "**old**" and the "**new**" Schottky notation.
  - The "*old*" Schottky notation used special graphical symbols, like black circles or squares which are not available in **HTML** anyway.
  - So we only give the *new* Schottky notation in direct comparison with the Kröger-Vink notation, again for the example **NaCl** with **Ca** impurities, i.e.
    A = Na^+, B = Cl^-, C = Ca^{++}.

	A on B site	**A-vacancy**	**A-interstitial**
Schottky (new)	$Na\|Cl\|^{\cdot\cdot}$	$\|Na\|^{/}$	$Na^{\cdot}$
Kröger-Vink	$Na_{Cl}^{\cdot\cdot}$	$V_{Na}^{/}$	$Na_i^{\cdot}$

- So far, the difference between the Schottky notation and the Kröger-Vink notation seems superficial. The important difference, however, becomes clear upon writing down defect reactions. Lets look at the formation of Frenkel and Schottky defects in the two notations.

	Frenkel defects	**Schottky defects**
Schottky (new)	$\|Ag\|^{/} + Ag^{\cdot} = 0$	$\|Ag\|^{/} + \|Cl\|^{\cdot} + AB = 0$
Kröger-Vink	$Ag_{Ag} + V_i = V^{/}_{Ag} + Ag^{\cdot}_i$	$Ag_{Ag} + Cl_{Cl} = V^{/}_{Ag} + V^{\cdot}_{Cl} + AB$

- In words, the Schottky notation says:
  - For *Frenkel defects*: A negatively charged **Ag** vacancy plus a positively charged **Ag** interstitial gives zero.
  - For *Schottky defects*: A neatively charged **Ag** vacancy plus a positively charged **Cl** vacancy plus a **AgCl** "lattice molecule" gives zero.
- This is clear enough for these simple cases, but not as clear and easy as the Kröger-Vink notation.
- But, and that is the big advantage, we can apply the mass action law directly to the reactions in the Schottky notation.
  - This is not directly obvious. After all, Frenkel defects, e.g., do not only *appear* to be linked (where there is an interstitial, there is also a vacancy), but actually *are* linked if the defects are charged (otherwise there would be neither net charge in the crystal, or we would have to invoke electrons or holes to compensate the ionic charge - but then we would have to include those into the reaction equation).
  - *Theoretically*, however, you can introduce one more vacancy or one more interstitial into a crystal with a given concentration of each and look at the change of the free enthalpy, i.e. the chemical potential of the species under consideration. The independence condition does not require that it is *easy* to change individual concentrations, only that it is *possible*!
  - If you do neglect the energy associated with charge (i.e. you look at the chemical and not the electrochemical potential), the answers you get will not contain the coupling between the defects and you have to consider that separately. We will see how this works later.

Now, why don't we use just the Schottky notation and forget about Kröger-Vink? We asked that question before; the answer hasn't changed: If we look at more complicated reactions, e.g. between point defects in an ionic crystal, a gas on its outside, and with electrons and holes for compensating charges, it is *much easier* to formulate possible reaction in the Kröger-Vink notation. The trick now is, to convert your reaction equations from the Kröger-Vink structure elements to the Schottky building elements. There is a simple recipe for doing this.

Converting Kröger-Vink to Schottky

All we have to do, is to combine the two *structure elements* of Kröger-Vink that refer to the *same place* in the lattice and view the combination as a *building element*.

Lets first look at an example and then generalize. Consider the Frenkel disorder in **AgCl**. Using structure elements, we write

$$Ag_{Ag} + V_i = V'_{Ag} + Ag^{\cdot}_i$$

- Combining the terms referring to the same place in the lattice (with the actual defects always as the first term in the combination) yields

$$(V'_{Ag} - Ag_{Ag}) + (Ag^{\cdot}_i - V_i) = 0$$

- Now all we have to do is to write down the corresponding Schottky notation and identify the terms in brackets with the Schottky structure elements. We see that

$$(V'_{Ag} - Ag_{Ag}) = |Ag|'$$

$$(Ag^{\cdot}_i - V_i) = Ag^{\cdot}$$

We can generalize this into a "**translation table**":

	A on B site	**A-vacancy**	**A-interstitial**	**AB molecule**	**C on B site**	**Free electron**	**hole**
	All defects neutral					***Always charged***	
Schottky (new) *Building elements*	A|B|	|A|	A	AB	C|B|	e'	$h^{\cdot}$
Kröger-Vink *Structure elements*	A_B	V_A	A_i	AB	C_B	e'	$h^{\cdot}$
Combined structure elements = Building elements	A_B - B_B	V_A- A_A	A_i - V_i	AB	C_B - B_B	e'	$h^{\cdot}$

The General Recipe for Point Defect Reactions and First Example

If you now stick to the full formalism, what you do is:

- **1.** Write down the possible reaction with structure elements (= Kröger-Vink notation).
- **2.** Translate that to building elements (= Schottky notation).
- **3.** Consider charges and check for electroneutrality. If the sum of all charges is not zero, search for mistakes and if there are none, *throw in holes or electrons*, as your system provides.
- **4.** Use the mass action law with the final (building element) equation (including holes or electrons).

First Example: Schottky Defects

We take the Schottky defect as a fitting elementary example and go through the movements:

- **1.** Kröger-Vink *structure element* equation

$$Ag_{Ag} + Cl_{Cl} = V'_{Ag} + V^{\cdot}_{Cl} + AB$$

- After rearranging (remember, the defect comes first!) so we can use the translation table, we have

$$(V'_{Ag} - Ag_{Ag}) + (V^{\cdot}_{Cl} - Cl_{Cl}) + AB = 0$$

- **2.** Switch to *building elements* using the the expressions in brackets; we have

$$|Ag|' + |Cl|^{\cdot} + AB = 0$$

- **3**. *Charge neutrality* demands

$$\Sigma(\text{pos. charge}) = \Sigma(\text{neg. charge})$$

$$[|Ag|'] = [|Cl|^{\cdot}]$$

- **4.** *The mass action* law now gives

$$\frac{[|Ag|'] \cdot [|Cl|^{\cdot}]}{[AB]} = \exp - \frac{\Sigma \nu_i \mu_i^0}{kT}$$

- And this leads to

$$[|Ag|'] = [|Cl|^{\cdot}] = [AB] \cdot \exp - \frac{\Sigma \nu_i \mu_i^0}{kT}$$

- With μ_i = standard chemical potentials of the two vacancies and a lattice molecule, resp., and ν_i = stoichiometric coefficients of the reaction (**1**,**1**, and **– 1** in our case).

This sure looks strange compared to the formula derived in the "physical" way. *But it is the same*. Lets see why.

- First, the activity (or concentration) of the lattice molecule **AB** is simply **[AB] = 1** since it is nothing but the number of mols of **AgCl** molecules in one mol of **AgCl**; i.e. it is **= 1**. This gives us

$$[|Ag|'] \cdot [|Cl|^{\cdot}] = \exp - \frac{\Sigma \nu_i \mu_i^0}{kT}$$

- Now lets look at the energies in the exponent. As always, the energy scale is relative. From whatever zero point you measure your energy to make an **Ag** vacancy or a **Cl** vacancy, you must subtract the energy of the **AB** molecule as measured in that system. If you take it to be zero - which then defines the energy origin of your standard system - the standard chemical potentials of the two vacancies are just the usual formation energies.

Note that as in treatment given before, the mass action law alone does *not* specify the vacancy concentration, *only their product*.

- Only invoking the electroneutrality condition, which demands **[|Ag|'] = [|Cl|** $^{\cdot}$**]**, allows to compute the individual concentrations. Writing H^+ and H^- for the formation enthalpies of the positively or negatively charged vacancy, resp., we obtain the familiar result

$$[|Ag|^{/}] = [|Cl|^{\cdot}] = \exp -\frac{H^{+} + H^{-}}{2kT}$$

Second Example: Frenkel Defects

First we write down the reaction with *structure elements* (= Kröger-Vink notation).

$$A_A + V_i = A_i^{\bullet} + V_A^{/}$$

- After rearranging (remember, the defect comes first!) so we can use the translation table, we have

$$(A_i^{\cdot} - V_i) + (V_A^{/} - A_A) = 0$$

- **2.** The expressions in brackets are the *building elements*, we have

$$|A|^{\cdot} + |A|^{/} = 0$$

- **3**. *Charge neutrality* demands

$$|A|^{\cdot} = |A|^{/}$$

- **4.** *The mass action* law now gives
- *A problem*? What are we going to do with the "**0**"? Well, there really is no problem with the zero - just take the mass action law as it is

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

- Nowhere was it required that in the product there must be terms with negative stoichiometry coefficients ν_i. This gives us

$$[A^{\cdot}] \cdot [|A|^{/}] = \exp -\frac{G_F}{kT}$$

- And we identify G^0 with the formation energy G_F of a Frenkel pair as before.
- Together with the charge neutrality condition we have

$$[A^{\cdot}] = [|A|^{/}] = \exp -\frac{G_F}{2kT}$$

- *almost* the familiar result - *except that we do not have the factor* $(N/N')^{1/2}$.
- OK - we do have a problem, but *not with the zero*. Where did we lose the factor $(N/N')^{1/2}$?

Lets look at equilibrium another way. We do not involve the mass action law but go one step back to the equilibrium condition for the chemical potentials: $\mu_{A^{\bullet}} = \mu_{|A|^{/}}$. We write the chemical potentials in the standard form and obtain

$$\mu_{A\cdot} = \mu^0_{A\cdot} + kT \cdot \ln \frac{n_{A\cdot}}{N'}$$

$$\mu_{|A|/} = \mu^0_{|A|/} + kT \cdot \ln \frac{n_{|A|/}}{N}$$

- For equilibrium we now obtain (if you wonder at the n/N and n/N', consult the link)

$$n_{A\cdot} \cdot n_{|A|/} = N \cdot N' \cdot \exp - \frac{\mu^0_{A\cdot} + \mu^0_{|A|/}}{kT}$$

- Charge neutrality tells us that

$$n_{A\cdot} = n_{|A|/} = n_{FP}$$

- For the concentration of Frenkel pairs $c_{FP} = n_{FP}/N$ we now obtain the correct old formula

$$c_{FP} = \left(\frac{N}{N'}\right)^{1/2} \cdot \exp - \frac{\mu^0_{A\cdot} + \mu^0}{2kT} = \left(\frac{N}{N'}\right)^{1/2} \cdot \exp - \frac{G_{FP}}{2kT}$$

Aha! Applying the mass action law uncritically causes a problem: The standard chemical potentials of vacancies and interstitials were for *different* standard conditions:

- In one case (the *vacancies*) the standard condition was for adding N vacancies to the system, in the other case (the *interstitials*) it was for adding $N' = N \cdot l$ interstitials (and l being some factor taking into account that there are more positions for interstitials than for vacancies in a crystal).
- If that appears to be incredibly complicated and prone to errors - *that's because it is*! But take comfort: You get used to it, and working with it is not all that difficult after overcoming an intial "energy barrier".

Some Remarks to Practical Work

Many books and other texts to not dwell extensively on the fine differences between notations, problems with the standard condition in the chemical potentials, meaning of reaction equations and so on - they write down a reaction equation, in the worst case a mix of Kröger-Vink and Schottky notations, throw in electrons or holes right away to achieve charge neutrality, and write down the mass action law in the form

$$\prod_i (a_i)^{\nu_i} = \exp \frac{-G}{kT} = K(T) = \text{const} \cdot \exp - \frac{G}{kT}$$

- And not much attention is given to the constant $K(T)$ in front of the exponential.
- Even though it's *faulty thermodynamics*, let's see what happens if we do that for Frenkel defects in the Kröger-Vink notation:

The reaction equation was

$$Ag_{Ag} + V_i = V^{/}_{Ag} + Ag^{\cdot}_i$$

$$Ag_{Ag} + V_i - V^{/}_{Ag} - Ag^{\cdot} = 0$$

- The mass action law uncritically applied gives

$$\frac{[Ag_{Ag}] \cdot [V_i]}{[V_i^{/}] \cdot [Ag_i^{\cdot}]} = const \cdot \exp - \frac{G'}{kT}$$

- As long as the defect concentration is small compared to the concentrations of atoms and lattice sites, we may simply equate

$$[Ag_{Ag}] = 1$$

$$[V_i] = 1$$

- Which leaves us with

$$[V_i^{/}] \cdot [Ag_i^{\cdot}] = const \cdot \exp - \frac{G}{kT}$$

- With $G = -G'$, but that is irrelevant - we simply know that the exponential always has a minus sign for the reactions we are interested in and that G must be the formation enthalpy of a Frenkel pair.

That is the correct result, expressed in Kröger-Vink terms. What that means is that you don't have to worry all that much about the finer details as long as you are not terribly interested in the exact value of the constant in front of the exponential - you will mostly get your reaction equation right!

- Luckily, there are only a few fundamental reaction equations involving point defects - everything else can be expressed as linear combinations of the fundamental reactions (like Frenkel and Schottky defect equilibrium) - after some initiation, you will feel quite comfortable with defect reactions.

As shown above for Frenkel defects, it is often advisable not to use the mass action law directly, but to go one step back and use the equilibrium condition for the chemical potentials. This gives not only a clearer view of what constitutes the standard conditions, but also circumvents a number of other problems associated with the law of mass action (if you really want to know, consult the advanced module accessible by the link).

2.4.4 Systematics of Defect Reactions in Ionic Crystals and Brouwer Diagrams

- There are essentially three fundamental situations for defects in ionic crystals.
  - *Intrinsic defects*; i.e. the defects present in the bulk of a crystal in thermodynamic equilibrium. This includes the Frenkel and Schottky defects considered before, but also some other kinds not yet discussed.
  - *Defect Doping*; i.e. the intentional manipulation of defect types and concentrations by the incorporation of specific impurities into the (bulk of the) crystal.
  - *Defect reactions at interfaces*, e.g. the incorporation of atoms or molecules from the "outside" into the crystal via defects - or the opposite, the loss of crystal atoms to the outside world generating defects in the crystal.
- There are not always clear distinctions, but lets look at these three cases separately for a start.

Intrinsic Defects in Ionic Crystals

- This case includes all defect situations that one could find in a perfect ionic crystal. Besides Schottky and Frenkel disorder or any mixture of the two, we could have many more defects - any combination of interstitials and vacancies for any kind of atom in the crystal is admissible.
  - One example: In an **AB** crystal, instead if two oppositely charged vacancies (the Schottky defects), we could also have two interstitials of the two kinds of atoms carrying different charge.
  - This kind of defect is called an "**Anti Schottky defect**", it would be formed according to the (Kröger-Vink) reaction

$$AB + 2V_i = A_i^{\cdot} + B_i^{/}$$

  - Following our recipe, we obtain after rearranging and "translating" to Schottky notation:

$$(A_i^{\cdot} - V_i) + (B_i^{/} - V_i) = AB$$
$$A^{\cdot} + B^{/} = AB$$

- This is essentially the same result as for the regular Schottky defects. However, in using the mass action law, we would have to use the formation energies for interstitials (and take care of the additional degrees of freedom for arranging interstitials as in the case of Frenkel defects).
  - In reality, the formation energies of interstitials are mostly larger than those for vacancies; this is certainly true for the "big" negatively charged anion interstitial.
  - Anti Schottky defects therefore have not been observed as the dominating defect type so far. But they are not only perfectly feasible, but also always present - only their concentration is so small that it can not be measured; consequently they do not play a role in anything interesting connected with defects.
- We could also conceive of "**anti site defects**", i.e. **A**-atoms on **B**-places and vice versa; A_B and B_A, of combinations like a V_A and A_B, and of plenty more intrinsic defects for more complicated crystals, e.g. for $YBa_2Cu_3O_7$ (the famous first high temperature superconductor with a critical temperature larger than the boiling point of liquid N_2).
  - We could, moreover, include isoelectronic impurity atoms into the list; e.g. **K** instead of **Na**; **Ba** instead of **Ca**, or **F** instead of **Cl**, which could be incorporated into a crystal without the need to change anything else. A little dirt, after all, is always "intrinsic", too.
  - And *all* the reactions that are conceivable will *occur*. The only difference is that some might be frequent and some might be rare - and it is often sufficient to only consider the dominating reaction.
- This teaches us a *major lesson*, especially with respect to the upcoming paragraphs
  - There are far too many defect types and reactions principally possible in simple (ionic) crystals (not to mention complicated ones) for a priori treatments of all possible effects. We must invoke some *additional* information (such as anion interstitials being unlikely) to simplify the situation to a level where it can be handled.
  - For the intrinsic defects mentioned so far, this was easy and has been done all along. It will become an important guiding principle for the other two cases, however.
- But we are not yet done with intrinsic defects: If we look at *semiconducting* ionic or compound crystals, we may have to include *electrons and holes* in our defect systematization. Let's look at this, first assuming that electrons and holes are the *only* defects.
  - Implying a band structure and always using the Boltzmann approximation for the tail of the proper Fermi distribution, we might denote the generation of an electron in the conduction band in complete analogy to the Kröger-Vink system.by

$$e_V^{/} + h_C^{\cdot} = e_C^{/} + h_V^{\cdot}$$

- Rearranging gives

$$(e_C^{/} - h_C^{\cdot}) + (h_V^{\cdot} - e_V^{/}) = 0$$

- The obvious translation to Schottky notation yields

$$e^{/} + h^{\cdot} = 0$$

- Adding electroneutrality, i.e. $e^{/} = h^{\cdot}$, makes the analogy to Frenkel defects complete, and we obtain

$$e^{/} = h^{\cdot} = c \cdot \exp - \frac{E}{kT}$$

There is a big difference, however. We know that the constant c in front of the exponential is the effective density of states N_{eff} (or more precisely $[N^C_{eff} \cdot N^V_{eff}]^{½}$) and the formation energy is given by $E = E_g/2$, with E_g = band gap.

- This is knowledge that comes from quantum theory and there is simply *no way* to deduce this from classical thermodynamics.

However, the mass action law is based on considering the minimum of the free enthalpy - a principle that is always valid. It is only the chemical potential of electrons and holes that cannot be directly expressed in the standard form. Mass action, however, remains valid.

Defects and Doping

Lets now consider some typical doping reactions. Most common and important is the doping of semiconductors with substitutional impurities, i.e. **P**, **As**, or **B** in **Si**.

- If we call the substitutional dopant atoms **D** (for **D**onor) or **A** (for **A**cceptor), we may express the doping reaction, i.e. the exchange of electrons or holes with the bands, as follows

$$D = D^{\cdot} + e^{/}$$

$$A = A^{/} + h^{\cdot}$$

- Kröger-Vink and Schottky notation are identical in this case (figure it out!), and we have the mass action law

$$\frac{[D]}{([D^{\cdot}] \cdot [e^{/}])} = k_D$$

$$[e^{/}] = \frac{[D]}{[D^{\cdot}]} \cdot k_D^{-1}$$

- Lets consider a simple situation with just one type of doping, say donors, with a concentration **[D]**, giving us $[e^{/}] = k_D^{-1}[D]/[D^{\cdot}]$ as stated above. But we have no holes so far. We need some other reaction to produce holes, what comes to mind is

$$h^{\cdot} + D = D^{\cdot}$$

$$\Rightarrow \frac{[h^{\cdot}] \cdot [D]}{[D^{\cdot}]} = k_A$$

- Can we get the "universal" mass action law for semiconductors, as we know it from semiconductor physics from this, i.e.

$$[e^{/}] \cdot [h^{\cdot}] = n_i^2$$

  - If we form the product with the equation from above we obtain

$$[h^{\cdot}] \cdot [e^{/}] = k_D^{-1} \cdot k_A$$

  - We have the mass action law as we know it with, however, unspecified constant. In order to obtain the absolute concentrations, we need one more equation, which, in the absence of other charged defects, is supplied by the electroneutrality condition

$$[h^{\cdot}] + [D^{\cdot}] = [e^{/}]$$

- OK, we are on safe grounds again. But there seems to be a certain ambiguity. Instead of the reaction $h^{\cdot} + D = D^{\bullet}$, we also could have chosen the "normal" intrinsic reaction $e^{/} + h^{\bullet} = 0$ from above.
  - So what is it? *Both*, of course. The equations above are dominating at low temperatures where thermal carrier generation can be neglected, i.e. not too high temperatures, the other one dominates at high temperatures.
  - But *both* occur independent of each other and, since there is only *one* equilibrium value for the respective concentrations, *both* must give the same numerical values for the same quantity if evaluated.
- This teaches us an important lesson for the treatment of defect equilibria:
  - Since the mass action law and the electroneutrality condition supply only *two* equations for possibly more than two unknown defect types, *any* sensible reaction equation that comes to mind and contains the unknown quantities can be used to supply the required additional information! The equilibrium concentration of defect type **i** is always the same - no matter in which equation it comes up!
- As we see, even for pure semiconductors it is possible to describe the electron-hole equilibrium in terms of reaction equations.
  - But the notion of chemical potentials becomes somewhat strained for calculating concentrations or "activities". Considering densities of states and distribution functions (Fermi distribution in full generality or Boltzmann distribution in the proper approximation) may be more advantageous as long as only electrons and holes are considered.
- Now lets look at a different kind of doping: We intentionally change the vacancy concentration, i.e. we **dope a crystal with vacancies**.
  - In a simple example we may look back at the introductory paragraph of this subchapter, and consider a reaction equation for the incorporation of **Ca** into a **NaCl** crystal (in Schottky notation right away - can you figure out the Kroeger-Vink notation?)

$$Ca^{\cdot}_{Na} + V^{/}_{Na} = 0$$

  - In words: A **Ca-V_{Na}** pair is introduced (or taken out) of the crystal.
  - The mass action law demands that $[Ca^{\cdot}_{Na}] \cdot [V^{/}_{Na}] =$ constant; and charge neutrality is conserved if $[V^{/}_{Na}] = [Ca^{\bullet}_{Na}]$.
  - If, and *only* if there is no other way to achieve charge neutrality, e.g. by generating electrons or holes, we will now produce *vacancies* by incorporating a doping element in perfect analogy to producing electrons or holes in semiconductors.
- Of course, we could also have incorporated **Ca** by generating $Ca_i^{/}$ interstitials, a mix of vacancies and interstitials, or even worse, a mix of vacancies, interstitials, holes and electrons. All those possible reactions *will* occur and we cannot know a priori what will dominate. In a real case we must use some additional knowledge as pointed out above if we want to get results.

- Well, "we" do know that vacancy doping can indeed be achieved in this way in **ZrO_2** doped with e.g. **Y_2O_3** or **CaO**, generating **$V_O^{\bullet\bullet}$**, i.e. doubly positively charged oxygen vacancies. This is a particularly relevant example, because it is part of the working principle of the oxygen sensor in your car exhaust system that feeds the controller of the car engine in order to keep emissions at the lowest possible level.
- The technical importance is the same as in semiconductors: Whatever the intrinsic defect concentrations might be in the perfect intrinsic material, with doping you have a more or less fixed concentration of vacancies that can be far larger than the intrinsic concentration and may not depend sensitively on temperature.

Great if you need lots of vacancies because you want to make an ionic conductor where the conductivity depends on the diffusion of oxygen via a vacancy mechanism. However, vacancy doping is not such a hot issue at low temperatures (like room temperature) if the vacancies - and therefore the oxygen ions, too - are not mobile at reasonable temperatures - in contrast to electrons and holes which get rather more mobile with decreasing temperature.

- But if there is *some* mobility, you have now increased the *ionic* conductivity by orders of magnitude - exactly as you increase the *electronic* conductivity in semiconductors by doping.
- Moreover, you *know* the concentration of the vacancies, and within some parameter range, you can treat it as *constant* which means you can remove **[V]** from the business end of the mass action law and multiply it into the general reaction constant. Life is easier.

There are more examples for technical uses of doping, but we will now consider the *third* basic reactions, the defect reactions at interfaces.

Defect Reactions at Interfaces

Considering that *all* chemical reactions between a solid and anything else occur at the surface of the solid, i.e. at the solid-gas, solid-liquid, or solid- solid interface, this headline covers a good part of general chemistry.

- Indeed, it has become clear in recent years that reactive solid-solid interfaces generate or consume point defects. However, here we will only look at reactions between a gas and a (simple) ionic crystal as they are used in sensor technology. In other words, we consider the possible reactions between a **MeX** crystal and a **X_2(g)** gas in a first simplified treatment. We do not consider charges for the time being, to keep things simple.
- The crystal may then incorporate **X_2(g)** (or emit it) via several reactions which we can easily formulate with the the Kröger-Vink notation:

$$\text{1.} \qquad \tfrac{1}{2}X_2(g) + V_X = X_X$$

- i.e. an atom of the gas occupies a fitting empty place (= vacancy) of the crystal.

$$\text{2.} \qquad \tfrac{1}{2}X_2(g) + V_i = X_i$$

- i.e. an atom of the gas occupies a fitting empty place (= vacancy) in the interstitial lattice and is now an interstitial.

$$\text{3.} \qquad \tfrac{1}{2}X_2(g) + V_M = X_M$$

- i.e. an atom of the gas occupies a (probably not well fitting) empty place (= vacancy) in the metal ion lattice and is now an anti-site defect.
- And there will be even more possibilities as is shown below.

The thing to note once more is: If these reactions *can* occur, they *will* occur - independently of each other. Only their probabilities (or reaction rates) are (wildly) different, and we are well off if we know (or can make an educated guess) at the dominating reaction. And the equilibrium concentrations **$[D_i]$** of some defect type **D_i**, no matter in which equation it appears, are always identical in equilibrium.

- In other words, we should know

1. What kind of defect situation dominates in our **MeX** crystal (Frenkel- or Schottky defects etc.); i.e. which reaction constant is smallest?
2. How can charge neutrality be achieved (only with ions and defects; only with electrons and holes, or in a mixture)? In other words are we dealing with an ionic conductor, a semiconductor, or a mixed case? Of course, the answer to this question may well depend on the temperature.
3. Is there intentional (or unintentional) ionic or electronic doping that imposes specific conditions on the defect situation?

- This only looks hopeless, but rejoice, it is not - albeit for a sad reason: After all, we are not so much interested in defects per se, but in their uses. Typically, we want to do something "electrical" - make a better battery, a fuel cell, a sensor giving of a voltage or a current in response to the stuff to be sensed - and this demands that we use only ionic crystals that are **ionic conductors** of some sort.
  - Unfortunately, not too many ionic crystal are useful ionic conductors; it's just a handful of crystal families. And only those families we have to know. The search for a real good (and affordable) ionic conductor at room temperature is still on - if you find it, you may not make the Nobel prize, but certainly a great deal of money.
  - So all it takes is to study some **4** or **5** typical cases which contain practically everything encountered in ionic defect engineering.
- The paradigmatic case is an undoped crystal with Frenkel defects and some semiconducting properties. We start assuming that the electron/hole concentration is far smaller than the Frenkel defect concentration.
  - Thermal equilibrium of the **MeX** crystal by itself thus means that we have

$$[M^{\cdot}_i] = [V^{/}_M]$$

$$[e'] = [h^{\cdot}] << [V_M]$$

- Now we establish equilibrium with the gas $\mathbf{X_2(g)}$. It could be $\mathbf{H_2}$, $\mathbf{O_2}$, $\mathbf{Cl_2}$, $\mathbf{F_2}$, whatever.
  - How are **X** atoms to be incorporated? Surprisingly, perhaps, none of the three possibilities given above is the preferred reaction. We do not have **X**- vacancies ($\mathbf{V_X}$) available in our case, and we are not going to generate very unlikely **X** interstitials ($\mathbf{X_i'}$) or anti site defects. We want to incorporate **X** on a regular lattice site.
  - Now you see why the Kröger-Vink notation is useful. Playing around a bit with what you have (and putting in charge neutrality right away), gives

$$\tfrac{1}{2}\, X_2(g) + M_M = MX + V^{/}_M + h^{\cdot}$$

  - In words: An **X**-atom takes out a **M** atom from its position on the crystal surface, forming a **MX** molecule that is added to the crystal somewhere, leaving back a vacancy on a **M**-site and a hole.
  - This is not so easily expressible in Schottky notation (try it), but leads easily to the mass action law noticing that $\mathbf{[M_M] = [MX] =}$ const $\mathbf{= 1}$ and thus not needed in the "business end" of the mass action law.
  - We could come up with other reaction equations achieving the same result; but the one we have is good enough for the time being.
- We thus obtain two "master" equations for this case, one from mass action and one from charge neutrality exactly along the lines discussed before.
  - Mass action law

$$\frac{[X_2(g)]^{1/2}}{[V^{/}_M]\cdot[h^{\cdot}]} = \text{const.}$$

  - Charge neutrality

$$[V^{/}_M] + [e^{/}] = [M_i^{\cdot}] + [h^{\cdot}]$$

  - We have *two* equations for the *four* unknowns $\mathbf{V^{/}}$, $\mathbf{e^{/}}$, $\mathbf{M_i^{\cdot}}$, and $\mathbf{h^{\cdot}}$ which determine the electrical conductivity σ as a function of the concentration (or partial pressure) of the gas $\mathbf{[X_2(g)]}$ via

$$\sigma([X_2(g)]) = \Sigma_i\,(q_i \cdot c_i \cdot \mu_i)$$

  - With q = charge carried by the defect **i**, c_i = concentration and μ_i = mobility of defect **i**, resp.
- Again, we need to have additional information about our system if we want quantitative relations between a measurable parameter like the defect-dependent conductivity.
  - Doping, as described above, could be helpful. It would provide a more or less constant value for e.g. the vacancy or electron concentration and thus remove one (or two) unknowns.
  - Without that, however, we have to resort to case studies making reasonable assumptions and considering the important quantities for the task at hand. As an example, if we want to measure the $\mathbf{[X_2(g)]}$ concentration, we are not so much interested in the absolute value of σ, but in its change with the gas concentration, $\mathbf{d\sigma/d[X_2(g)]}$.

- That implies that we are mainly interested in those defects which react sensitively to concentration changes of $[X_2(g)]$.

For our example we postulated the two conditions

$$[M^{\cdot}_i] = [V^{/}_M]$$

$$[e^{/}] = [h^{\cdot}] << [V_M^{/}]$$

- This is valid as long as we are considering stoichiometric **MX** which is neither losing nor adding **X**. In other words, the crystal is kept at the **stoichiometric point** - at a certain partial pressure of X_2.
- It is now important to notice that the reaction with the outside gas at partial pressures different from that belonging to the stoichiometric point changes the stoichiometry - no matter how you look at it. For ambient (or standard) pressure, there is no reaction and the stoichiometry is perfect - we are at the stoichiometric point. For large partial pressures of X_2, we will produce $MX_{1+\delta}$, for low pressures $MX_{1-\delta}$.

Now lets see what happens if we work around the stoichiometric point. The absolute concentration of the vacancies and interstitials may change a little, and this means that the electron and hole concentration has to change exactly the same amount in order to maintain charge neutrality.

- However, since we assumed that the absolute concentration of the electrons and holes at the stoichiometric point is much smaller than that of the vacancies and interstitials, the *relative change* of $[e^{/}]$ and $[h^{\cdot}]$ is much larger than that of $[M_i^{\cdot}]$ and $[V^{/}_M]$.
- Accordingly, we may assume that $[M_i^{\cdot}]$ and $[V^{/}_M] \approx$ **constant** around the stoichiometric point, i.e. within a certain range of partial pressures of X_2 below and above the standards pressures which simplifies the relevant equations to

$$\frac{[X_2(g)]^{1/2}}{[h^{\cdot}]} = \text{const.}$$

- With a different value of the constant, however.
- The concentration of holes and electrons, on the other hand, changes markedly, but their *absolute* concentrations are still much smaller than that of vacancies and interstitials.

This leads us to an *extremely important generalization*:

- As far as the *mass action law* is concerned, only *variable* concentrations, i.e. concentrations that are not (approximately) constant, count. The absolute concentration is of no special importance - it becomes part of the reaction constant.
- As far as *charge neutrality* is concerned, only *absolute* concentrations count. Minority carriers can simply be neglected in a first approximation.

This give a first direct result: We can write dow the following simplified mass action law and electroneutrality condition:

Mass action	$[h^{\cdot}] = \text{const} \cdot [X_2(g)]^{1/2}$
Electro-neutrality	$[V^{/}_M] = [M_i^{\cdot}]$

How about the electron concentration? Since our approximations imply that there is no interaction between the point defects and the electrons and holes, we must have $[h^{\cdot}] \cdot [e'] =$ const. and thus

$$[e^{/}] = \text{const} \cdot [X_2(g)]^{-1/2}$$

- But we will derive that result now by using different reaction equation just to show that in equilibrium you always must obtain the same results.

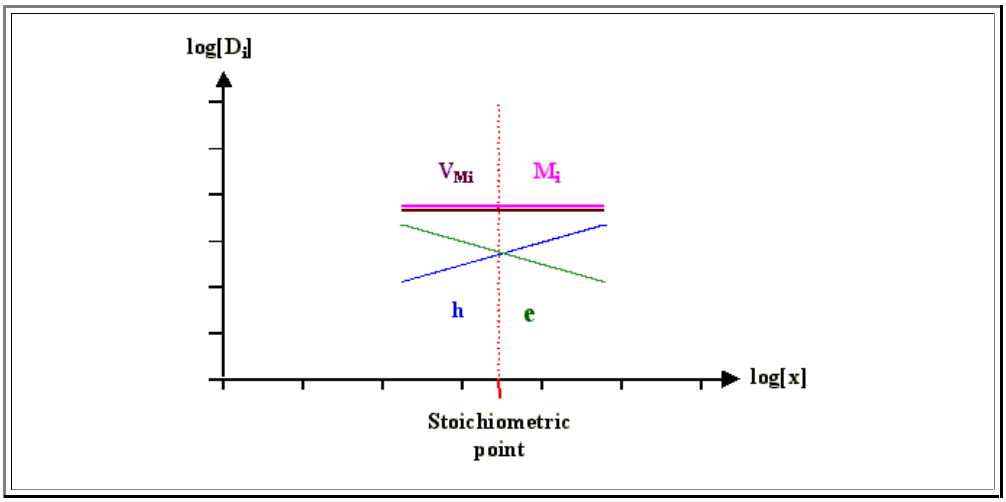
Lets consider the reaction

$$\tfrac{1}{2}\, X_2(g) + M^{\cdot}_i + e^{/} = MX + V_i$$

- In words: A positively charged metal interstitial plus an electron and a gas atom form a crystal molecule and a vacancy on the interstitial lattice.
- Neglecting $M_i^{\cdot}$, **MX**, and V_i with the same arguments as before, we have, as we know that it must be

$$[e^{/}] = \text{const} \cdot [X_2(g)]^{-1/2}$$

In a **log [i] - log [X_2(g)]** plot we have straight lines with a slope of **1/2** for holes, **– 1/2** for electrons and **0** for the interstitials and vacancies, respectively. This looks like this:



- Without looking at the reaction constants, we know that the cross over of the $e^{/}$ and $h^{\cdot}$ concentration lines must be at the stoichiometric point.
- It is clear that for large deviations from the stoichiometric point the approximations used are no longer valid. For very small or very large partial pressures of X_2, we now may consider the other two possible extremes by simply extrapolating the lines in the illustration:
- **1.** For very large partial pressures of X_2 the $e^{/}$ concentration becomes negligible while the hole concentration becomes comparable to the point defect concentration. Charge neutrality can only be maintained by decreasing the positively charged metal interstitials and increasing the negatively charged vacancies. In the extreme, we may only consider $[V^{/}_M] = [h^{\cdot}]$ for charge neutrality. Inserting that in the reaction equation from above, we have .

$$\tfrac{1}{2}\, X_2(g) + M_M = MX + 2h^{\cdot}$$

- which gives the mass action equation

$$\frac{[X]^{1/2}}{[h^{\cdot}]^2} = \text{const}$$

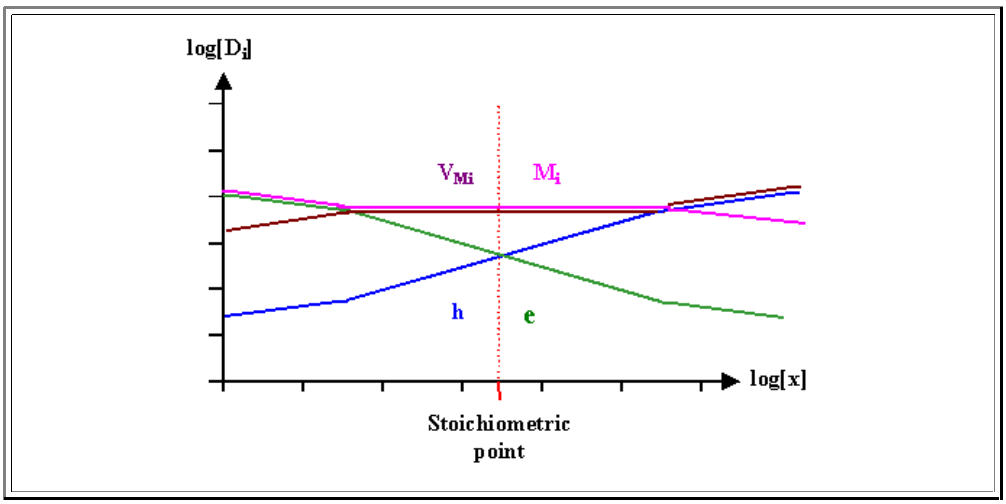
$$[h^{\cdot}] = \text{const.} \cdot [X]^{1/4}$$

- For very low partial pressures we obtain exactly along the same line of arguments $[e^{/}] = \text{const.} \cdot [X]^{-1/4}$.

With these relations, we may also calculate the concentrations of the minority point defects by simply inserting the above equations in the appropriate reaction equation and applying mass action which yields

High pressure side	$[M_i^{\cdot}] = \text{const} \cdot [X]^{-1/4}$
Low pressure side	$[V^{/}_M] = \text{const} \cdot [X]^{1/4}$

- Putting everything together in on single graph, we obtain a schematic **Kröger-Vink** or **Brouwer diagram**:



- Of course, the change-over would be smooth in reality; and we cannot tell easily where it will occur. It is also obvious that there are no discontinuities of the concentration curves, which tells us something about the "const." in the mass action equation.
- In the consideration above we did not assign values to the "const." and carry it through. That might be a interesting exercise one of those days.

In any case, we now have seen how Kröger-Vink reaction equations, mass action, charge neutrality and some additional knowledge or educated guesses allow to come up with a pretty good idea of what will happen in a reaction involving point defects.

- The possibilities of electronic and ionic doping together with the temperature dependence of point defect equilibria now give us a powerful tool for designing materials for specific applications.
- "**Ionics**", in research and application, is slowly coming into its own. Together with the good old "electronics" it may well open up new fields for materials scientists.

3. Point Defects and Diffusion

3.1 Diffusion Primer

3.1.1 Diffusion and Point Defects

3.1.2 Recapitulation of Ficks Laws and Random Walk

3.1.3 Coupling Phenomenological Laws to Single Atomic Jumps

3.1.2 Essentials to Chapter 3.1: Diffusion Primer

3.2 Diffusion Mechanism

3.2.1 Atomic Mechanisms

3.2.2 Self-Diffusion

3.2.3 Diffusion of Impurity Atoms

3.2.4 Essentials to Chapter 3.2 Diffusion Mechanisms

3.3 Experimental Approaches to Diffusion Phenomena

3.3.1 Determination of Diffusion Profiles

3.3.2 Essentials to Chapter 3.3 Experimental Approaches to Diffusion Phenomena

3. Point Defects and Diffusion

3.1 Diffusion Primer

3.1.1 Diffusion and Point Defects

- Point defects generally are mobile - at least at high temperatures. They are the vehicles that make the atoms of the crystal mobile - point defects are the cause of **solid state diffusion**.
- Many products of modern technology depend on solid state diffusion and thus on point defects. Some examples are:
  - **Microelectronics** and **Optoelectronics**.
  - Solid state *Sensors*, e.g. the oxygen sensor regulating the emissions of your car.
  - Solid state *batteries*,*accumulators* and *fuel cells*.
  - *High strength materials*.
- The concentration of point defects, their specific kind (including impurity atoms), their migration parameters, equilibrium or non-equilibrium conditions and the atomic mechanisms of diffusion determine what you get in a specific solid state experiment involving diffusion.
  - Small wonder that many diffusion phenomena are not yet totally clear!
- If we use the term "diffusion", we always and exclusively mean that *something* moves around in a random fashion,. i.e. does a **random walk**.
  - The "*something*" in the context of this lecture is an *atom* moving around randomly in a crystal. Generally, it can be *any* particle or even quasi-particle we like - atoms, molecules, electrons, positrons, photons, phonond, excitons, ...

3.1.2 Recapitulation of Ficks Laws and Random Walk

- Lets quickly go over the basic laws of diffusion which were discovered by Adolf **Fick** on a phenomenological base long before point defects were known. The starting point is **Ficks 1. Law**, stating:
  - The **flux** $\boldsymbol{j}$ of diffusing *particles* (not necessarily atoms) is proportional to the *gradient* of their *concentration*, or

$$j_i = -D \cdot \nabla c_i$$

  - The index **i** refers to the particular particle with number **i** observed; $\boldsymbol{D}$ is the diffusion coefficient of that particle.
- Note that even this purely phenomenological description applies to everything - e.g. liquids - as long as we discuss *diffusion* and not, e.g., some kind of flow.
  - This means that the underlying dynamics of the particles on an atomic scale is essentially random walk.
  - The derivation of the simple continuum equation above from the primary events of random scattering (causing random walk) of many discrete particles takes a lot of averaging. If you don't know how it's done (or forgot), do consult the proper (english) modul of "Introduction to Materials Science II" (and the links from this modul).
- If there are several interacting particles, the formulation of Ficks **1.** law must be more general, we have

$$j_i = -M \cdot \nabla \mu_i$$

  - With $\boldsymbol{\mu}$ = chem. potential; $\boldsymbol{M}$ = mechanical mobility.
  - Since the gradient of the chemical potential may be different from zero even for constant concentrations, special effects as, e.g., **uphill diffusion** are contained within this formulation.
- The next basic equation is the **continuity equation**. It states:
  - *Changes* of the particle concentration within a volume element must express the *difference* of what goes in to what goes out - we have *conservation of the particle number* here. In mathematical terms this means

$$\frac{\partial c}{\partial t} = -\operatorname{div} j$$

- This is if course only true as long as no particles are generated or annihilated (as, e.g., in the case of electrons or holes in an illuminated semiconductor).

- Combining the two equations from above we obtain **Ficks 2. law**:
  - The *temporal change* in concentration at a given point is proportional to the ***2nd*** *derivative of the concentration*, or

$$\frac{\partial c}{\partial t} = \mathrm{div}\,(D \cdot \nabla c) = D \cdot \Delta c$$

  - With the final equation being valid only for D **= const.**

- Ficks equations look innocent enough, but solutions of the rather simple differential equations forming Ficks laws are, in general not all that simple! They do follow some general rules, however:
  - They involve almost always statistical functions, as well they should, considering that diffusion is a totally statistical process at the atomic level.
  - The solutions to heat conducting problems are quite similar, as well they should, because the conduction of heat can be treated as a diffusion phenomena. (it actually *is* a diffusion phenomena).
  - This link gives some more information about Fick's laws and standard solutions.

- Many diffusion phenomena can be dealt with on the phenomenological base of Ficks laws. All that is required, is to know the diffusion coefficient and its dependence on temperature and possibly other variables - you do not have to know anything about the *atomic mechanisms* involving point defects to solve diffusion problems.
  - It turns out, however, that complex diffusion problems - e.g. the simultaneous diffusion of **B** and **P** in **Si** can not be modeled adequately without knowing the atomic mechanisms and their interaction. This explains the impetus behind major efforts to unravel the precise mechanisms of diffusion in **Si** and other semiconductors.

- All we have to know about **random walk** is the general relation between the average distance $\langle r^2 \rangle$ covered by a randomly "walking" object (which we often also call **diffusion length** L) , the number N of steps made, and the (average) distance r_0 covered in one step.
  - For three-dimensional random walk we have quite generally

$$\langle r^2 \rangle = L^2 = r_0^2 \cdot 3N$$

3.1.3 Coupling Phenomenological Laws to Single Atomic Jumps

- We now must link the phenomenological description of diffusion (that only works on averages and thus only if many particles are considered) with the basic diffusion event, the single jump of a single atom or defect.
- We describe the *net flux* of particles as the difference in the number of particle jumps to the left and to the right. With the jump frequency ν we obtain Ficks **1.** law with an expression for the diffusion coefficient (for cubic crystals), a detailed derivation is given in the link.

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

  - With a = lattice constant, ν = jump frequency, i.e. the number of *jumps per second* from one position to a neighboring one .
  - g is the **geometry factor** of the lattice type considered. It takes into account that considering all jumps that are possible in the given lattice, only some have a component in the x-direction. Its definition is

$$g = \tfrac{1}{2} \cdot \Sigma_i \frac{\Delta x_i}{a}$$

  - g is always about **1** as you will find out doing the exercise, so we will not consider it any more.

- The jump frequency $\nu = N/t$ (= number of jumps N per second) is given by

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

- with G_m = free enthalpy for the jump or for the migration of the atom or defect. ν_0 is the frequency of "attempts" to overcome the enthalpy barrier for a jump; it is, of course, the vibration frequency of the lattice atoms, i.e. around 10^{13} Hz.

This gives us the second important parameter set describing a property of a point defect, namely its **migration energy** and **entropy**

- All we have to do is to express $G_m = H_M - TS_M$, with H_M = migration enthalpy (or -energy), and S_M = migration entropy.
- The magnitude of the migration entropy will be comparable to the formation entropy because it has the same roots. It is thus around **1 k** for "normal" crystals.

Combining everything, we obtain an expression for the diffusion coefficient D in terms of the migration energy:

$$D = D_0 \cdot \exp - \frac{H_M}{kT}$$

- Where all constant (or nearly constant) factors have been included in D_0. Some numerical values are given in the link.

These formulas relate the atomic properties of defects to the diffusion coefficient from Ficks laws.

There is one more expression of prime importance when it comes to diffusion. It brings together statistical considerations from looking at random walk (which is exactly what a vacancy does) as given above with the diffusion coefficient. All we have to do is to express N by the diffusion coefficient.

- What we get is the famous **Einstein** - **Smoluchowski** relation (for **3**-dimensional diffusion).

$$D = \frac{L^2}{6\tau}$$

- With L = mean square displacement or **diffusion length**, τ = time since start of the diffusion (or, if the particle "dies", e.g. by recombination in the case of minority carriers in semiconductors, its **lifetime**).

Einstein derived this in **1905** in a slightly more general form:

$$D = \frac{\langle r^2 \rangle}{g^* \cdot \tau}$$

- With r = vector between "start" and "stop" of the diffusing particle for the time τ; $\langle r^2 \rangle$ is thus the average of the square of the mean displacement (this is something different from the square of the average!), and **g*** is some factor "in the order of **1**", i.e. **2**, or **6**, depending if the diffusion is **1 -**, **2 -** or **3 -**dimensional and what kind of symmetry (cubic, etc.) is involved.

Now let's do an exercises:

Exercise 3.1-1
Calculate geometry factors

Exercise 3.3-1
Quick Questions to 3.

3.1.2 Essentials to Chapter 3.1: Diffusion Primer

There is no technology without diffusion and no "*high*" technology without *controlled* diffusion.

– Diffusion =

Fick's first law is the foundation of phenomenological diffusion.

$$1. \quad j_i = -D \cdot \nabla c_i$$

$$2. \quad \frac{\partial c}{\partial t} = \text{div}\,(D \cdot \nabla c) = D \cdot \Delta c$$

- Fick's second law is simply the continuity equation for diffusing entities (without changing the total particle number).

- Diffusion is synonymous with "random walk". The basic equation for random walk relates the diffusion length L to the number of jumps N and the (average) distance a covered in one jump.

$$L^2 = a^2 \cdot 3N$$

The relation between the atomic point of view and the phenomenological point of view goes back to Einstein; ν is the jump frequency N/t.

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

$$= D_0 \cdot \exp - \frac{H_M}{kT}$$

- The important parameter for atomic diffusion is now the migrations enthalpy H_M of the atom (or better defect) under consideration, and, somewhat less important, the pre-exponential factor D_0 that contains the migration entropy S_M and the lattice parameters.

If we combine the equations for D with the one for random walk, we obtain the Einstein-Smolukowski relation

$$D = \frac{L^2}{6t}$$

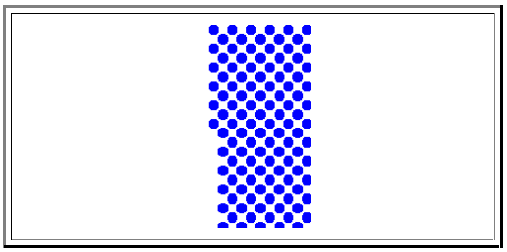
- Read backwards it tells us that the diffusion lengtt L is given by the square root of diffusion coefficient D times diffusion time t.

3.2 Diffusion Mechanism

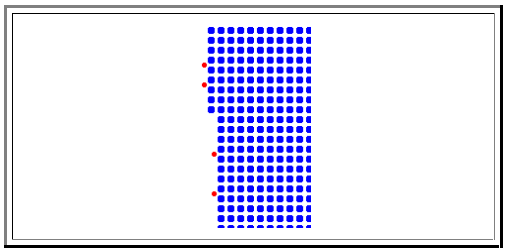
3.2.1 Atomic Mechanisms

There are several atomic mechanisms that lead to the movement of atoms. By far the most prominent are the vacancy mechanism and the direct interstitial mechanism. How they work can be seen in the animations:

Simulation of the vacancy mechanism



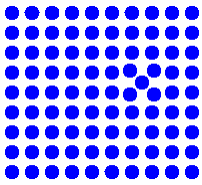
Simulation of the direct interstitial mechanism



- Note a fundamental difference! If you consider the diffusion of a *particular* atom (any blue one of your choice for the vacancy mechanism, any one of the red ones for the interstitial mechanism), your selected atom *always* moved a bit in the second case, but *may not have done anything in the first case*. The diffusion of a particular lattice atom by a vacancy mechanism, while inextricably linked to the movements of vacancies, is not the same as vacancy diffusion, but something different!
- In other words: if a vacancy has made ***N*** jumps by moving around in the lattice, ***N*** atoms will also have made a jump. However, not necessarily ***N*** *different* atoms, because some *individual* atoms may have made more than **1** jump. If we look at any *particular* atom, there is no way of telling if ti has made a jump or not. At best we can give some probability.
- This leads to a major conclusion: While the diffusion of a *particular* lattice atom by a vacancy mechanism is inextricably linked to the movements of *many vacancies*, its specific movement is principally different from the movement of a single vacancy.

Other mechanisms which are quite rare but nonetheless potentially important in semiconductors are:

Indirect interstitial mechanism
for self-interstitials

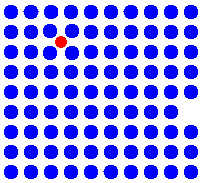


- The simulation shows the elementary step: A self-interstitial (shown in light blue for easier identification) pushes a lattice atom into the interstitial lattice. The net effect is the migration of an self-interstitial from one interstitial site to an different one.
- The mechanism is totally different from the regular interstitial mechanism. If in a thought experiment you mark a *specific* self-interstitial atom (paint it red), it will move a lot with the direct interstitial mechanism, but hardly at all with the indirect one.

The **"kick-out" mechanism**
for impurity atoms

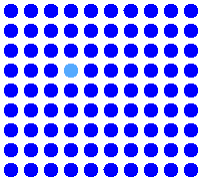
- Interstitial impurity atoms move rather fast by a direct interstitial mechanism, until they eventually displace a lattice atom. This is shown in the simulation. We now have a self-interstitial (that may or may not be very mobile) and a rather immobile substitutional impurity atom, which may now diffuse with one of the other (slow) mechanisms.
- The total effect of the diffusion now is caused by the superpositon of *two* (usually very different) mechanisms. **Au** in **Si**, and possibly some other impurities, diffuse in this fashion.

The **Frank-Turnbull mechanisms** (or **dissociative mechanism**).



- This is the pendant to the kick-out mechanism. Here the diffusing impurity atom does not dislodge a lattice atom, but gets trapped in a vacancy, whereupon it is almost immobile. The total effect may be quite similar to the kick-out mechanism.
- Which mechanism - Frank-Turnbull or kick-out - is operative, is difficult to find out. We must expect that in materials containing predominantly vacancies, the Frank-Turnbull mechanism will occur for some impurities, while the kick-out mechanism may be operative in materials with sizeable concentrations of interstitials.

Various direct diffusion mechanisms



- Shown is *one* variant, a direct exchange of places between two atoms. Other variants are exchanges involving more that **2** atoms (a whole "ring" that "rotates").
- Direct mechanisms are every now and then suggested in the literature to account for some new diffusion phenomena, but so far do not seem to occur in crystals.
- They may, however, play a role (in analogous form) when considering diffusion in amorphous materials.

- Then we *could* have:
  - The **"Extended interstitial" mechanism**
    This is a possibility not yet discussed or observed. It is mentioned just to show that there might be more atomic mechanisms than have been discovered so far. Imagine an extended interstitial moving through a crystal. The **10** or so atoms "inside" the extended interstitial move around a bit while the interstitial passes through and may end up on lattice places different from the ones where they were - they have moved! It is totally unknown if this effect plays a role in **Si**, but it well might occur at high temperatures.
  - And, maybe, there could be more?
- Again, it is important to keep in mind that you must clearly keep apart the movement of the "vehicle" - the vacancy, interstitial, etc. - and the movement of the atom(s) whose diffusion is of interest to you!

Exercise 3.3-1

Quick Questions to 3.

3.2.2 Self-Diffusion

- We ask ourselves how the regular atoms of a crystal diffuse. In the case of crystals with two or more different atoms, we have to answer this question for each kind of atom separately.
- The answer is easiest for a simple (mono)vacancy mechanism in simple elemental cubic crystals. The **self-diffusion coefficient** is given by $g \cdot a^2$ times the number of jumps per sec that the diffusing particles make.
  - Since only lattice atoms that have a vacancy as a neighbor can jump, or, in other words, the number of lattice atoms jumping per sec is identical to the number of vacancies jumping per sec, we obtain for the diffusion coefficient of self-diffusion by a simple vacancy mechanism the following equations:

$$D_{SD} = c_V \cdot D_V$$

$$D_{SD} = g \cdot a^2 \cdot n_0 \cdot \exp -\frac{G_m}{kT} \cdot \exp -\frac{G_F}{kT}$$

$$D_{SD} = g \cdot a^2 \cdot \nu_0 \cdot \exp \frac{S_M}{k} \cdot \exp -\frac{H_m}{kT} \cdot \exp \frac{S_F}{k} \cdot \exp -\frac{H_F}{kT}$$

$$D_{SD} := D^* \cdot \exp -\frac{H_m + H_F}{kT}$$

  - G_m is the free enthalpy for a jump, i.e. the free enthalpy barrier that must be overcome between two identical positions in the lattice.
  - In words: All the material dependent constants (including the migration and formation entropy) have been lumped together in D^*; and the exponential now contains the *sum* of the migration and formation energy of a vacancy.
- Lets discuss this equation a bit:
  - As mentioned before, we need an entropy of migration as a parameter of a point defect. In summary we need four parameters correlated with an intrinsic point defect to describe its diffusion behavior (if we discount the vibration frequency).
  - But only two parameters, the formation energy and the migration energy are of overwhelming importance.
  - Everything else may be summarized in a (more or less) constant pre-exponential factor D^* which contains the entropies. Since the entropies may be temperature dependent (for **Si** this is probably the case), you must look at bit closer at your calculations if you are interested in precise diffusion data.
  - An **Arrhenius**-representation (**lg** D vs. $1/T$) will give a straight line, the slope is given by $H_M + H_F$. The pre-exponential factor determines the intersection with the axis and is thus measurable.
  - Since it is much easier to measure diffusion coefficients compared to point defect densities, the *sum* $H_M + H_F$ for point defects is mostly much better known than the *individual* energies. Some values are given in the link.
  - Self-diffusion vial self-interstitials follows essentially the same laws.
- For self-diffusion in **Si**, we find the following (rather small) values : $D_{SD} = (10^{-21} \text{ — } 10^{-16})\ m^2/s$ in the relevant temperature regime. Detailed data in an Arrhenius plot for for self-diffusion in **Si** can be found in the link; some numbers for **Si** self-diffusion as well as the migration parameters of vacancies and interstitials and a few elements are also illustrated.
- Now for an exercise. Self-diffusion means that the atoms in a crystal change their position. After some time *all* atoms will have changed their positions at least once.
  - Our crystal just lying there, somehow changes identity. What does it mean? How long does it take? Do the exercise **3.2-1**!

Exercise 3.2-1
Crystal Identity

Exercise 3.3-1
Quick Questions to 3.

3.2.3 Diffusion of Impurity Atoms

Diffusion of Impurity Atoms via Vacancies or Self-Interstitials

- In this case it is especially important not to confuse the *vehicle* (a point defect except in the case of the direct interstitial diffusion) with the *diffusing impurity atom*.
  - What we want to have are the diffusion data for the impurity atom, not for the vehicle! We will not go into details at this point, but it will come up later again.

Direct diffusion

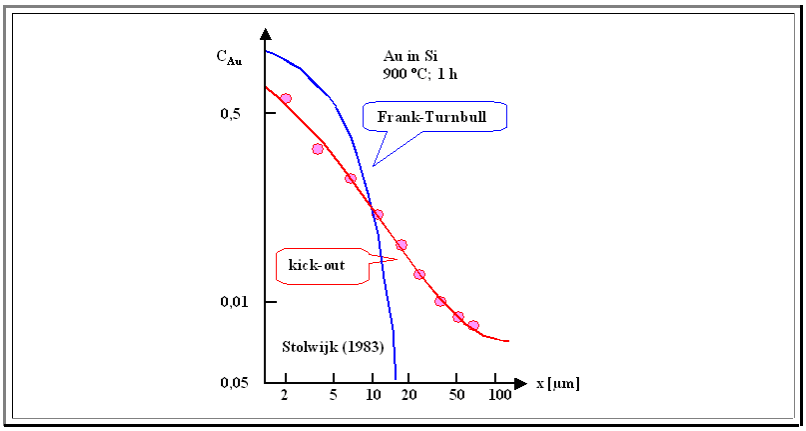
- This is the simplest mechanism, it does not need point defects. It only works for *interstitial impurity atoms*. The diffusion coefficient D_{dir} is simply given by

$$D_{dir} = g \cdot a^2 \cdot \nu_0 \cdot \exp\frac{S_m}{k} \cdot \exp -\frac{H_m}{kT}$$

  - With H_m, S_m being the enthalpy and entropy of the jumping impurity atom.
- This is in most cases a sufficiently good approximation. The host lattice only enters in the form of the lattice factor ***g*** (and to some extent in the migration entropy), but plays no other role. This is of course a source of possible aberrations, because the ideal lattice implied in this case does normally not exist. Good examples for directly diffusing impurities are **O** in **Si**, but there are many other elements.
  - The following table (taken from K. **Graff**, Metal Impurities in **Si**-Device Fabrication; Springer Series in Mat. Science **24**) gives a few more examples.

Interstitial Impurity Atom Diffusion in **Si**					
Metal	H_{Sol} [eV]	S_{Sol}[k]	H_m [eV]	D^*[m^2/s]	T-regime [^0C]
Ti	3,05	4,22	1,79	$1{,}45 \times 10^{-2}$	950 - 1200
Cr	2,79	4,7	0,99	$1{,}0 \times 10^{-2}$	900 - 1250
Mn	2,80	7,11	0,6	$5{,}7 \times 10^{-4}$	900 - 1200
Fe	2,94	8,2	0,68	$1{,}3 \times 10^{-3}$	30 - 1200
Co	2,83	7,6	0,53	$4{,}2 \times 10^{-3}$	900 - 1100
Ni	1,68	3,2	0,47	$2{,}0 \times 10^{-3}$	800 - 1300
Cu	1,49	2,4	0,43	$4{,}7 \times 10^{-3}$	400 - 900

  - H_{Sol} and S_{Sol} denote the **solubility enthalpies** and -**entropies**. This is the extrinsic point defect quivalent of the formation enthalpy and entropy of intrinsic point defects.
- Here are some links illustrating impurity diffusion:
  - From a review of U. **Gösele**, we take the Arrhenius plot for many impurity atoms in **Si**.
  - In this link we have a somewhat unusual way of showing impurity diffusion in **Si**.
- Here some specialities: Frank-Turnbull and Kick-out Mechanism in **Si**
  - **U. Gösele** essentially "invented" the kick-out mechanism around **1985** and demonstrated that it not only provided an alternative to the already known Frank-Turnbull Mechanism, but does most likely control the diffusion of ***Au***, ***Pt***, and possibly ***Zn*** in **Si**.
  - The two mechanisms, although similar in many ways, lead under certain circumstances to very different diffusion profiles, as shown in the graph below taken from a *review article* (*Fast Diffusion in Semiconductors*) accessible through the link.



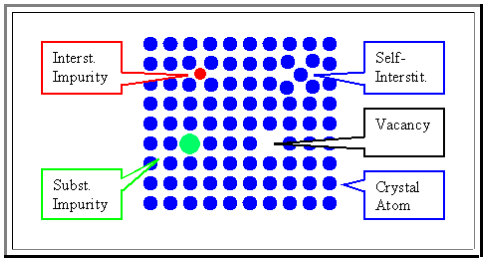
- Often it is quite difficult to decide from the data what kind of mechanism is operative. During the last few years, some research groups studying diffusion in **GaAs** came to the conclusion that the diffusion mechanisms in **GaAs** might be very different from what was accepted before (invoking, e.g., a kick-out mechanism).
- There is no evidence so far that impurity diffusion in crystals of any kind of crystal involves a **direct mechanism**. Direct mechanisms, however, are periodically suggested in the literature and should not be ruled out per se.

So, what is the *Message* of this sub-chapter?

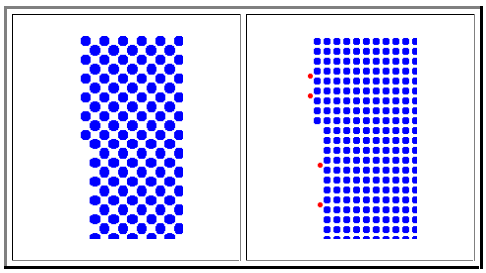
- Many diffusion phenomena, especially in semiconductors or more complicated crystals, are still not very well understood. Precise modeling of diffusion, however, depends sooner or later on using the correct mechanisms.
- Data from diffusion measurements are always (sometimes "encrypted") data on point defects.

3.2.4 Essentials to Chapter 3.2 Diffusion Mechanisms

- Considering diffusion in crystals we have exactly three basic cases
  - **1.** An interstitial impurity atom diffuses in the crystal=impurity diffusion.
  - **2.** A substitutional impurity atom diffuses in the crystal=impurity diffusion.
  - **3.** An atom of the crystal diffuses in the crystal=self-diffusion.
- Case **2.** and **3.** are impossible without diffusion "vehicles", i.e. vacancies (and on occasion self-interstitials).



- Diffusion mechanisms are the atomic mechanisms that are capable of moving atoms. The most important ones are:
  - Vacancy mechanism. Accounts for most of cases **2.** and **3.** from above in simple crystals,
  - Direct interstitial mechanisms. Accounts for almost all of case **1.**



- Some more complex mechanisms exist (and are of prime importance) in **Si** (and possibly other semiconductors and somewhat more complex crystals)
  - "Kick-out" mechanism, impurity and self-diffusion via self-interstitials, ...

Wait and see!
And keep an open mind

- In any case we need the migration enthalpy H_m and entropy S_m of the "jumping" entities to obtain the diffusion coefficient D of the process
  - Typical values are - like always, it seems - in the **1 eV** (better: **0.5 eV - 3 eV**) and **1 k** region, respectively.
  - Question to ponder: How long does it take for all atoms of crystal to be somewhere else; i.e. not at the original position? (Exercise 3.2-1)

$$D_{dir} = g \cdot a^2 \cdot \nu_0 \cdot \exp\frac{S}{k} \cdot \exp -\frac{H_m}{kT}$$

$$D_{SD} = c_V \cdot D_V = D^* \cdot \exp -\frac{H_m + H_F}{kT}$$

3.3 Experimental Approaches to Diffusion Phenomena

3.3.1 Determination of Diffusion Profiles

General Remarks

- In a typical diffusion experiment, some impurity atoms are introduced into a host by first putting them (ideally) with δ - distribution at the surface.
  - After annealing for a specified time at a specified temperatures, some diffusion of the impurity atoms will have produced a **diffusion profile**, i.e. a smooth curve of the concentration ***c*** vs. depth ***x*** in the sample (usually plotted as **lg(*c*)** - ***x*** curve).
  - Some experimental care is necessary. Simply depositing the impurity atoms on the surface of the host crystal may not lead to any results, because e.g. an impenetrable oxide layer may prevent any diffusion of the impurity atoms into the crystal. "Shooting" the impurity atoms into a surface-near area via ion implantation will overcome that problem, but may create its own problems by generating point defects which change the regular diffusion behavior.
  - There are some well established standard methods for measuring the diffusion profiles after a successful diffusion experiment (see below). Fitting the profile to the applicable solution of Ficks law will provide two results:

    1. The numerical value of the diffusion coefficient ***D*** for the set of parameters considered.
    2. The validity of Ficks law for this case as evidenced by the quality of the fit.
  - Of course, any "macroscopic" method for measuring profiles relies on having a profile on a, lets say, **10 µm** scale in the first place, i.e. each impurity atom must have made *many* individual jumps.
  - This will in most cases only happen at sufficiently high temperatures. Waiting a long time is not very effective; this is immediately clear if looking at diffusion phenomena in a slightly different way.
- How then do we get experimental data at small concentrations or small numbers of jumps? The answer is: use radioactive **tracer atoms** as the diffusing atoms, that can be found and identified in extreme small concentrations!

Tracer Techniques

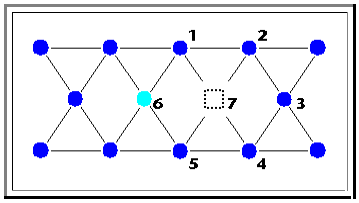
- Radioactive tracer atoms can be easily detected whenever they decay, emitting some high energy radiation.
  - If the half-life time of the tracer used is relatively small (but still large enough to allow an experiment before the tracer has vanished), a large percentage of the tracer atoms can be detected by their decay products - typically α, β, or γ-rays. We thus may have an extremely high detection efficiency, many orders of magnitude below the detection limits of standard methods.
- Lets consider the general way a tracer experiment is done:
  - Deposit a thin layer of the atoms that are to diffuse on the (very clean) host crystal. Some of those atoms should be a suitable radioactive isotope of the species investigated. Use any deposition technique that works for you (evaporation, sputtering techniques, sol-gel techniques ("painting it on")...), but make sure that the deposition technique does not alter your substrate (sputtering, e.g., may produce point defects) and that you have no "barrier layer" between the substrate and the thin layer.
  - Anneal for a suitable time at a specific temperature.
  - Remove thin layers from the surface (ideally one atomic layer after the other) by, e.g. sputtering techniques, anodic oxidation and chemical stripping, ultramicrotomes, chemical dissolution, ...).
  - Measure the radioactivity of each layer.
  - With the known half-life of the tracer and the time since the deposition of the layer, calculate how many tracer atoms are in your layer. From the measurement of many layers a concentration profile of the tracer atoms results.
- The rest is conventional: Fit the profile against a standard solution of Ficks law or against your own solution and extract the diffusion coefficient for the *one* temperature used. This gives *one* data point. And then:
  - Repeat the experiment for several other temperatures, collecting data points for different temperatures.
  - From an Arrhenius representation of the measured diffusion coefficients you obtain D_0 *and* an activation energy for the tracer diffusion if your data are on a (halfway) straight line.
- If this sounds tedious, it's because it is! You appreciate why students doing a master or PhD thesis are so essential to research. Still, nothing beats tracer experiments when it comes to sensitivity and accuracy.
  - There is, however, a basic problem that we have to discuss if you want to extract information about the vehicle of the tracer diffusion, i.e. about the vacancies or, in some cases, interstitials from a tracer experiment. This is always the case when dealing with self-diffusion.

- The diffusion coefficient of the tracer atom is not necessarily identical with the diffusion coefficient for self-diffusion as defined for the vehicles - usually vacancies.

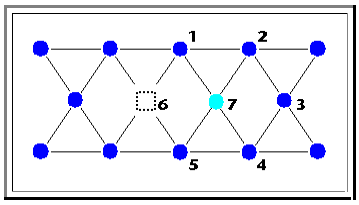
The reason for this is that the tracer is a *specific* atom, while we look at *many* vacancies that help it along - and we must not confuse the vehicle wih the diffusing impurity (or tracer) atom, as noted before. In particular, the jumps of the tracer atom may be *correlated* with the jumps of the individual vacancy coming by.

- In other words, whereas a particular vacancy may (and usually does) jump around in a perfect *random walk* pattern (i.e. each jump contributes to the mean square displacement of the vacancy), the *tracer atom* may *not* move randomly!

Lets look at a simple example for a two-dimensional vacancy diffusion mechanism.



The tracer atom is marked in light blue,
it has a vacancy as a neighbor, a jump is possible.



Vacancy and tracer atom have exchanged their positions.

- Next, the vacancy will jump again - with equal probability on one of the **6** surrounding atom sites - so it is truly doing a random walk. And *one* of those jumps goes back to position **7**, with exactly the same probability as to the other available sites.
- The "viewpoint" of the tracer atom, however, is different. It will jump back to site **6** with a *higher probability* than to the sites **1 - 5** because a vacancy *is available on **6***, whereas for the other sites the passing of some *other* vacancy must be awaited. There is a *correlation between jump **1** and jump **2** - there is no random walk*. The jumps back will lead to wrong values of the mean square displacement, because this combination does not add anything and occurs more frequently as it would for a truly random walk.

The correlation effects between individual jumps of the tracer atom and the random jumps of vacancies can be calculated by a rigid theory of diffusion by individual jumps, but here I won't go into that.

- As a result, these **correlation effects** (in all dimensions and for all lattice types) can be dealt with by defining a **correlation factor *f*** that must be introduced into the equations coupling the tracer diffusion to the vacancy diffusion.

We define a correlation coefficient f_{1V} that allows to correlate the diffusion coefficient for the (vacancy driven) self-diffusion, $D_{SD}(T)$, *as measured by a tracer experiment*, to the diffusion coefficient for self-diffusion, $D_{SD}(\text{Theo})$ *as given by theory* via the equation

$$D_{SD}(T) = f_{1V} \cdot D_{SD}(\text{Theo})$$

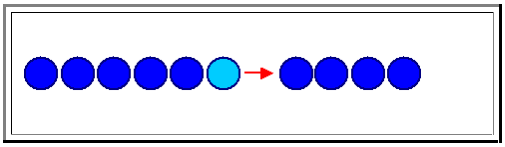
- As an example for a real correlation factor we look at $f_{1V}(\text{cub})$, the correlation factor for self-diffusion mediated by single vacancies in a cubic lattice. It is given in a good approximation by

$$f_{1V}(\text{cub}) \approx 1 - \frac{2}{z} = \begin{matrix} 5/6 \text{ fcc} \\ 3/4 \text{ bcc} \end{matrix}$$

- With z = number of nearest neighbors.

To illustrate the correlation phenomena, suppose that $f = 0$. In this case, even for wildly moving vacancies ($D_{SD} >> 0$), the tracer atoms would not move - we would not observe any diffusion.

- This case is fully realized for *one-dimensional* diffusion, where it is also easy to see what happens - just consider a chain of atoms with one vacancy:



- The vacancy may move back and forth the chain like crazy - the tracer atom (light blue) at most moves between two position, because on the average there will be just as many vacancies coming from the right (tracer jumps to the left) than from the left (tracer jumps to the right).

Correlation coefficients can be *calculated* - as long as the diffusion mechanism and the lattice structure are known. They are, however, very difficult to *measure* which is unfortunate, because they contain rather direct information about the mechanism of the diffusion. The calculations, however, are not necessarily easy.

- Impurity atoms, which may have some interaction with a vacancy, may show complicated correlation effects because in this case the vacancy, too, does no longer diffuse totally randomly, but shows some correlation to whatever the impurity atom does.
- If a kick-out mechanism is active, the tracer atom might quickly be found immobile on a lattice site, whereas another atom - which however will not be detected because it is not radioactive - now diffuses through the lattice. The correlation factor is very small.

Some examples for correlation coefficients are given in the table for a simple vacancy mechanism (after **Seeger**). The correct value from extended calculations is contrasted to the value from the simple formula given above.

Lattice type	coordination number z	$f_{1V} \approx 1 - 2/z$	f_{1V} (correct)
One dim. lattice			
Chain	2	0	0
Two dim. lattices			
hex. close packed	6	0.6666	0,56006
square	4	0,5	0,46694
Three dim. lattice			
cub. primitive	6	0,6666	0,65311
Diamond	4	0,5	0,5
bcc	8	0,75	0,72722
fcc	12	0,83	0,78146

Other Methods for Measuring Diffusion Coefficients

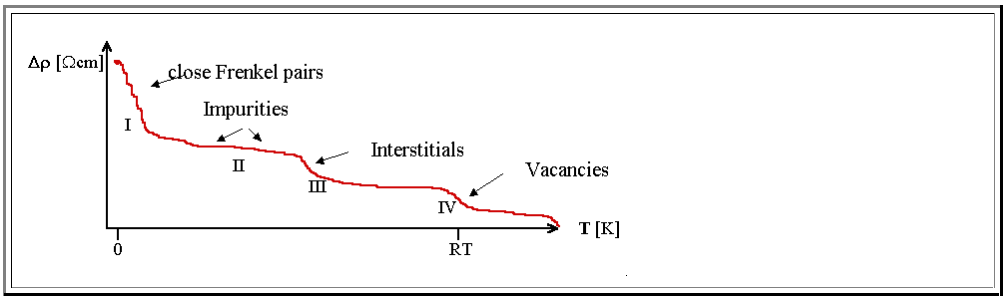
There is a ***plethora*** of methods, some are treated in other lecture courses. In what follows a few important methods are just mentioned.

Concentration Profile Measurements

- **Secondary Ion Mass Spectrometry (*SIMS*)** for direct measurements of atom concentrations. This is the most important method for measuring diffusion profiles of dopants in Si (and other semiconductors).
- **Rutherford Backscattering (*RBS*)** for direct measurements of atom concentrations.
- Various methods for measuring the *conductivity* as a function of depth for semiconductors which corresponds more or less directly to the concentration of doping atoms. In particular:
  - Capacity as a function of the applied voltage ("***C(U)***") for ***MOS*** and junction structures)
  - Spreading resistance measurements
  - Microwave absorption.
- Local **growth kinetics** of defects, e.g. the precipitation of an impurity, contain information about the diffusion, e.g.
  - Growth of oxidation induced stacking faults in **Si**
  - Impurity -"free" regions around grain boundaries (because the impurities diffused into the grain boundary where they are trapped).
- An example for a "diffusion denuded" zone along grain boundaries can be seen in the illustration

Annealing experiments (See also chapter 4.2.1)

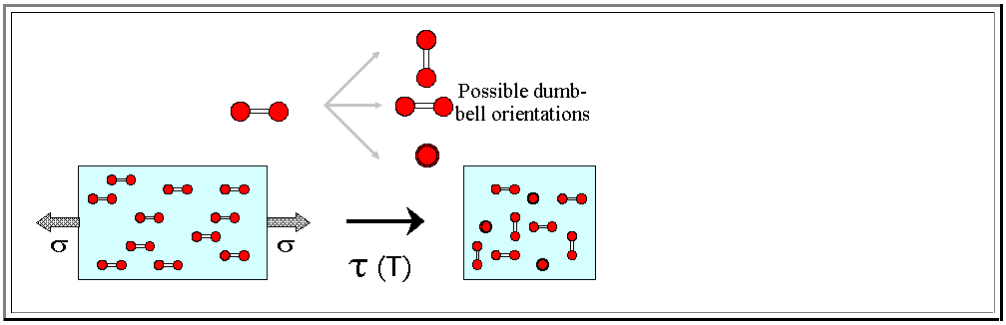
- These experiments are in a class of their own. In this case point defects which have been rendered immobile in a large *supersaturation*, e.g. by rapid cooling from high temperatures, are made mobile again by controlled annealing at specified temperatures. Since they tend to disappear - by precipitation or outdiffusion - measuring a parameter that is sensitive to point defects - e.g. the residual resistivity - will give kinetic data.
- A classical experiment produces supersaturated point defects by irradiation at low temperatures with high-energy electrons (**a few MeV**). The energy of the electrons must be large enough to displace single atoms - Frenkel pairs may be formed - but not large enough to produce extended damage "cascades".
- Annealing for a defined time at a specified temperature will remove some point defects which is monitored by measuring the residual resistivity - always at the same very low temperature (**usually 4K**). Repeating the sequence many times at increasing temperatures gives an annealing curve. A typical annealing curve may look like this:



- What "impurities" means in this context is left open. They may form small complexes, interact with nearby vacancies or interstitials, or whatever.
- The interpretation of the steps in the annealing curves as shown above is not uncontested. The "Stuttgart school" around **A. Seeger** has a completely different interpretation, invoking the "crowdion", than the (more or less) rest of the world.

Methods measuring single atomic jumps

- This ultimate tool can be used if the point defects have rather low symmetry. The best example is the dumbbell configuration of the interstitial or interstitial carbon in **Fe**
- In the classical experiment the crystal is uniaxially deformed at not too low temperatures. The dumbbells will, given enough time, orient themselves in the direction of tensile deformation (there is more space available, so the energy is lower) and thus carry some of the strain. We have more dumbbells in one of the three possible orientations than in the two other ones (see below)

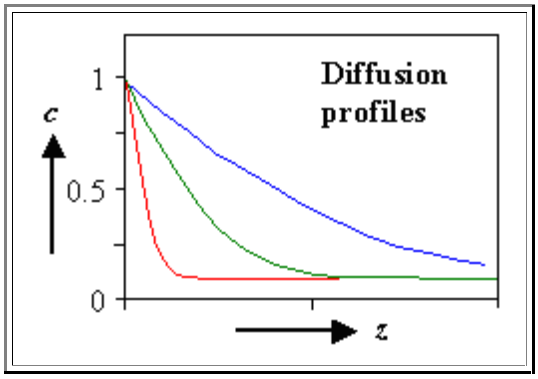


- The tensile stress is now suddenly relieved. Besides the purely and instantaneous elastic relaxation, we will now see a slow and temperature dependent additional relaxation because the dumbbells will randomize again. The time constant of this process directly contains the jump frequency for dumbbells. This effect, which exists in many variants, is called "**Snoek effect**".
- If you do not use a static stress, but a periodic variation with a certain frequency ω, you have a whole new world of experimental techniques!

Last, there are methods which monitor the *destruction (or generation) of some internal order* in the material. The prime technique is **Nuclear Magnetic Resonance** (***NMR***), which monitors the decay of nuclear magnetic moments which were first oriented in a magnetic field and then disordered by atomic jumps, i.e. diffusion. The **Mößbauer effect** may be used in this connection, too.

3.3.2 Essentials to Chapter 3.3 Experimental Approaches to Diffusion Phenomena

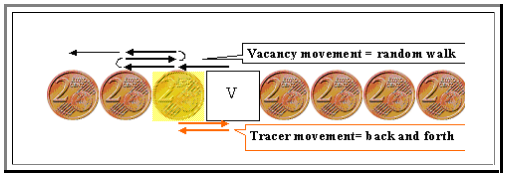
- It's easy in principle: You produce and measure a diffusion profile.
  - Put whatever is supposed to diffuse on the crystal surface (make sure you cope properly with the "dirt" or oxide on the surface).
  - Let it diffuse at a defined ***T*** for a defined time ***t***.
  - Measure the diffusion profile "somehow".
  - Fit to a solution of Fick's law = one data point for ***D*(*T*)**.
  - Repeat at different temperatures until you gave enough data points for an (Arrhenius) ***D*(*T*)** plot.



- Use isotopes of the material in question for self-diffusion measurements.

- While the intrinsic point defect serving as diffusion vehicle will do a perfectly random walk, the diffusing atom may *not*.
  - There is a correlation coefficient ***f*** linking measured and theoretical diffusion coefficients.

$$D_{SD}(T) = f_{1V} \cdot D_{SD}(\text{Theo})$$



  - The correlation coefficient ***f*** is **= 0** for **1dim**. diffusion, around **1/2 - 2/3** for **2dim.** diffusion (e.g. in the base plane of hexagonal lattices) and around **2/3 - 3/4** for **3dim.** diffusion.

- There are many other ways to obtain diffusion data, none fool-proof and all money and/or time expensive.

4. Experimental Techniques for Studying Point Defects

4.1 Point Defects in Equilibrium

4.1.2 Essentials to Chapter 4.1: Experimental Techniques for Studying Point Defects in Equilibrium

4.2 Point Defects in Non-Equilibrium

4.2.2 Essentials to Chapter 4.2: Experimental Techniques for Studying Point Defects in Non-Equilibrium

4.3. Specialities

4. Experimental Techniques for Studying Point Defects

4.1 Point Defects in Equilibrium

Differential Thermal Expansion Method

- How can we measure *directly* the type and concentration of point defects and, if we do it as function of temperature, extract the *formation energies* and *formation entropies*?
  - Simple question - but there is essentially only one *direct* method: Measure the change of the lattice constant ***a***, i.e. Δa, and the change in the specimen dimension, Δl, (one dimension is sufficient) simultaneously as a function of temperature.
  - What you have then is the differential thermal expansion method also called the $\Delta l/l - \Delta a/a$ method.
  - This method was invented by **Simmons** and **Balluffi** around **1960**.
- The basic idea is that $\Delta l/l - \Delta a/a$ (with l = length of the specimen $= l(T$, **defects)**) contains the regular thermal expansion *and* the dimensional change from point defects, especially vacancies.
  - This is so because for every vacancy in the crystal an atom must be added at the surface; the total volume of the vacancies must be compensated by an approximately equal additional volume and therefore an additional Δl.
  - If we subtract the regular thermal expansion, which is simply given by the change in lattice parameter, whatever is left can *only* be caused by point defects. The difference then gives directly the vacancy concentration.
  - For a cubic crystal with negligible relaxation of the atoms into the vacancy (so the *total* volume of the vacancy provides added volume of the crystal), we have

$$3\left(\frac{\Delta l}{l} - \frac{\Delta a}{a}\right) = c_V - c_i$$

  - With c_V = vacancy concentration, c_i = interstitial concentration.
  - We have to take the *difference of the concentration* because interstitial atoms (coming from a vacancy) do *not* add volume.
- This is quite ingenious and straightforward, but not so easy to measure in practice.
  - The measurements of both parameters have to be very precise (in the 10^{-5} range); you also may have to consider the double vacancies.
  - But successful measurements have been made for most simple crystals including all important metals, and it is this method that supplied the formation energies and entropies for most important materials.
- The link shows a successful measurement of $\Delta l/l - \Delta a/a$ for **Ag + 4% Sb**.
- Some values mostly obtained with that method are shown in the following table (after **Seeger**):

Element	c_V at T_m	H_F [eV]	S_F [k]
Cu	2×10^{-4}	1,04	0,3
Ag	$1{,}7 \times 10^{-4}$	0,99	0,5
Au	$7{,}2 \times 10^{-4}$	0,92	0,9
Al	9×10^{-4}	0,65	0,8
Pb	$1{,}7 \times 10^{-4}$	0,5	0,7
Na	7×10^{-4}		
Li	4×10^{-4}		
Cd	$6{,}2 \times 10^{-4}$		
Kr	3×10^{-3}		

Si	no values, $\Delta l/l - \Delta a/a = 0$ even at ultra-high precision

Positron Annihilation

- A somewhat exotic, but still rather direct method is measuring the time constant for positron ***annihilation*** as a function of temperature to obtain information about vacancies in thermal equilibrium.
  - What you do is to shoot positrons into your sample and measure how long it takes for them to disappear by annihilation with an electron in a burst of γ - rays. The time from entering the sample to the end of the positron is its (mean) life time τ.
  - It is rather short (about 10^{-10} seconds), but long enough to be measured, *and it varies with the concentration of vacancies* in the sample. Since electrons are needed for annihilation and a certain overlap of the wave functions has to occur, the life time τ is directly related to the average electron concentration available for annihilation.
  - A nice feature of these technique is that the positron is usually generated by some radioactive decay event, and then announces its birth by some specific radiation emitted simultaneously. Its death is also marked by specific γ rays, so all you have to do is to measure the time between two special bursts of radiation.
- Vacancies are areas with low electron densities. Moreover, they are kind of attractive to a positron because they form a potential well for a positron - once it falls in there, it will be trapped for some time.
  - Since an average life time of 10^{-10} **s** is large enough for the positron, even after it has been thermalized, to cover rather large distances on an atomic scale, some positrons will be trapped inside vacancies and their percentage will depend on the vacancy concentration.
  - Inside a vacancy the electron density is smaller than in the lattice, the trapped positrons will enjoy a somewhat longer life span. The average life time of all positrons will thus go up with an increasing number of vacancies, i.e. with increasing temperature.
- This can be easily quantified in a good approximation as follows.
  - Lets assume that on the average we have n_0 (thermalized) positrons in the lattice, split into n_1 "free" positrons, and n_2 positrons trapped in vacancies; i.e.

$$n_0 = n_1 + n_2$$

  - The free positrons will either decay with a fixed rate λ given by $\lambda_1 = 1/\tau_1$, (with τ_1 = (average) lifetime), or are trapped with a probability ν by vacancies being present in a concentration c_V.
  - The trapped positrons then decays with a rate λ_2 which will be somewhat smaller then λ_1 because it lives a little longer; its average lifetime is now τ_2.
  - The change in the partial concentration then becomes

$$\frac{dn_1}{dt} = -(\lambda_1 + \nu \cdot c_V) \cdot n_1$$

$$\frac{dn_2}{dt} = -\lambda_2 \cdot n_2 + \nu \cdot c_V \cdot n_1$$

- This system of coupled differential equation is easily solved (we will do that as an exercise), the starting conditions are

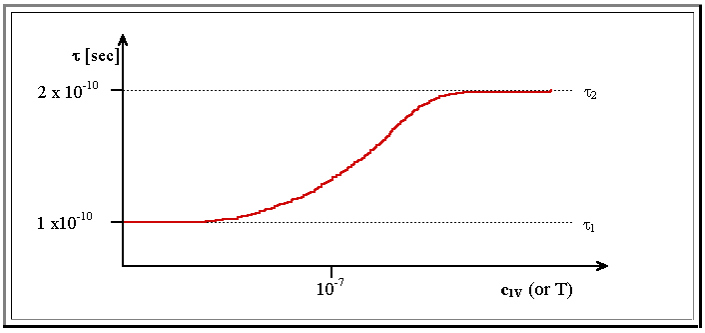
$$n_1(t = 0) = n_0$$

$$n_2(t = 0) = 0$$

  - The average lifetime τ, which is the weighted average of the decay paths and what the experiment provides, will be

$$\tau = \tau_1 \cdot \left(\frac{1 + \tau_2 \cdot \nu \cdot c_V}{1 + \tau_1 \cdot \nu \cdot c_V} \right)$$

The probability ν for a positron to get trapped by a vacancy can be estimated with relative ease, the following principal "**S" - curve** is expected. By now, it comes as no surprise that no effect was found for **Si**.



The advantage of positron annihilation experiments is its relatively high sensitivity for low vacancy concentrations (**10^{-6} - 10^{-7}** is a good value), the obvious disadvantage that a quantitative evaluation of the data needs the trapping probability, or cross section for positron capture.

Some examples of real measurements and further information are given in the links:
Life time of positrons in Ag
Life time of positrons in Si and Ge.
Paper (in German): Untersuchung von Kristalldefekten mit Hilfe der Positronenannihilation

- A large table containing values for **H_F** as determined by positron annihilition (and compared to values obtained otherwise) can be found in the link

Exercise 4.1-1

Derive the Formula for τ

More Direct Methods for Measuring Point Defect Properties

There isn't much. Some occasionally used methods are

- Measurements of the *resistivity*. Very suitable to ionic crystals if the mechanism of conduction is ionic transport via point defects. But you never know for sure if you are measuring intrinsic equilibrium because "doping" by impurities may have occurred.
- *Specific heat* as a function of ***T***. While there should be some dependence on the concentration of point defects, it is experimentally very difficult to handle with the required accuracy.
- Measuring *electronic noise*. This is a relatively new method which relies on very sophisticated noise measurements. It is more suited for measuring diffusion properties, but might be used for equilibrium conditions, too. The illustration in the link shows a noise measurement obtained upon annealing frozen-in point defects.

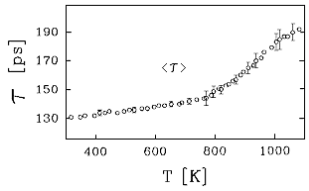
However, the view presented above (and in the chapters before) is not totally unchallenged. There are serious scientists out there who claim that things are quite different, especially with respect to equilibrium concentrations of vacancies in refractory metals, because the formation entropy is much higher than assumed.

- The method of choice to look at this is *calorimetry* at high temperature, i.e. the measurement of the specific heat. A champion of this viewpoint is Y. **Kraftmakher**, who just published a book to this point.

4.1.2 Essentials to Chapter 4.1: Experimental Techniques for Studying Point Defects in Equilibrium

- Essentially we have two rather direct methods
  - Differential Thermal Expansion (or $\Delta l/l - \Delta a/a$ -method).
  - .Positron annihilation
- Both methods will not give results if the vacancy concentration at the melting point is below, roughly, 10^{-7}.

$$3\left(\frac{\Delta l}{l} - \frac{\Delta a}{a}\right) = c_V - c_i$$



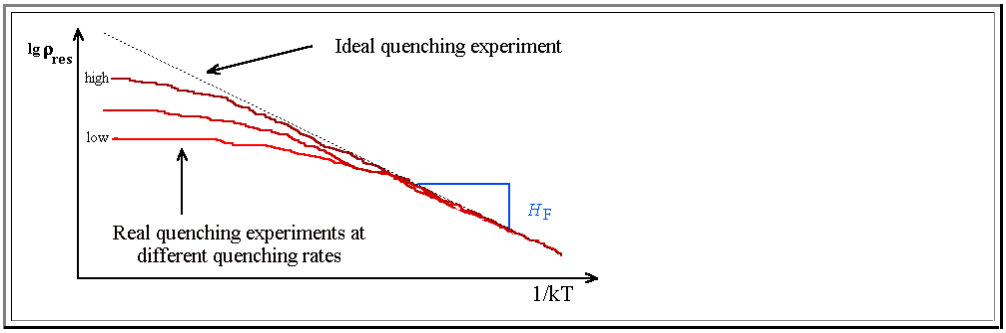
$$\langle\tau\rangle = \tau_2 \cdot \frac{1 + \tau_{2V} \cdot c_V}{1 + \tau_{1V} \cdot c_V}$$

- Most numbers for point defects in metals and some other crystals were obtained by these two methods.
- There are many other methods, but always either limited to certain crystals, expensive, hard to evaluate, and so on.
- In essence, there are still no reliable and undisputed numbers for, e.g., the formation and migration enthalpies for vacancies (and interstitials) in **Si** or other semiconductors like **GaAs**.

4.2 Point Defects in Non-Equilibrium

Quenching Experiments

- The basic idea behind these techniques is simple: if you have *more* point defects than what you would have in thermal equilibrium, it should be easier to detect them. There are several methods, the most important one being ***quenching*** from high temperatures. Lets look at this technique in its *extreme* form:
  - A *wire* of the material to be investigated is heated to some desired (high) temperature ***T*** in liquid and *superfluid* **He II** (i.e. a liquid with a "∞" large heat conduction) to the desired temperature (by passing current through it). Astonishingly, this is easily possible because the **He**-vapor produced acts as a very efficient thermal shield and keeps the liquid **He** from exploding because too much heat is transferred.
  - After turning off the heating current, the specimen will cool extremely fast to **He II** temperature (≈ **1K**). There is not much time for the point defects being present at the high temperature in thermal equilibrium to disappear via diffusion; they are to a large percentage "**frozen-in**". The frozen-in concentration can now be determined by e.g. measuring the **residual resistivity** ρ_{res} of the wire, the link gives an old example.
  - The residual resistivity is simply the resistivity found around **0 K**. It is essentially dominated by defects because scattering of electrons at phonons is negligible.
- There are, however, many *problems* with the quenching technique.
  - The *quenching speed* (≈ 10^4 **°C/s** with the **He II** technique) may still be too small to definitely rule out agglomeration off point defects (look at exercise 4.2-1). The cure for this problem is to repeat the experiments at different quenching speeds and to extrapolate to infinite quenching speed. What you will see for e.g. the residual resisitivity ρ_{res} may look like the schematic representation below.



  - We assumed in a fairly good approximation that $\rho_{res} \propto c_V$; so we should get Arrhenius behaviour for ρ_{res}.
  - Recorded is the ρ_{res} in an Arrhenius plot as a function of the emperature ***T*** from which it was quenched. If you get a decent piece of a straigth line you can deduce the vacnacy formation enthalpy.
- *Plastic deformation* is the next big problem.
  - The unavoidable large temperature gradients introduced by quenching produce large mechanical stress which may cause severe plastic deformation or even fracture of the specimen. Plastic deformation, in turn, may severley distort the concentrations of point defects and fracture of a sample simply terminates an experiment.
- Finally, *impurities*, always there, may influence the results.
  - Since impurities may drastically influence the residual resistance, measurements with "dirty" specimens are always open to doubt. In addition, it is not generally easy to avoid in-diffusion of impurity atoms at the high temperatures needed for the experiment.
  - Quenching experiments with **Si**, for example, did not so far give useful data. If any "good" curves were obtained, it was invariably shown (later) that the results were due to impurity in-diffusion (usually **Fe**).
- The illustration in the link gives an example for the processes occurring during quenching for **Au** obtained by calculations and demonstrates the difficulties in extracting data from raw measurements.

Exercise 4.2-1

Diffusion during cooling

Other Methods

- *If all else fails*: try to find *agglomerates of point defects* looking at your specimen with the **transmission electron microscope** (***TEM***), with **X**-ray methods or with any other method that is applicable.
  - Accept local equilibrium: Don't cool too fast, allow time for agglomerates to form. Conclude from the type of agglomerate, from their density and size, and whatever additional information you can gather, what kind of point defect with what concentration was prevalent.
  - This is rather indirect and qualitative, but:
- It gives plenty of information. There are many examples where **TEM** contributed vital information to point defect research. Especially, it was **TEM** that gave the first clear indication that self-interstitials play a role in thermal equilibrium in **Si** and some rough numbers for formation energies and migration energies (Föll and Kolbesen **1978**).
- In the link an example of the agglomerates of self-interstitials as detected by **TEM** is given. The major experimental problem in this case was to find the agglomerates. Their density is very low and at the required magnification huge areas had to be searched.

- A very new way of looking at point defects is to use the **scanning tunneling microscope** (***STM***) and to look at the atoms on the surface of the sample. This idea is not new; before the advent of the **STM field ion microscopy** was used with the same intention, but experiments were (and are) very difficult to do and severely limited.
  - One idea is to investigate the surface after fracturing the quenched sample in-situ under ultra-high vacuum (***UHV***) conditions. This would give the density of vacancies on the fracture plane from which the bulk value could be deduced.
  - An interesting set of **STM** images of point defects in **GaAs** from recent research is given in the link.
  - Vacancies can be seen, but there are many problems: The image changes with time - the density of point defects goes *up*! Why - who knows?
  - The interpretation of what you see is also difficult. In the example, several kinds of contrasts resulting from vacancies can be seen, probably because they are differently charged or at different depth in the sample (**STM** also "sees" defects one or two layers below the top layer). It needs detailed work to interprete the images as shown in the link.
  - More recent pictures show the surface of **Si** or **Pt**, including point defects, in astonishing clarity. But we still will have to wait a few more years to see what contributions **STM** will be able to make towards the understanding of point defects.

4.2.2 Essentials to Chapter 4.2: Experimental Techniques for Studying Point Defects in Non-Equilibrium

- Non-equilibrium can be obtained in several ways; one always tries to have point defect concentrations far above equilibrium.
  - *Quenching*, i.e. freezing-in some equilibrium concentration (or some non-equilibrium concentration) at the low temperature T_{quench} that was present at the temperature T.
  - Irradiation (e.g. with electrons) that mostly produce vacancy - interstitial pairs in a concentration given by the irradiation intensity and thus will be above thermal equilibrium.
- After the point defects have been frozen-in, i.e. immobilized, you measure a property that is sensitive to point defects, most prominently the conductivity at low temperatures, and then study how this property changes upon annealing, i.e. letting your point defects achieve equilibrium (= disappear).
  - If you started from equilibrium, you will get equilibrium concentration and diffusion data that must be separated "somehow".
  - If you started from a non-equilibrium concentrations, you will *only* get diffusion data, i.e. migration enthalpies and entropies.

What happens during cooling down - rapidly or otherwise?

Question to ponder:

How far can a point defect move during cooling or what is the total diffusion length L_{total}?

Exercise 4.2-1

L_{total} determines

- How well quenching works
- Density of agglomerates
- Size of agglomerates

4.3. Specialities

Special Methods for Ionic Crystals

- In ionic crystals, experimental investigations must follow different routes.
  - The ∆l/l - ∆a/a method will not work *by definition* for Frenkel defects, where the concentrations of vacancies and interstitials are identical and the volume change zero.
  - It might work for Schottky defects and mixed defects. In the latter case, however, it will not be possible to obtain information for the individual point defect types involved because the measurement only gives integral numbers.
  - Quenching is difficult if not impossible, because ionic crystals are usually bad heat conductors; this will limit the quenching speed to useless values. In addition, ionic crystals tend to be brittle and they usually fracture upon quenching.
  - Positrons will also be trapped by the negatively charged ions, the technique is not applicable.
  - And last but not least: it is quite unlikely that what you find are *equilibrium* numbers anyway, because point defects in ionic crystals are so sensitive to deviations from stochiometry and so on.
- Fortunately, *there are methods specific for ionic and oxide crystals*; most prominent is the measurement of the ionic conductivity which is often mediated by point defects and therefore can be used to gather information about point defects.
  - Spectroscopic methods (ionic crystal are often transparent) may be applied, too.

Other Methods

- Since most properties of crystals are structure sensitive, many more methods exist that give some information about point defects. In what follows we give a list of some tools (which might be elaborated upon in due time):
  - **Deep level transient spectroscopy** (***DLTS***). This is a standard method for the investigation of impurity atoms in semiconductors.
  - **Electron spin resonance** (***ESR***)
  - **Infra red spectroscopy** (**IR** spectroscopy); especially in the form of **Fourier-transform IR-spectroscopy** (**FTIR**). The method of choice to investigate **O** and **C** in **Si**.

5. Dislocations

5.1 Basics

5.1.1 Burgers and Line Vector

5.1.2 Volterra Construction and Consequences

5.1.3 Essentials to Chapter 5.1: Dislocations - Basics

5.2 Elasticity Theory, Energy , and Forces

5.2.1 General Remarks and Basics of Elasticity Theory

5.2.2 Stress Field of a Straight Dislocation

5.2.3 Energy of a Dislocation

5.2.4 Forces on Dislocations

5.2.5 Interactions Between Dislocations

5.2.6 Essentials to 5.2: Dislocations - Elasticity Theory, Energy , and Forces

5.3 Movement and Generation of Dislocations

5.3.1 Kinks and Jogs

5.3.2 Generation of Dislocations

5.3.3 Climb of Dislocations

5.3.4 Essentials to 5.3: Movement and Generation of Dislocations

5.4 Partial Dislocations and Stacking Faults

5.4.1 Stacking Faults and Close Packed Lattices

5.4.2 Dislocation Reactions Involving Partial Dislocations

5.4.3 Some Dislocation Details for Specific Lattices

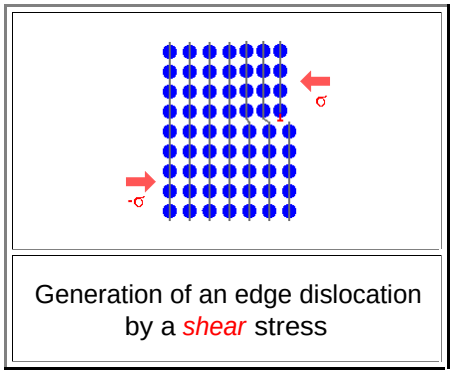
5.4.4 Essentials to 5.4 Partial Dislocations and Stacking Faults

5. Dislocations

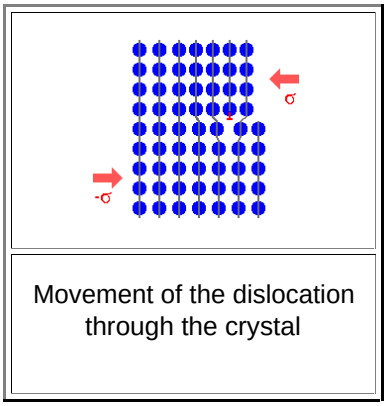
5.1 Basics

5.1.1 Burgers and Line Vector

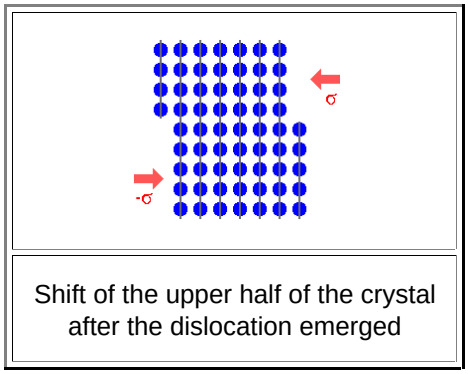
- The ***smelting*** and ***forging*** of metals marks the beginning of civilization - the *art* of working metals was for thousands of years the major "high tech" industry of our ancestors.
  - Trial and error over this period of time lead to an astonishing degree of perfection, as can be seen all around us and in many museums. In the state museum of **Schleswig-Holstein** in **Schleswig**, you may admire the **damascene blades** of our **Viking** ancestors.
  - Two kinds of iron or steel were welded together and forged into a sword in an extremely complicated way; the process took several weeks of an expert smith's time. All this toil was necessary if you wanted a sword with better properties than those of the ingredients. The damascene technology, shrouded in mystery, was needed because the vikings didn't know a thing about defects in crystals - exactly like the Romans, Greek, Japanese (india) Indians, and everbody else in those times.
  - You might enjoy finding and browsing through several modules to this topic which are provided "on the side" in this Hyperscript.
- Exactly *why* metals could be plastically deformed, and *why* the plastic deformation properties could be changed to a very large degree by forging (and magic?) without changing the chemical composition, was a mystery for thousands of years.
  - No explanation was offered before **1934**, when **Taylor**, **Orowan** and **Polyani** discovered (or invented?) independently the **dislocation**.
  - A few years before (**1929**), U. **Dehlinger** (who, around **1969** tried to teach me basic mechanics) almost got there, he postulated so-called "*Verhakungen*" as lattice defects which were supposed to mediate plastic deformation - and they were almost, but not quite, the real thing.
- It is a shame up to the present day that the discovery of the basic scientific principles governing metallurgy, still the most important technology of mankind, did not merit a **Nobel prize** - but after the war everything that happened in science before or during the war was eclipsed by the atomic bomb and the euphoria of a radiantly beautiful nuclear future. The link pays tribute to some of the men who were instrumental in solving one of the oldest scientific puzzles of mankind.
- Dislocations can be perceived easily in some (mostly two-dimensional) structural pictures on an atomic scale. They are usually introduced and thought of as extra lattice planes inserted in the crystal that do not extend through all of the crystal, but end in the dislocation line.
  - This is shown in the schematic **three-dimensional view** of an edge dislocations in a cubic primitive lattice. This beautiful picture (from Read?) shows the inserted half-plane very clearly; it serves as the quintessential illustration of what an *edge* dislocation looks like.
- Look at the picture and try to grasp the concept. *But don't forget*
  - **1.** There is *no such crystal* in nature: All real lattices are more complicated - either not cubic primitive or with more than one atom in the base.
  - **2.** The *exact structure* of the dislocation will be more complicated. *Edge* dislocations are just an extreme form of the possible dislocation structures, and in most real crystals would be split into "partial" dislocations and look much more complicated.
- We therefore must introduce a more general and necessarily more abstract definition of what constitutes a dislocation. Before we do that, however, we will continue to look at some properties of (edge) dislocations in the simplified atomistic view, so we can appreciate some elementary properties.
  - *First*, we look at a simplified but principally correct rendering of the connection between **dislocation movement** and **plastic deformation** - the elementary process of metal working which contains all the ingredients for a complete solution of all the riddles and magic of the smith's art.



Generation of an edge dislocation by a *shear* stress



Movement of the dislocation through the crystal



Shift of the upper half of the crystal after the dislocation emerged

- This sequence can be seen **animated** in the link

This calls for a little exercise

Exercise 5.1-1

Find the mistakes

What the picture illustrates is a simple, but far-reaching truth:

Plastic deformation proceeds - atomic step by atomic step - by the

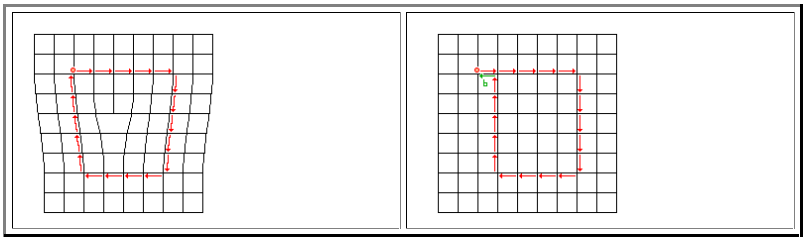
generation and movement of dislocations

- The whole art of forging consists simply of manipulating the *density* of dislocations, and, more important, their *ability of moving* through the lattice.
- After a dislocation has passed through a crystal *and* left it, the lattice is complely restored, and no traces of the dislocation is left in the lattice. Parts of the crystal are now shifted in the plane of the movement of the dislocation (left picture). This has an interesting consquence: *Without dislocations, there can be no elastic stresses whatsoever in a single crystal!* (discarding the small and very localized stress fields around point defects).

We already know enough by now, to deduce some elementary properties of dislocations *which must be generally valid*.

- **1.** A dislocation is **one-dimensional defect** because the lattice is *only* disturbed along the **dislocation line** (apart from small elastic deformations which we do not count as defects farther away from the core). The dislocation line thus can be described at any point by a **line vector *t(x,y,z)***.
- **2.** In the **dislocation core** the bonds between atoms are *not* in an equilibrium configuration, i.e. at their minimum enthalpy value; they are heavily distorted. The dislocation thus must possess *energy* (per unit of length) and *entropy*.
- **3.** Dislocations *move* under the influence of external forces which cause internal stress in a crystal. The area swept by the movement defines a plane, the **glide plane**, which always (by definition) contains the dislocation line vector.
- **4.** The movement of a dislocation *moves the whole crystal* on one side of the glide plane relative to the other side.
- **5.** (Edge) dislocations could (in principle) be generated by the *agglomeration of point defects*: self-interstitial on the extra half-plane, or vacancies on the missing half-plane.

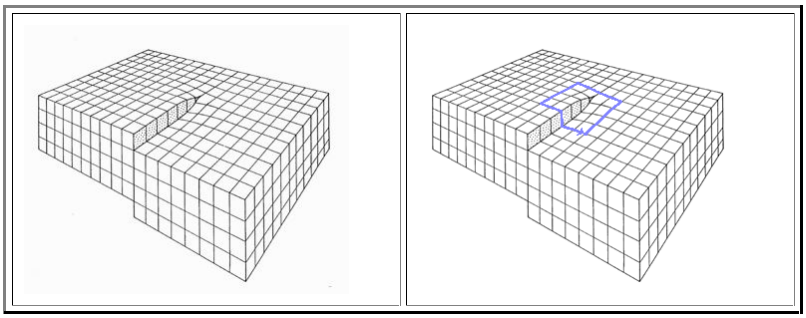
Now we add a new property. The fundamental quantity defining an arbitrary dislocation is its **Burgers vector *b***. Its atomistic definition follows from a **Burgers circuit** around the dislocation in the real crystal, which is illustrated below



- *Left picture*: Make a closed circuit that encloses the dislocation from lattice point to lattice point (later from atom to atom). You obtain a closed chain of the base vectors which define the lattice.
- *Right picture*: Make exactly the same chain of base vectors in a perfect reference lattice. *It will not close*.

- The special vector needed for closing the circuit in the reference crystal is *by definition* the *Burgers vector* ***b***.

- It follows that the *Burgers vector* of a (perfect) dislocation is of necessity a *lattice vector*. (We will see later that there are exceptions, hence the qualifier "perfect").
- *But beware!* As always with conventions, you may pick the *sign* of the Burgers vector at will.
  - In the version given here (which is the usual definition), the closed circuit is around the dislocation, the Burgers vector then appears in the reference crystal.
  - You could, of course, use a closed circuit in the reference crystal and define the Burgers vector around the dislocation. You also have to define if you go clock-wise or counter clock-wise around your circle. You will always get the same vector, but the sign will be different! And the sign is very important for calculations! So whatever you do, *stay consistent!*. In the picture above we went clock-wise in both cases.
- Now we go on and learn a new thing: There is a second *basic* type of dislocation, called **screw dislocation**. Its atomistic representation is somewhat more difficult to draw - but a Burgers circuit is still possible:



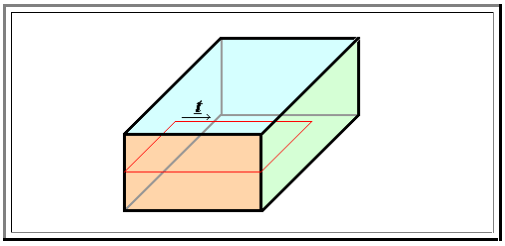
  - You notice that here we chose to go *clock-wise* - for no particularly good reason
- If you imagine a walk along the non-closed Burges circuit, which you keep continuing round and round, it becomes obvious how a **screw dislocation** got its name.
  - It also should be clear by now how Burgers circuits are done.
  - But now we will turn to a more formal description of dislocations that will include *all possible cases*, not just the extreme cases of pure edge or screw dislocations.

Exercise 5.1-3

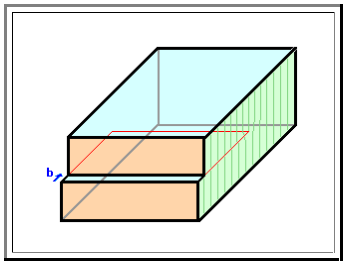
Quick Questions to 5.1

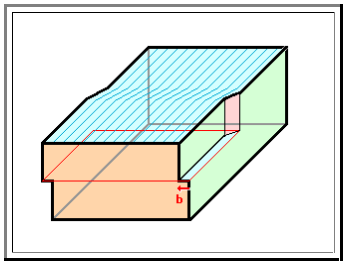
5.1.2 Volterra Construction and Consequences

- We now generalize the present view of dislocations as follows:
  - **1.** Dislocation lines may be *arbitrarily curved* - never mind that we cannot, at the present, easily imagine the atomic picture to that.
  - **2.** *All lattice vectors* can be Burgers vectors, and as we will see later, even vectors that are *not* lattice vectors are possible. A general definition that encloses all cases is needed.
- As ever so often, the basic ingredients needed for "making" dislocations existed before dislocations in crystals were conceived. **Volterra**, coming from the mechanics of the continuum (even crystals haven't been discovered yet), had defined all possible basic deformation cases of a continuum (including crystals) and in those elementary deformation cases the basic definition for dislocations was already contained!
  - The link shows Volterra' basic deformation modes - three can be seen to produce *edge* dislocations in crystals, one generates a *screw* dislocation.
  - Three more cases produce defects called "**disclinations**". While of theoretical interest, disclinations do not really occur in "normal" crystals, but in more unusual circumstances (e.g. in the two-dimensional lattice of flux lines in superconductors) and we will not treat them here.
- Volterra's insight gives us the tool to define dislocations in a very general way. For this we invent a little contraption that helps to imagine things: the "**Volterra knife**", which has the property that you can make any conceivable cut into a crystal with ease (in your mind). So lets produce dislocations with the Volterra knife:
- **1.** *Make a cut*, any cut, into the crystal using the Volterra knife.
  - The cut is always defined by some *plane* inside the crystal (here the plane indicated by he red lines).
  - The cut does not have to be on a flat plane, but we also do not gain much by making it "warped". The picture shows a flat cut, mainly just because it is easier to draw.
  - The cut is by necessity completely contained within a *closed line*, the *red cut line* (most of it on the outside of the crystal).
  - That part of the cut *line* that is *inside* the crystal will define the line vector ***t*** of the dislocation to be formed.

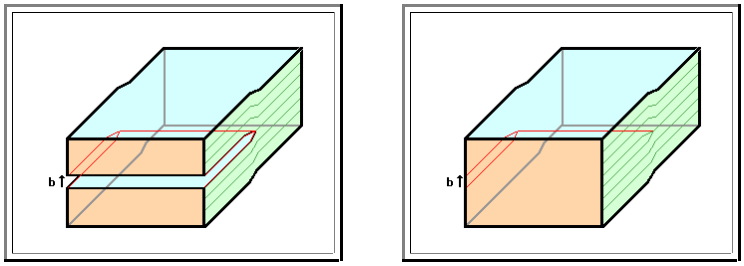


- **2.** *Move the two parts of the crystal* separated by the cut relative to each other by a *translation vector of the lattice*; allowing elastic deformation of the lattice in the region around the dislocation line.
  - The translation vector chosen will be the *Burgers vector* ***b*** of the dislocation to be formed. The sign will depend on the convention used. Shown are movements leading to an edge dislocations (left) and a screw dislocation (right).

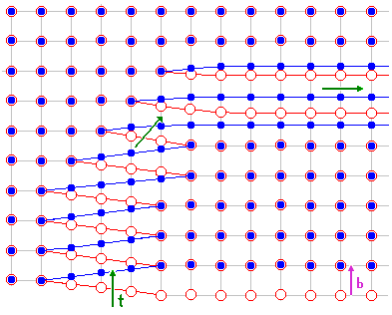




- ***3.*** *Fill in material* or take some out, if necessary.
  - This will *always* be necessary for obvious reasons whenever your chosen translation vector has a component perpendicular to the plane of the cut.
  - Shown is the case where you have to fill in material - always preserving the structure of the crystal that was cut, of course.
  - *Left*: After cut and movement. *Right*: After filling up the gap with crystal material.

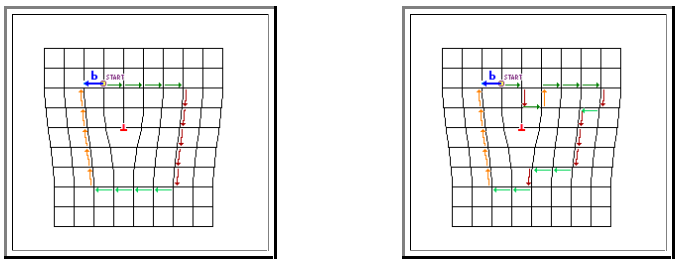


- **4.** *Restore the crystal* by "welding" together the surfaces of the cut.
  - Since the displacement vector was a *translation vector of the lattice*, the surfaces will fit together *perfectly* everywhere - except in the region around the dislocation line defined as by the cut line.
- A one-dimensional defect was produced, defined by the *cut line* (= line vector ***t*** of the dislocation) and the *displacement vector* which we call **Burgers vector *b***.
  - It is rather obvious (but not yet proven) that the Burgers vector defined in this way is identical to the one defined before. This will become totally clear in the following paragraphs.
- From the Volterra construction of a dislocation, we can not only obtain the simple edge and screw dislocation that we already know, but *any* dislocation. Moreover, from the Volterra construction we can immediately deduce a new list with more properties of dislocations:
  - **1.** The *Burgers vector* for a given dislocation is always the same, i.e. it does not change with coordinates, because there is only *one* displacement for every cut. On the other hand, the *line vector* may be different at every point because we can make the cut as complicated as we like.
  - **2.** Edge- and screw dislocations (with an angle of **90°** or **0°**, resp., between the Burgers- and the line vector) are just *special cases* of the general case of a **mixed dislocation**, which has an arbitrary angle between **b** and **t** that may even change along the dislocation line. The illustration shows the case of a curved dislocation that changes from a pure edge dislcation to a pure screw dislocation.



We are looking at the plane of the cut (sort of a semicircle centered
in the lower left corner). Blue circles denote atoms just below, red circles atoms just above the cut. Up on the right the dislocation is
a pure edge dislocation, on the lower left it is pure screw. In between
it is mixed. In the link this dislocation is shown moving in an animated illustration.

  - **3.** The Burgers vector must be independent from the precise way the Burgers circuit is done since the Volterra construction does not contain any specific rules for a circuit. This is easy to see, of course:



Old circuit

Two arbitrary alternative Burgers circuits.
The colors serve to make it easier
to keep track of the steps.

- **4.** A dislocation *cannot end* in the interior of an otherwise perfect crystal (try to make a cut that ends internally with your Volterra knife), but only at
  - a crystal surface
  - an internal surface or interface (e.g. a grain boundary)
  - a **dislocation knot**
  - **on itself** - forming a **dislocation loop**.
- **5.** If you do not have to add matter or to take matter away (i.e. involve interstitials or vacancies), the Burgers vector **b** *must be in the plane of the cut* which has two consequences:
  - The cut plane must be planar; it is defined by the line vector and the Burgers vector.
  - The cut plane is the *glide plane* of the dislocation; only in this plane can it move without the help of interstitials or vacancies.

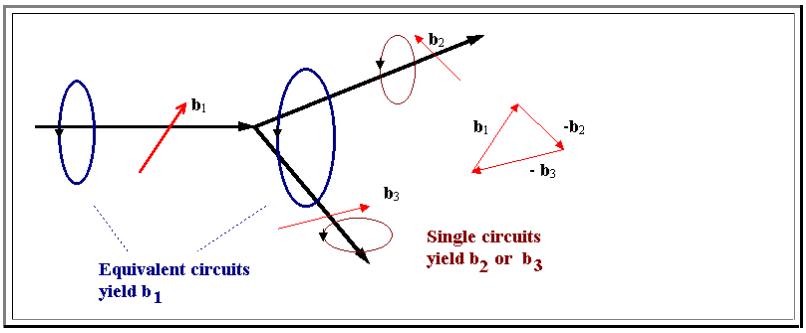
  .
- The glide plane is thus the plane spread out by the Burgers vector ***b*** *and the line vector* ***t***.
- **6.** Plastic deformation is promoted by the movement of dislocations in glide planes. This is easy to see: Extending your cut produces more deformation and this is identical to moving the dislocation!
- **7.** The magnitude of ***b*** (**= *b***) is a measure for the **"strength"** of the dislocation, or the amount of elastic deformation in the core of the dislocation.

A not so obvious, but very important consequence of the Voltaterra definition is

- **8.** At a **dislocation knot** the *sum of all Burgers vectors is zero*, $\Sigma \underline{b} = 0$, provided all line vectors point into the knot or out of it. A dislocation knot is simply a point where three or more dislocations meet. A knot can be constructed with the Volterra knife as shown below.

Statement **8**. can be proved in two ways: Doing Burgers circuits or using the Voltaterra construction twice. At the same time we prove the equivalence of obtaining ***b*** from a Burgers circuit or from a Voltaterra construction.

- Lets look at a dislocation knot formed by three arbitrary dislocations and do the Burgers circuit - always taking the direction of the Burgers circuit from a "right hand" rule



- Since the sum of the two individual circuits must give the same result as the single "big" circuit, it follows:

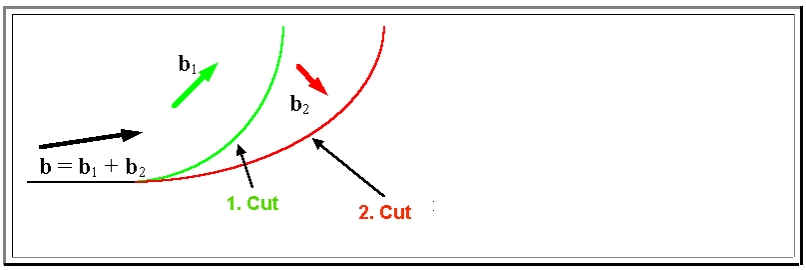
$$\underline{b}_1 = \underline{b}_2 + \underline{b}_3$$

- Or, more generally, after reorienting all ***t*** -vectors so that they point into the knot:

$$\Sigma_i \underline{b}_i = 0$$

Now lets look at the same situation in the Voltaterra construction:

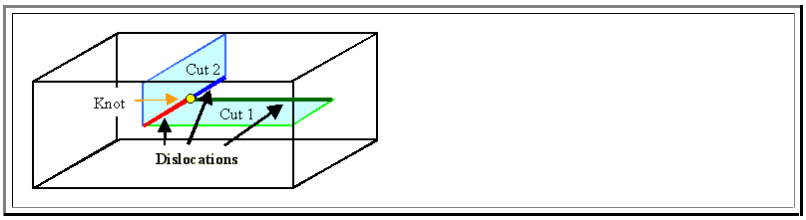
- We make a first cut with a Burgers vector $\underline{b}_1$ (the green one in the illustration below).
- Now we make a second cut in the same plane that extends partially beyond the first one with Burgers vector $\underline{b}_2$ (the red line). We have three different kinds of boundary lines: red and green where the cut lines are distinguishable, and black where they are on top of each other. And we have also produced a dislocation knot!



- Obviously the displacement vector for the black line, which is the Burgers vector of that dislocation, must be the sum of the two Burgers vectors defined by the two cuts: ***b*** = ***b*****1** + ***b*****2**. So we get the same result, because our line vectors all had the same "flow" direction (which, in this case, is actually tied to which part of the crystal we move and which one we keep "at rest").

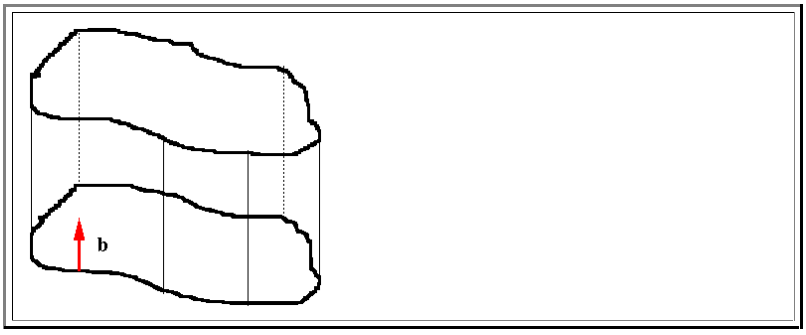
If we produce a dislocation knot by two cuts that are *not* coplanar but keep the Burgers vector on the cut plane, we produce a knot between dislocations that do not have the same glide plane. As an immediate consequence we realize that this knot might be *immobile* - it cannot move.

- A simple example is shown below (consider that the Burgers vector of the red dislocation may have a glide plane different from the two cut planes because it is given by the (vector) sum of the two original Burgers vectors!).



We can now draw some conclusion about how dislocations must behave in circumstances not so easy to see directly:

- Lets look at the glide plane of a *dislocation loop*. We can easily produce a loop with the Volterra knife by keeping the cut totally inside the crystal (with a *real* knife that could not be done). In the example the dislocation is an edge dislocation.
- The glide plane, always defined by Burgers and line vector, becomes a **glide cylinder**! The dislocation loop can move up or down on it, but no lateral movement is possible.



- What would the glide plane of a screw dislocation loop look like? Well there is no such thing as a *screw dislocation loop* - you figure that one out for yourself!
- A pure (straight) *screw* dislocation has no particular glide plane since ***b*** and ***t*** are parallel and thus do not define a plane. A screw dislocation could therefore (in principle) move on any plane. We will see later why there are still some restrictions.

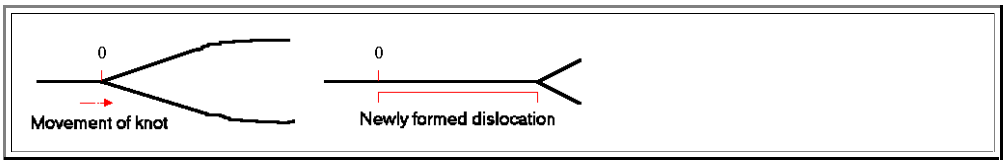
This leaves the touchy issue of the sign convention for the line vector ***t***. *This is important*! The sign of the line vector determines the sign of the Burgers vector, and the Burgers vector, including sign, is what you will use for many calculations. This is so because for a Burgers circuit you must define if you go clockwise or counter-clockwise around the line vector, using the right-hand convention. *We will go clockwise*!

- The easiest way of dealing with this is to remember that the sum of the Burgers vectors must be zero if all line vectors either point into the knot or away from it.
- As long as only three dislocations meet at one point, there is no big problem in being consistent in the choice of line vector and Burgers vectors, once you started assigning signs for the line vectors, you can throw in the Burgers vector. There is however no principal restriction to only three dislocations meeting at one point; in this case the situation is not always unambiguous; we will deal with that later. This is not as easy as it seems. We will do a little exercise for that.
- Last we define: The circuit is to close around the dislocation; the circuit in the reference crystal then defines the Burgers vector.

Exercise 5.1-2

Sign of Burgers- and Line Vectors

- We see that one can get pretty far with the purely geometric consideration of dislocations following a Voltaterra kind of construction. But some questions with respect to properties allowed by the Volterra construction remain open if we pose them for *real* crystals :
  - Are there real knots where **4**, **5**, **6**, or even more dislocations meet? We sure can produce them with the knife.
  - Are there really dislocations with all kinds of translation vectors, e.g. ***b*** **= *a*<100>** or ***b*** **= *a*<123>**? They are all allowed.
  - Is the geometry of a network arbitrary, i.e. are the angles between dislocations in a knot arbitrary?
  - Are real dislocations really arbitrarily curved?
- Then there are questions to which the Volterra construction has nothing to say in the first place:
  - What determines dislocation reactions, e.g. the formation of a new dislocation? A very simple reactions takes place, for example, whenever a knot moves as shown in the illustration below.



  - Do dislocations repel or attract each other? Or, more generally: How do they interact with other defects including point defects, other dislocations, grain boundaries, precipitates and so on?
- To be able to answer these questions, we have to consider the *elastic energy* of a dislocation; we will do this in the next chapter.

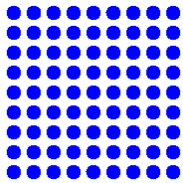
Exercise 5.1-3

Quick Questions to 5.1

5.1.3 Essentials to Chapter 5.1: Dislocations - Basics

Plastic deformation of crystals=movement of dislocations through the crystal.

- The distortion necessary to deform a crystal is localized in a **1**-dimensional defect=dislocation that moves through the crystal under the influence of (external shear) forces.
- "Discovery" of dislocations as source of plastic deformation=answer to one of the biggest and oldest scientific puzzles in **1934** (Taylor, Orowan and Polyani). No Noble prize!
- Movement of dislocations produce steps of atomic size characterized by a vector called "Burgers vector".
- Movement of dislocations occurs in a plane (=glide plane) and shifts the upper part of the crystal with respect to the lower part.

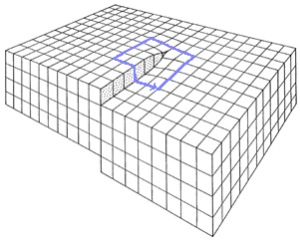


Dislocations are characterized by

1. Their Burgers vector ***b***=

- vector describing the step obtained after a dislocation passed through the crystal.
- Vector obtained by a Burgers circuit around a dislocation.
- Translation vector in the Volterra procedure.

- All definitions of ***b*** give identical results for a given dislocations; but watch out for sign conventions!
- By definition, ***b*** is always a translation vector ***T*** of the lattice.
- For energetic reasons ***b*** is usually the shortest translation vector of the lattice; e.g. ***b*=a/2** ⟨**110**⟩for the **fcc** lattice.



Burgers circuit for a screw dislocation

2. Their line vector ***t*(*x*,*y*,*z*)** describing the direction of the dislocation line in the lattice

- ***t*(*x*,*y*,*z*)** is an arbitrary (unit) vector in principle but often a prominent lattice direction in reality
- While the dislocation can be curved in any way, it tends to be straight (=shortest possible distance) for energetic reasons.

The glide plane by necessity must contain ***t*(*x*,*y*,*z*)** and ***b*** and is thus defined by the two vectors

- The angle α between ***t*(*x*,*y*,*z*)** and ***b*** determines the character or kind of dislocation:
- Note that any plane containing ***t*** is a glide plane for a screw dislocation.

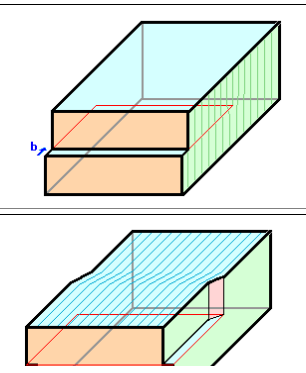
- **α=90 °**: Edge dislocation.
- **α=0 °**: Screw dislocation.
- **α=60 °**: "Sixty degree" dislocation.
- α=arbitrary : "Mixed" dislocation.

Dislocations have a large line energy E_{dis} per length and therefore are never thermal equilibrium defects

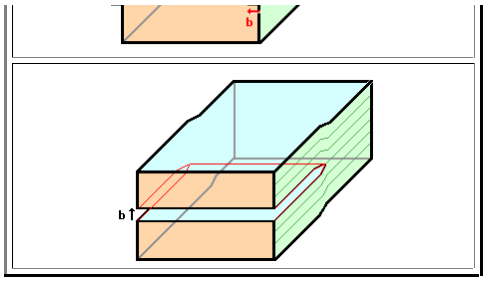
$E_{dis} \approx 5\ eV/|b|$

The formal Volterra definition of dislocations is very useful and extendable to more complex kinds of dislocations

- Cut into the *lattice* with a fictitious "Volterra" knife (make a plane cut to keep it easy), The cut line is a closed loop by necessity. The part of the cut line inside the crystal identifies the dislocation line.
- Move the part of the *lattice* above (or below - attention, signs change!!) the cut plane by an arbitrary lattice translation vector=Burgers vector of the dislocation. Add or remove *lattice* points as necessary (=remove or fill in atoms in the *crystal* going with the lattice).

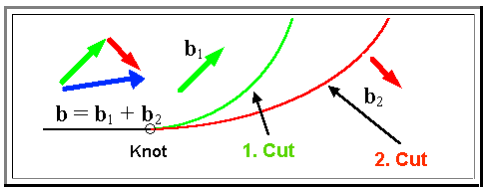


- Mend the *lattice* (or *crystal)* by "welding the upper part to the lower one. There will be a perfect fit by definition everywhere except along the dislocation line.
- Make the best arrangement of the atoms along the dislocation line by minimizing their energy (make best possible bonds).



- You now have formed a dislocation.
  - The procedure is "easily" extended to dislocations in ***n***-dim. lattices, to (special) dislocations with a Burgers vector not defined as translation vector of the lattice and to lattices more complex than a simple crystal lattice.

- Direct consequences are:
  - A dislocation cannot just end in the interior of a crystal
  - There is a "knot rule" for dislocation knots:
    S$\underline{b}$=0
    provided the signs of the line vectors follow a convention (all pointing to or away from the knot)
  - There can be all kinds of *dislocation loops* (just confine your fictitious cut to the lattice interior!)



- **Note:** "Simple" geometric considerations allow to deduce a lot about properties of dislocations

5.2 Elasticity Theory, Energy , and Forces

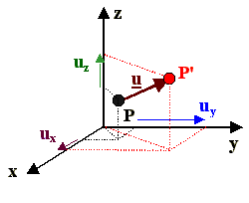
5.2.1 General Remarks and Basics of Elasticity Theory

General Remarks

- The **theory of elasticity** is quite difficult just for simple homogeneous media (no crystal), and even more difficult for crystals with dislocations - because the dislocation core cannot be treated with the linear approximations always used when the math gets tough.
  - Moreover, *relatively simple* analytical solutions for e.g. the elastic energy stored in the displacement field of a dislocation, are only obtained for an *infinite* crystal, but then often lead to infinities.
  - As an example, the energy of *one* dislocation in an otherwise perfect infinite crystal comes out to be infinite!
- This looks not very promising. However, for *practical* purposes, very simple relations can be obtained in good approximations.
  - This is especially true for the *energy per unit length*, the **line energy** of a dislocation, and for the forces between dislocations, or between dislocations and other defects.
- A very good introduction into the elasticity theory as applied to dislocation is given in the text book **Introduction to Dislocations** of **D. Hull** and **D. J. Bacon**. We will essentially follow the presentation in this book.
- The atoms in a crystal containing a dislocation are displaced from their perfect lattice sites, and the resulting **distortion** produces a **displacement field** in the crystal around the dislocation.
  - If there is a ***displacement*** field, we automatically have a **stress field** and a **strain field**, too. Try not to mix up displacement, ***stress*** and ***strain***!
- If we look at the picture of the edge dislocation, we see that the region above the inserted half-plane is in **compression** - the distance between the atoms is smaller then in equilibrium; the region below the half-plane is in **tension**.
  - The dislocation is therefore a source of **internal stress** in the crystal. In all regions of the crystal except right at the dislocation core, the stress is small enough to be treated by conventional *linear* elasticity theory. Moreover, it is generally sufficient to use *isotropic* theory, simplifying things even more.
  - If we know what is called the **elastic field**, i.e. the *relative* displacement of all atoms, we can calculate the force that a dislocation exerts on other dislocations, or, more generally, any interaction with elastic fields from other defects or from external forces. We also can then calculate the energy contained in the elastic field produced by a dislocation.

Basics of Elasticity Theory

- The first element of elasticity theory is to define the **displacement field** $\underline{u}(x,y,z)$, where $\underline{u}$ is a vector that defines the displacement of atoms or, since we essentially consider a continuum, the displacement of any point **P** in a strained body from its original (unstrained) position to the position **P'** in the strained state.



Displacement of ***P*** to ***P'***
by displacement vector $\underline{u}$

  - The displacement vector $\underline{u}(x, y, z)$ is then given by

$$\underline{u}(x, y, z) = \begin{matrix} u_x(x, y, z) \\ u_y(x, y, z) \\ u_z(x, y, z)] \end{matrix}$$

  - The components u_x , u_y , u_z represent projections of $\underline{u}$ on the ***x***, ***y***, ***z*** axes, as shown above.
- The vector field $\underline{u}$, however, contains not only uninteresting rigid body translations, but at some point **(*x,y,z*)** all the summed up displacements from the other parts of the body.
  - If, for example, a long rod is just elongated along the ***x***-axis, the resulting $\underline{u}$ field, if we neglect the contraction, would be

$$u_x = \text{const} \cdot x \qquad u_y = 0 \qquad u_z = 0$$

- But we are only interested in the *local* deformation, i.e. the deformation that acts on a volume element **d*V*** *after* it has been displaced some amount defined by the environment. In other words, we only are interested in the changes of the *shape* of a volume element that was a perfect cube in the undisplaced state. In the example above, *all* volume element cubes would deform into a rectangular block.

We thus resort to the local **strain** ϵ, defined by the nine components of the strain tensor acting on an elementary cube. That this is true for small strains you can prove for yourself in the next exercise.

- Applied to our case, the nine components of the **strain tensor** are directly given in terms of the first derivatives of the displacement components. If you are not sure about this, activate the link.
- We obtain the **normal strain** as the diagonal elements of the strain tensor.

$$\epsilon_{xx} = \frac{du_x}{dx} \qquad \epsilon_{yy} = \frac{du_y}{dy} \qquad \epsilon_{zz} = \frac{du_z}{dz}$$

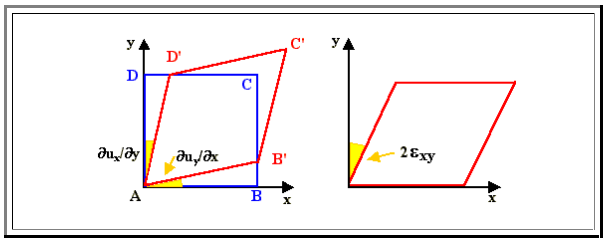
- The **shear strains** are contained in the rest of the tensor:

$$\epsilon_{yz} = \epsilon_{zy} = \tfrac{1}{2} \cdot \left(\frac{du_y}{dz} + \frac{du_z}{dy} \right)$$

$$\epsilon_{zx} = \epsilon_{xz} = \tfrac{1}{2} \cdot \left(\frac{du_z}{dx} + \frac{du_x}{dz} \right)$$

$$\epsilon_{xy} = \epsilon_{yx} = \tfrac{1}{2} \cdot \left(\frac{du_x}{dy} + \frac{du_y}{dx} \right)$$

Within our basic assumption of linear theory, the magnitude of these components is **<< 1**. The normal strains simply represent the fractional change in length of elements parallel to the ***x***, ***y***, and ***z*** axes respectively. The physical meaning of the shear strains is shown in the following illustration



- A small area element ***ABCD*** in the ***xy*** plane has been strained to the shape ***A B' C' D'*** without change of area. The angle between the sides ***AB*** and ***AD***, initially parallel to ***x*** and ***y***, respectively, has decreased by $2\epsilon_{xy}$. By rotating, but not deforming, the element as shown on the right-hand side, it is seen that the element has undergone a simple shear. The simple shear strain often used in engineering practice is $2\epsilon_{xy}$, as indicated.
- The volume ***V*** of a small volume element is changed by strain to

$$V + \Delta V = V \cdot (1 + \epsilon_{xx}) \cdot (1 + \epsilon_{yy}) \cdot (1 + \epsilon_{zz})$$

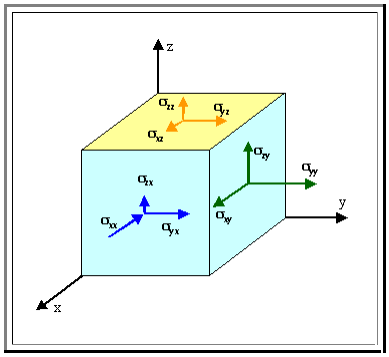
- The fractional change in volume Δ, known as the *dilatation*, is therefore

$$\Delta = \frac{\Delta V}{V} = \epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz}$$

- Note that Δ is independent of the orientation of the axes ***x, y, z***

Simple elasticity theory links the strain experienced in a volume element to the forces acting on this element. (*Difficult* elasticity theory links the strain experienced in any volume element to the forces acting on the macroscopic body!). The forces act on the surface of the element and are expressed as **stress**, i.e. as force per area. Stress is propagated in a solid because each volume element acts on its neighbors.

- A complete description of the stresses acting therefore requires not only specification of the magnitude and direction of the force but also of the orientation of the surface, for as the orientation changes so, in general, does the force.
- Consequently, nine components must be defined to specify the state of stress. This is shown in the illustration below.



- We have a volume element, here a little cube, with the components of the stress shown as vectors
- Since on any surface an arbitrary force vector can be applied, we decompose it into **3** vectors at right angles to each other. Since we want to keep the volume element at rest (no translation and no rotation), the sum of all forces and moments must be zero, which leaves us with **6** independent components.
- The stress vectors on the other **3** sides are exactly the opposites of the vectors shown
- A picture of the components of the strain tensor would look exactly like this, too, of course.

- The component σ_{ij}, where **i** and **j** can be ***x***, ***y***, or ***z***, is defined as the force per unit area exerted in the **+ i** direction on a face with outward normal in the **+ j** direction by the material *outside* upon the material *inside*. For a face with outward normal in the **– j** direction, i.e. the bottom and back faces in the figure above, σ_{ij} is the force per unit area exerted in the **– i** direction. For example, σ_{yz} acts in the positive ***y*** direction on the top face and the negative ***y*** direction on the bottom face.
- The six components with **i** unequal to **j** are the *shear* stresses. It is customary to abbreviate shear stresses with τ. In dislocation studies τ without an index then often represents the shear stress acting on the *slip plane* in the *slip direction* of a crystal.
- As mentioned above, by considering moments of forces taken about ***x***, ***y***, and ***z*** axes placed through the centre of the cube, it can be shown that rotational equilibrium of the element, i.e. net momentum **= 0**, requires

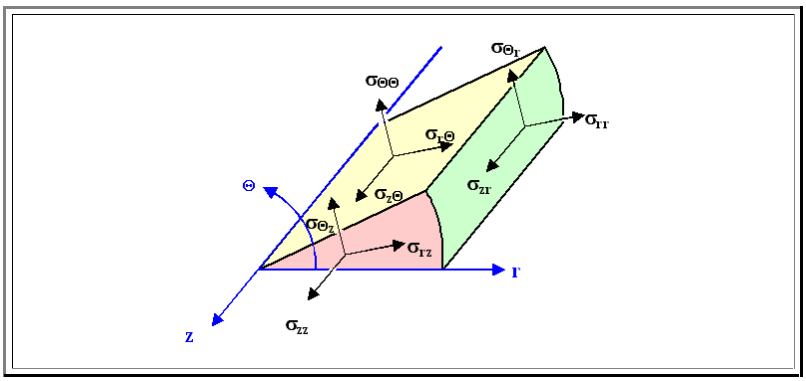
$$\tau_{zy} = \tau_{zy} \qquad \tau_{zx} = \tau_{xz} \qquad \tau_{xy} = \tau_{yx}$$

- It therefore does not matter in which order the subscripts are written.
- The three remaining components σ_{xx}, σ_{yy}, σ_{zz} are the *normal* components of the stress. From the definition given above, a *positive* normal stress results in *tension*, and a *negative* one in *compression*. We can define an **effective pressure** acting on a volume element by

$$p = -\frac{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}}{3}$$

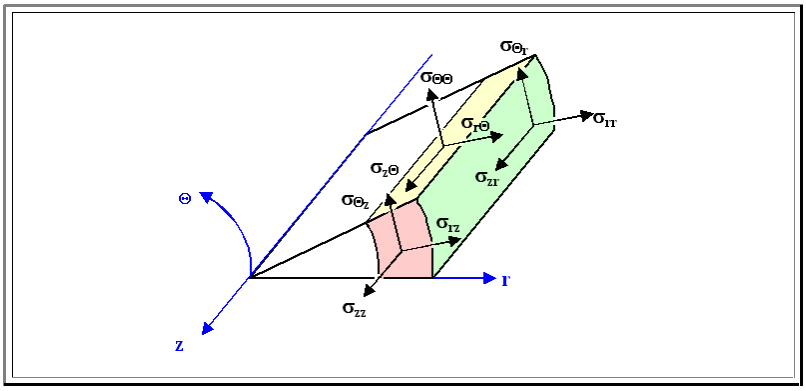
For some problems, it is more convenient to use **cylindrical polar coordinates** (***r***, θ, ***z***).

- This is shown below; the proper volume element of cylindrical polar coordinates, is essentially a "*piece of cake*".



Is it a piece of cake indeed? Well - *no*!

- The picture above is straight form the really good book of **Introduction to Dislocations** of **D. Hull** and **D. J. Bacon**, but it is a little bit wrong. But since it was only used a basic illustration, it did not produce faulty reasoning or equations.
- The picture below shows it right - think about it a bit yourself. (Hint: Imagine a situation, where you apply an uniaxial stress and try to keep your volume element in place)



- The stresses are defined as shown above; the stress $\sigma_{z,\,\theta}$, e.g., is the stress in ***z***-direction on the θ plane. The second subscript **j** thus denotes the plane or face of the "slice of cake" volume element that is perpendicular to the axis denote by the sunscript - as in the cartesian coordinate system above. The yellow plane or face thus is the θ plane, green corresponds to the ***r*** plane and pink denotes the ***z*** plane

Material Laws Relating Stress and Strain

In the simplest approximation (which is almost always good enough) *the relation between stress and strain is taken to be linear*, as in most "material laws" (take, e.g. "**Ohm's** law", or the relation between electrical field and polarization expressed by the dielectric constant); it is called "**Hooke's law**".

- Each strain component is linearly proportional to each stress component; in full generality for *anisotropic* media we have, e.g.

$$\epsilon_{11} = a_{11}\sigma_{11} + a_{22}\sigma_{22} + a_{33}\sigma_{33} + a_{12}\sigma_{12} + a_{13}\sigma_{13} + a_{23}\sigma_{23}$$

- For symmetry reasons, not all a_{ij} are independent; but even in the worst case (i.e. triclinic lattice) only **21** independent components remain.
- For *isotropic* solids, however, only *two* independent a_{ij} remain and Hooke's law can be written as

$$\sigma_{xx} = 2G \cdot \epsilon_{xx} + \lambda \cdot (\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz})$$

$$\sigma_{yy} = 2G \cdot \epsilon_{yy} + \lambda \cdot (\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz})$$

$$\sigma_{zz} = 2G \cdot \epsilon_{zz} + \lambda \cdot (\epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz})$$

$$\sigma_{xy} = 2G \cdot \epsilon_{xy}$$

$$\sigma_{yz} = 2G \cdot \epsilon_{xy}$$

$$\sigma_{zx} = 2G \cdot \epsilon_{zx}$$

- The two remaining material parameters λ and ***G*** are known as **Lamé constants**, but ***G*** is more commonly known as the **shear modulus**.

It is customary to use different elastic moduli, too. But for isotropic cubic crystals there are always only *two* independent constants; if you have more, some may expressed by the other ones.

- Most frequently used, and most useful are **Young's modulus**, ***Y*** , **Poisson's ratio**, ν, and the **bulk modulus**, ***K***. These moduli refer to simple deformation experiments:
- Under uniaxial, normal loading in the longitudinal direction, ***Young's modulus Y*** equals the ratio of longitudinal stress to longitudinal strain, and ***Poisson's ratio*** ν equals the negative ratio of lateral strain to longitudinal strain. The ***bulk modulus K*** is defined to be $-p/\Delta V$ (***p*** = pressure, ΔV = volume change). Since only two material parameters are required in Hooke's law, these constants are interrelated by the following equations

$$Y = 2G \cdot (1 + \nu)$$

$$\nu = \frac{\lambda}{2 \cdot (\lambda + G)}$$

$$K = \frac{Y}{3 \cdot (1 - 2\nu)}$$

- Typical values of Y and ν for metallic and ceramic solids are in the ranges $Y = (40\text{-}600)\ GNm^{-2}$ and $\nu = (0.2 - 0.45)$, respectively.

Elastic Energy

- A material under strain contains elastic energy - it is just the sum of the energy it takes to move atoms off their equilibrium position at the bottom of the potential well from the binding potential. Since energy is the sum over all displacements time the force needed for the displacement, we have:
  - **Elastic strain energy E_{el}** per unit volume = one-half the product of stress times strain for each component. The factor **1/2** comes from counting twice by taking each component. Thus, for an element of volume dV, the elastic strain energy is

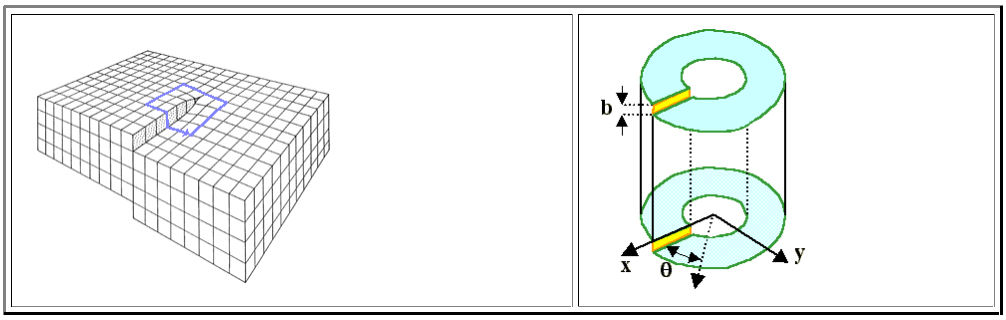
$$dE_{el} = \tfrac{1}{2} \cdot \sum_{i=x,y,z} \sum_{j=x,y,z} \sigma_{ij} \cdot \epsilon_{ij} \cdot dV$$

  - For polar or cylindrical coordinates we would get a similar formula.
- We do not actually have to calculate the energy with this formula (be grateful), but you must remember: If we have the stress field, we can calculate the strain field. If have both, we can calculate the energy.
  - And if we have the energy we have (almost) everything! Minimizing the energy gives the equilibrium configuration; gradients of the energy with respect to coordinates give forces, and so on.

5.2.2 Stress Field of a Straight Dislocation

Screw Dislocation

The elastic distortion around a straight screw dislocation of infinite length can be represented in terms of a cylinder of elastic material deformed as defined by Volterra. The following illustration shows the basic geometry.



- A screw dislocation produces the deformation shown in the left hand picture. This can be modeled by the Volterra deformation mode as shown in the right hand picture - except for the core region of the dislocation, the deformation is the same. A radial slit was cut in the cylinder parallel to the **z**-axis, and the free surfaces displaced rigidly with respect to each other by the distance ***b***, the magnitude of the Burgers vector of the screw dislocation, in the ***z***-direction.
- In the core region the strain is very large - atoms are displaced by about a lattice constant. *Linear* elasticity theory thus is not a valid approximation there, and we must exclude the core region. We then have no problem in using the Volterra approach; we just have to consider the core region separately and add it to the solutions from linear elasticity theory.

The elastic field in the dislocated cylinder can be found by *direct inspection*. First, it is noted that there are no displacements in the ***x*** and ***y*** directions, i.e. $u_x = u_y = 0$.

In the ***z***-direction, the displacement varies smoothly from **0** to ***b*** as the angle θ goes from **0** to **2π**. This can be expressed as

$$u_z = \frac{b \cdot \theta}{2\pi} = \frac{b}{2\pi} \cdot \tan^{-1}(y/x) = \frac{b}{2\pi} \cdot \arctan(y/x)$$

Using the equations for the strain we obtain the **strain field** of a *screw* dislocation:

$$\epsilon_{xx} = \epsilon_{yy} = \epsilon_{zz} = \epsilon_{xy} = \epsilon_{yx} = 0$$

$$\epsilon_{xz} = \epsilon_{zx} = -\frac{b}{4\pi} \cdot \frac{y}{x^2 + y^2} = -\frac{b}{4\pi} \cdot \frac{\sin\theta}{r}$$

$$\epsilon_{yz} = \epsilon_{zy} = \frac{b}{4\pi} \cdot \frac{x}{x^2 + y^2} = \frac{b}{4\pi} \cdot \frac{\cos\theta}{r}$$

The corresponding **stress field** is also easily obtained from the relevant equations:

$$\sigma_{xx} = \sigma_{yy} = \sigma_{zz} = \sigma_{xy} = \sigma_{yx} = 0$$

$$\sigma_{xz} = \sigma_{zx} = -\frac{G \cdot b}{2\pi} \cdot \frac{y}{x^2 + y^2} = -\frac{G \cdot b}{2\pi} \cdot \frac{\sin\theta}{r}$$

$$\sigma_{yz} = \sigma_{zy} = \frac{G \cdot b}{2\pi} \cdot \frac{x}{x^2 + y^2} = \frac{G \cdot b}{2\pi} \cdot \frac{\cos\theta}{r}$$

In cylindrical coordinates, which are clearly better matched to the situation, the **stress** can be expressed via the following relations:

$$\sigma_{rz} = \sigma_{xy}\cos\theta + \sigma_{yz}\sin\theta$$

$$\sigma_{\theta z} = -\sigma_{xz}\sin\theta + \sigma_{yz}\cos\theta$$

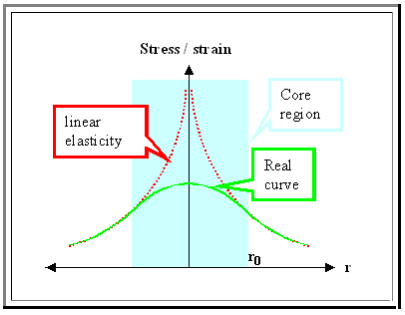
- Similar relations hold for the **strain**. We obtain the simple equations:

$$\epsilon_{\theta z} = \epsilon_{z\theta} = \frac{b}{4\pi r}$$

$$\sigma_{\theta z} = \sigma_{z\theta} = \frac{G \cdot b}{2\pi r}$$

The elastic distortion contains no tensile or compressive components and consists of pure shear. $\sigma_{z\theta}$ acts parallel to the ***z*** axis in radial planes of constant θ and $\sigma_{\theta z}$ acts in the fashion of a torque on planes normal to the axis. The field exhibits complete radial symmetry and the cut thus can be made on any radial plane θ= constant. For a dislocation of *opposite* sign, i.e. a left-handed screw, the signs of all the field components are *reversed*.

There is, however, a serious problem with these equations:



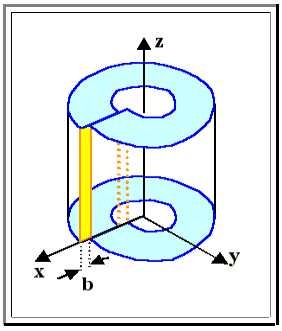
- The stresses and strains are proportional to **1/*r*** and therefore ***diverge to infinity as r → 0*** as shown in the schematic picture on the left.
- This makes no sense and therefore the cylinder used for the *calculations* must be *hollow* to avoid ***r*** - values that are too small, i.e. smaller than the core radius r_0.
- Real crystals, of course, do (usually) *not* contain hollow dislocation cores. If we want to include the dislocation core, we must do this with a more advanced theory of deformation, which means a non-linear atomistic theory. There are, however, ways to avoid this, provided one is willing to accept a bit of empirical science.
- The picture simply illustrates that strain and stress are, of course, smooth functions of ***r***. The fact that linear elasticity theory can not cope with the core, does not mean that there is a real problem.

How large is radius r_0 or the extension of the **dislocation core**? Since the theory used is only valid for small strains, we may equate the core region with the region were the strain is larger than, say, **10%**. From the equations above it is seen that the strain exceeds about **0,1** or **10%** whenever $r \approx b$. A reasonable value for the **dislocation core radius** r_0 therefore lies in the range ***b*** to **4b**, i.e. $r_0 \geq$ **1 nm** in most cases.

Edge Dislocation

The stress field of an edge dislocation is somewhat more complex than that of a screw dislocation, but can also be represented in an isotropic cylinder by the Volterra construction.

- Using the same methodology as in the case of a screw dislocation, we replace the edge dislocation by the appropriate cut in a cylinder. The displacement and strains in the ***z***-direction are zero and the deformation is basically a "plane strain".
- It is not as easy as in the case of the screw dislocation to write down the **strain field**, but the reasoning follows the same line of arguments. We simply look at the results:



$$\sigma_{xx} = -D \cdot y \frac{3x^2 + y^2}{(x^2 + y^2)^2}$$

$$\sigma_{yy} = D \cdot y \frac{x^2 - y^2}{(x^2 + y^2)^2}$$

$$\sigma_{xy} = \sigma_{yx} = D \cdot x \frac{x^2 - y^2}{(x^2 + y^2)^2}$$

$$\sigma_{zz} = \nu \cdot (\sigma_{xx} + \sigma_{yy})$$

$$\sigma_{zz} = \sigma_{zx} = \sigma_{yz} = \sigma_{zy} = 0$$

- We used the abbreviation $D = Gb / 2\pi (1 - \nu)$.

- The stress field has, therefore, both dilational and shear components. The largest normal stress is σ_{xx} which acts parallel to the Burgers vector. Since the slip plane can be defined as $y=0$, the maximum compressive stress (σ_{xx} is negative) acts immediately above the slip plane and the maximum tensile stress (σ_{xx} is positive) acts immediately below the slip plane.
  - The effective pressure (given by the sum over the normal components of the stress) is

$$p = \frac{2 \cdot (1 + \nu) \cdot D}{3} \cdot \frac{y}{x^2 + y^2}$$

  - We thus have compressive stress above the slip plane and tensile stresses below - just as deduced from the qualitative picture of an edge dislocation; graphical representation of the stress field of an edge dislocation is shown in the link.

- For edge dislocations (and screw dislocations too), the sign of the stress- and strain components *reverses* if the sign of the Burgers vector is reversed.
  - Again, we have to leave out the dislocation core; the core radius again can be taken to be about **1*b* - 4*b***

- We are left with the case of a **mixed dislocation**. This is not a problem anymore. Since we have a linear isotropic theory, we can just take the solutions for the edge- and screw **component** of the mixed dislocation and superimpose, i.e. add them.
  - As far as "simple" elasticity theory goes, we now have everything we can obtain. If better descriptions are needed, the matter becomes extremely complicated! But thankfully, this simple description is sufficient for most applications.

5.2.3 Energy of a Dislocation

With the results of the elasticity theory we can get approximate formulas for the **line energy** of a dislocation and the elastic interaction with other defects, i.e. the forces acting on dislocations.

- The energy of a dislocation comes from the elastic part that is contained in the elastically strained bonds outside the radius r_0 *and* from the energy stored in the core, which is of course energy sitting in the distorted bonds, too, but is not amenable to elasticity theory.
- The *total* energy per unit length E_{ul} is the sum of the energy contained in the elastic field, E_{el}, and the energy in the core, E_{core}.

$$E_{ul} = E_{el} + E_{core}$$

- *Do not confuse energies **E** with Youngs modulus **Y** which is often (possibly here) written as **E**, too! From the context it is always clear what is meant.*
- Using the formula for the strain energy for a volume element given before, integration over the total volume will give the total elastic energy E_{el} of the dislocation. The integration is easily done for the screw dislocation; in what follows the equations are always *normalized to a unit of length*.

$$dE_{el}(\text{screw}) = \pi \cdot r \cdot dr \cdot (\sigma_{\Theta z} \cdot \epsilon_{\Theta z} + \sigma_{z\Theta} \cdot \epsilon_{z\Theta}) = 4\pi \cdot r \cdot dr \cdot G \cdot (\epsilon_{\Theta z})^2$$

$$E_{el}(\text{screw}) = \frac{G \cdot b^2}{4\pi} \cdot \int_{r_0}^{R} \frac{dr}{r} = \frac{G \cdot b^2}{4\pi} \cdot \ln \frac{R}{r_0}$$

- The integration runs from r_0, the core radius of the dislocation to R, which is some *as yet undetermined* external radius of the elastic cylinder containing the dislocation. In principle, R should go to infinity, but this is not sensible as we are going to see.

The integration for the edge dislocation is much more difficult to do, but the result is rather simple, too:

$$E_{el}(\text{edge}) = \frac{G \cdot b^2}{4\pi(1 - \nu)} \cdot \ln \frac{R}{r_0}$$

- So, apart from the factor $(1 - \nu)$, this is the same result as for the screw dislocation.

Let us examine these equations. There are a number of interesting properties; moreover, we will see that there are very simple approximations to be gained:

1. *The total energy **U** of a dislocation is proportional to its length **L**.*

$$U = E_{ul} \cdot L = L \cdot (E_{el} + E_{core})$$

- Since we always have the principle of minimal energy (entropy does not play a role in this case), we can draw a important conclusion:
- *A dislocation tends to be straight* between its two "end points" (usually dislocation knots). That is a *first* rule about the direction a dislocation likes to assume.

2. *The line energy of an edge dislocation is always larger than that of a screw dislocation* since $(1 - \nu) < 1$. With $\nu \approx 1/3$, we have $E_{screw} \approx 0{,}66 \cdot E_{edge}$.

- This means that a dislocation tends to have as large a screw component as possible. This is a *second* rule about the direction a dislocation likes to assume *which may be in contradiction* to the first one. It is quite possible that a dislocations needs to zig-zag to have as much screw character as geometrially allowed - it then cannot be straight at the same time.

3. *The elastic part of the energy depends (logarithmically) on the crystal size* (expressed in R), *for an infinite crystal it is infinite (∞)!* Does this make any sense?

- Of course this doesn't make sense. Infinite crystals, however, do not make sense either. And in finite crystals, even in *big* finite crystals, the energy is finite!

- Moreover, in most real crystals it is *not* the outer dimension that counts, but the size of the grains which are usually quite small. In addition, if there are many dislocations with different signs of the Burgers vector, their strain fields will (on average) tend to cancel each other. So for *practical* cases we have a finite energy.

4. *The elastic part of the energy also depends (logarithmically) on the core radius* r_0.

5. *The energy is a weak function of the crystal (or grain size)* R.

- Taking an extremely small value for r_0, e.g. **0,1 nm**, we obtain for $\ln(R/r_0)$ an extreme range of **20,7 - 2,3** if we pick extreme values for R of **100 mm** or **1 nm**, respectively. A more realistic range for R would be **100 µm - 10 nm**, giving $\ln(R/r_0) = 13{,}8 - 4{,}6$. Grain size variations, within reasonable limits, thus only provide a factor of **2 - 3** for energy variations.

We will now deduce an approximation for the line energy that is sufficiently good for most purposes.

- We equate r_0 with the magnitude of the Burgers vector, $|\underline{b}|$. This makes sense because the Burgers vector is a direct measure of the "strength" of a dislocation, i.e. the strength of the displacement in the core region.
- We need a value for the energy in the core of the dislocations, which so far we have not dealt with. Since there is no easy way of calculating that energy, we could equate it in a first approximation with the energy of melting. That would make sense because the dislocation core is comparable in its degree of distortion to the liquid state. More sophisticated approaches end up with the best **simple** value:

$$E_{core} = \frac{G \cdot b^2}{2\pi}$$

- There are, however, other approaches, too.
- The total energy for $r_0 = b$ then becomes

$$E_{tot} = \frac{G \cdot b^2}{4\pi} \cdot \left(\ln \frac{R}{b} + 2 \right)$$

More generally, the following formula is often used

$$E_{tot} = \frac{G \cdot b^2}{4\pi(1-\nu)} \cdot \left(\ln \frac{R}{b} + B \right)$$

- with B = pure number best approximating the core energy of the particular case. Often $B = 1$ is chosen, leading to

$$E_{tot} = \frac{G \cdot b^2}{4\pi(1-\nu)} \cdot \left(\ln \frac{R}{b} + 1 \right)$$

$$= \frac{G \cdot b^2}{4\pi(1-\nu)} \cdot \left(\ln \frac{e \cdot R}{b} \right)$$

- For the last equation bear in mind that $\ln(e) = 1$.

The **ln** term is not very important. To give an example; it is exactly 4π for $e \cdot R = 3{,}88 \times 10^4 |b|$, i.e. for $R \approx$ **5 µm**; the total energy in this case would be $E_{tot} = 2G \cdot b^2$.

- In a very general way we can write

$$E_{tot} = \alpha \cdot G \cdot b^2$$

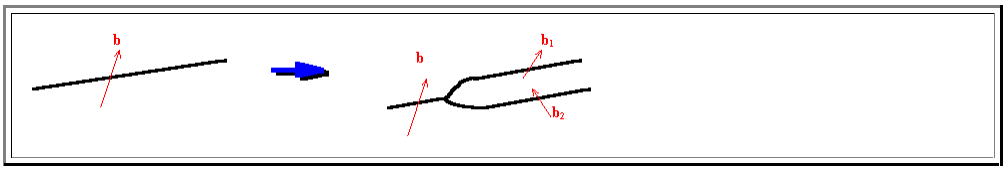
- And α (from measurements) is found to be $\alpha \approx$ **1,5 ....0,5**. If we do not care for factors in the order of unity, we get the final *very simple formula* for the **line energy** of a dislocation

$$E_{total} \approx G \cdot b^2$$

- With this expression for the line energy of a dislocation, we can deduce more properties of dislocations.
- **6.** *Dislocations always tend to have the smallest possible Burgers vector*.
  - Since for Burgers vectors $\underline{\boldsymbol{b}}_i$ larger than the smallest translation vector of the lattice and thus expressible by $\underline{\boldsymbol{b}}_i = \underline{\boldsymbol{b}}_1 + \underline{\boldsymbol{b}}_2$; $\underline{\boldsymbol{b}}_{1,2}$ = some shorter vectors of the lattice, we always have

$$(b_1)^2 + (b_2)^2 < b^2$$

  - A splitting into smaller Burgers vectors is therefore always energetically favorable.
- There are therefore no dislocations with large Burgers vectors!
  - If a dislocation would have a large Burgers vector, it would immediately split into two (or more) dislocations with smaller Burgers vectors.
  - This is *always* possible, because in the Volterra construction you can always replace *one* cut with the translation vector $\underline{\boldsymbol{b}}$ by *two* cuts with $\underline{\boldsymbol{b}}_1$ and $\underline{\boldsymbol{b}}_2$ so that $\underline{\boldsymbol{b}} = \underline{\boldsymbol{b}}_1 + \underline{\boldsymbol{b}}_2$.



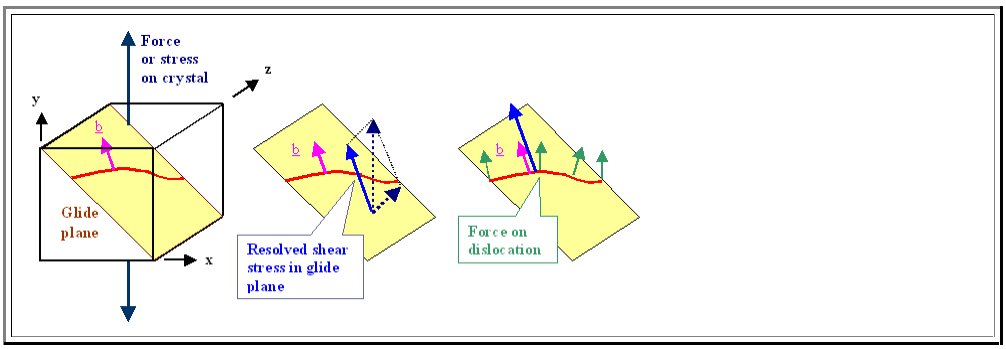
- **7.** *The line energy is in the order of **5 eV** per Burgers vector.*.
  - This makes dislocations automatically non-equilibrium defects. They will not come into being out of nothing like point defects for free enthalpy reasons.
- **8.** *The line energy (= energy per length) has the same dimension as a a force*, it expresses a **line tension**, i.e. a force in the direction of the line vector which tries to shorten the dislocation.
  - It actually *is* a force, as we can see from the definition of such a line tension $\boldsymbol{F}$:

$$F = -\frac{dU}{dL}$$

  - We thus may imagine a dislocation as a stretched rubber band, which tries to be as short as possible. But one should be careful not to overreach this analogy. Ask yourself: What keeps dislocations loops stable?

5.2.4 Forces on Dislocations

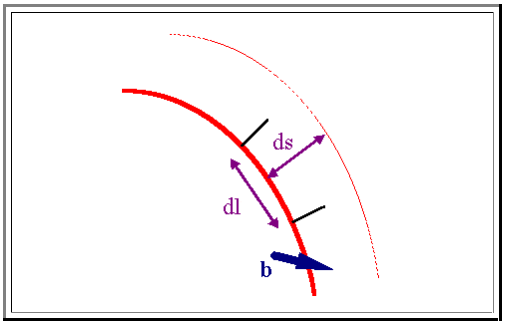
- Again, detailed calculations are complicated and must be done numerically in most cases. For practical usage, however, we will find simple approximations by using the energy formula already derived; this will be good enough for most cases.
  - First, we have to see that for the movement of a dislocation on its glide plane, we only need to consider the *shear stress on this plane*. This is so, because only force components lying *in* the glide plane of the dislocation can have any effect on dislocation motion in the glide plan. The normal components of the the stress in the glide plane system act perpendicular to the glide plane and thus will not contribute to the dislocation movement.
  - Both shear stress components in the glide plane act on the dislocation. Important, however, is only their combined effect in the direction of the Burgers vector, which is called the **resolved shear stress τ_{res}**; for simplicity it will just be called τ from now on.
  - However, while the resolved shear stress points into the direction of the Burgers vector, *the direction of the force component* acting on, i.e. moving the dislocation, *is always perpendicular to the line direction*! This is so because the force component in the line direction does not do anything - a dislocation cannot move in its own direction. or, if you like that better: If it would - nothing happens! The whole situation is outlined below



- Under the influence of the force ***F*** acting on the dislocation and which we want to calculate, the dislocation moves *and work* ***W*** *is done* given by ***W* = Force · distance**. Lets look at the ultimate work that can be done by moving one dislocation.
  - If the dislocation moves *in total* through the crystal on a glide plane with the area ***A***, the upper half of the crystal moves by ***b*** relative to the lower half which is the distance on which work has been done.
  - This only happens if a shear force acts on the crystal, and this force obviously does some **work *W***. This work is done bit by bit by moving the dislocation through the crystal, so we must identify the force that does work with the force ***F*** acting on the dislocation.
  - The acting shear stress in this case is then τ = Force ***F***/area ***A***. and force ***F*** is that component of the external force that is contained or "resolved" in the glide plane of the dislocation as discussed above.
- For the total work ***W*** done by moving half of the crystal a distance equal to the Burgers vector ***b*** we obtain

$$W = A \cdot \tau \cdot b$$

  - With ***A* · τ =** Force; ***b* =** Burgesvector = distance.
- We just as well can divide ***W*** into incremental steps **d*W***, the incremental work done on an incremental area that consists of an incremental piece **d*l*** of the dislocation moving an incremental distance **ds**, as shown below



  - The relation between the incremental work **d*W*** to the total work ***W*** then is just the ratio between the incremental area to the total area; we have

$$\frac{dW}{W} = \frac{ds \cdot dl}{A}$$

- Putting everything together, we obtain

$$dW = A \cdot \tau \cdot b \cdot \frac{dl \cdot ds}{A} = \tau \cdot b \cdot dl \cdot ds$$

  - An incremental piece of work **d*W*** can always be expressed as a force times an incremental distance **d*s***; i.e. **d*W* = *F* · d*s***. The force ***F*** acting on the incremental length **d*l*** of dislocation then obviously is ***F* = τ · *b* · d*l***.
  - If we now redefine the force on a dislocation slightly and refer it to the (incremental) *unit length* ***dl***, i.e. we take ***F* = F*/d*l***, we obtain a very **simple formula** for the *magnitude* of the force (*it is not a vector!*) acting on a unit length of a dislocation:

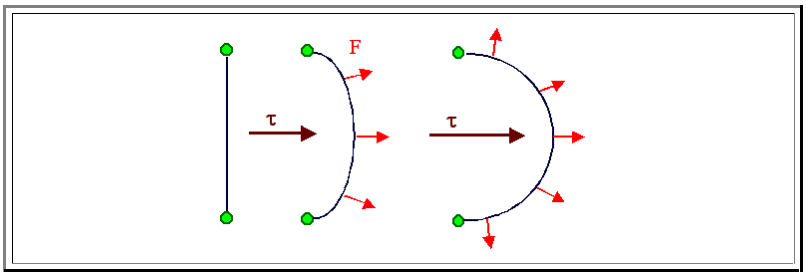
$$F^* = \tau \cdot b$$

- This is easy - but beware of the sign of the force! You must get *all the signs right* (Burgers vector, line vector, τ) to get the correct sign of the force! We also will drop the "*" in what follows, because as with all other properties of dislocations, it is automatically per unit length if not otherwise specified.
  - The important part is τ. It is the component of the shear strain in the glide plane in the direction of ***b***. This is normally not a known quantity but must be calculated, e.g. by a coordinate transformation of a given external stress tensor to a coordinate system that contains the glide plane as one of its coordinate planes.
- Again, we must realize that the Force ***F*** as defined above *is always perpendicular* to the dislocation line; even if **τ =** constant everywhere on the glide plane.
  - This is somewhat counterintuitive, but always imagine the limiting case of a pure edge and screw dislocation: The same external τ must exert a force on a screw dislocation that is perpendicular to the force on an edge dislocation to achieve the same deformation (think about that; looking at the pictures helps!)
  - This calls for a little exercise

Exercise 5.2-2

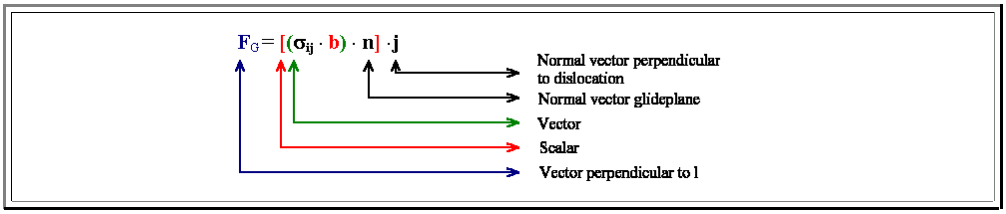
Forces on a dislocation

- In reality, dislocations can rarely move in total because they are usually firmly anchored somewhere. For a straight edge dislocation anchored at two points (e.g. at immobile dislocation knots) responding to a constant τ, we have the following situation.



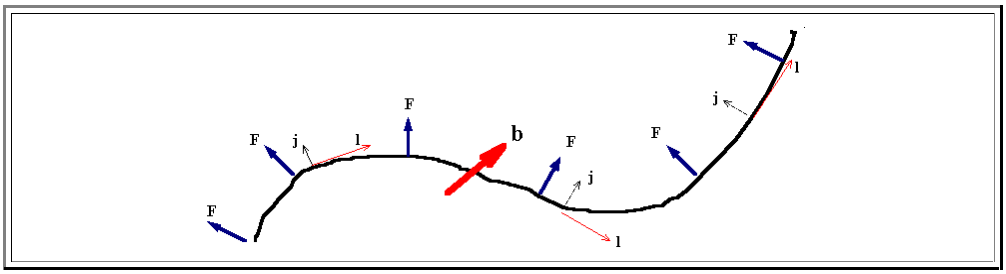
  - The forces resulting from the resolved shear stress (red arrows) will "draw out" the dislocation into a strongly curved dislocation (on the right). A mechanical equilibrium will be established as soon as the force pulling back the dislocation (its own line tension) exactly cancels the external force.
  - The middle picture shows an intermediate stage where the dislocation is still moving.
- It is possible to write down the force on a dislocation as a tensor equation which automatically takes care of the components - but this gets complicated:
  - First we need to express the force as a vector with components in the glide plane and perpendicular to it. We define ***F*** **= Force on a dislocation = (F_N, F_G)**

- With $\underline{F_G}$ = component in the glide plane, F_N = component vertical to the glide plane. Only $\underline{F_G}$ is of interest, it is given by



- Note that scalars, vectors and tensors are combined to form ultimately a vector. The colors of the brackets code the respective property as outlined in the margin.

The consequences of this equation and the quantities used are illustrated below. Also note that you have many ways to confuse signs!

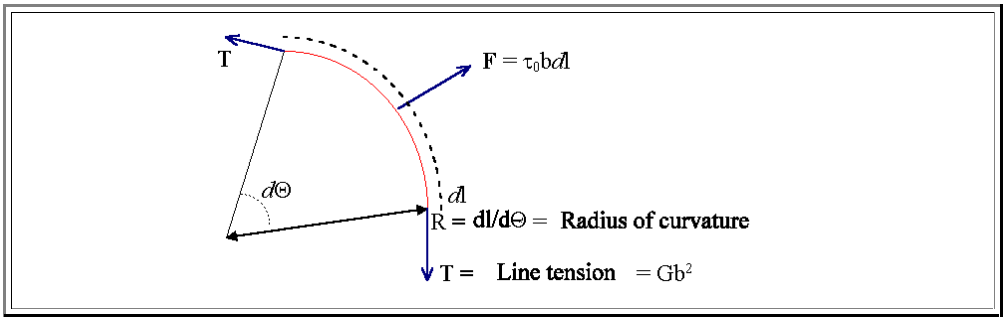


Using the formulas derived so far, we can find an important quantity, the *shear stress necessary to maintain a certain radius of curvature for a dislocation*.

- If we look at the illustration above, we see that for a certain stress, the force will draw the dislocation into a curved line, but for some configuration there will be a balance of force, because the line tension of the dislocation pulls back.

We can calculate the balance of power by looking at an incremental piece of dislocation with a radius of curvature R. The acting force $\underline{F}$ is balanced by the line tension $\underline{T}$.

- Let's assume we increase the radius R of an incremental curved piece by dl. The acting force needed for this is $\underline{F} = \tau \cdot b \cdot dl$ and we have $dl = R \cdot d\Theta$. The picture below shows a very large $d\Theta$ for clarity.



- The line tension $T = Gb^2$ is "pulling back", but only a small component T_{-F} is directly opposing $\underline{F}$.
- The component T_{-F} is given by

$$T = T \cdot \sin(d\Theta/2) \approx d\Theta/2 \quad \text{for small } d\Theta$$

- Since there are two components we have the balance of power

$$T \cdot d\Theta = Gb^2 \cdot d\Theta = \tau \cdot b \cdot dl = \tau \cdot b \cdot R \cdot d\Theta$$

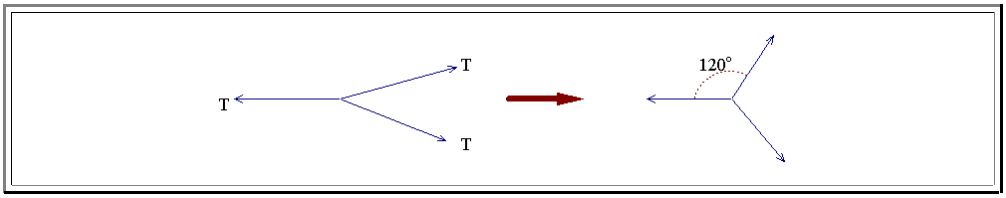
The equilibrium radius R_0 obtained for a shear stress τ_0 is thus

$$R_0 = \frac{Gb}{\tau_0}$$

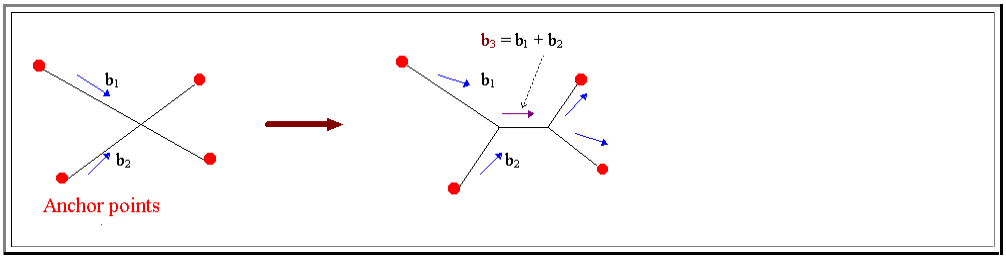
- This will have important consequences becuase the equation states that a dislocation will move "forever" if $\tau > Gb/R_{min}$ with R_{min} denoting some minimal radius of curvature that cannot be decreased anymore.

From looking at force balance, we now can answer the questions posed before for a dislocation network:

- The sum of the line tensions at a knot must be zero, too (or at least very small), otherwise the knot and the dislocations with it will move. We thus expect that **3**-knots will always show angles of (approximately) **120°**.

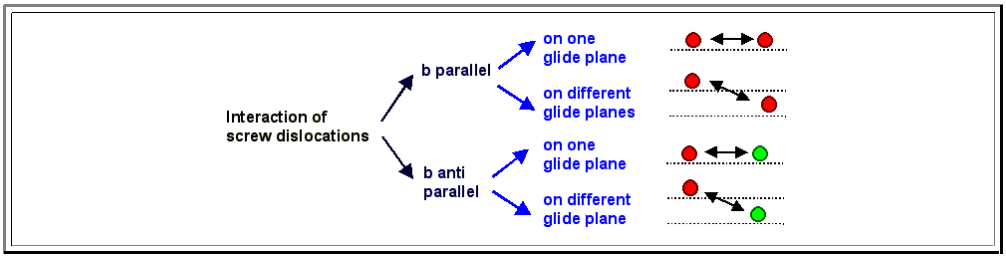


- Knots with more than three dislocations will, as a rule, split into **3**-knots, since otherwise there can be no easy balance of line tensions. In real cases, however, you must also consider the geometry of the anchor points (are they fixed, can they move?), the change of line energy with the character of the dislocation and the new total length of the dislocations.

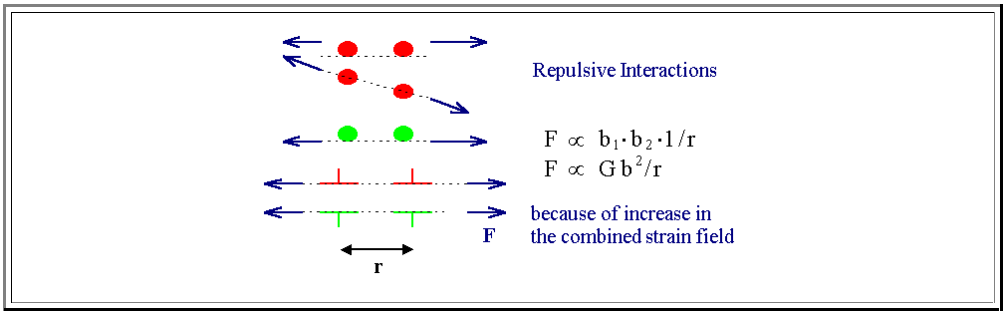


5.2.5 Interactions Between Dislocations

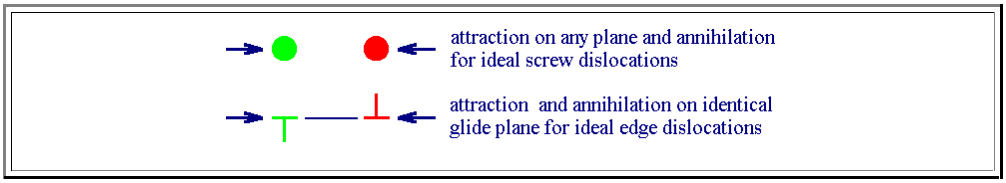
- We will first investigate the interaction between two *straight* and *parallel* dislocations of the same kind.
  - If we start with screw dislocations, we have to distinguish the following cases:



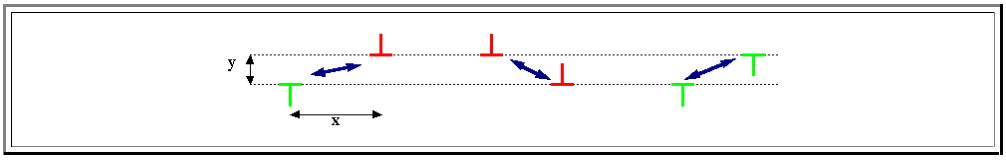
- In analogy, we next must consider the interaction of edge dislocations, of edge and screw dislocations and finally of mixed dislocations.
  - The case of mixed dislocations - the general case - will again be obtained by considering the interaction of the screw- and edge parts separately and then adding the results.
- With the formulas for the stress and strain fields of edge and screw dislocations one can calculate the resolved shear stress caused by *one* dislocation on the glide plane of the *other one* and get everything from there.
- But for just obtaining some basic rules, we can do better than that. We can classify some *basic cases* without calculating anything by just exploiting one obvious rule:
  - The superposition of the stress (or strain) fields of two dislocations that are moved toward each other can result in two basic cases:
  - **1.** The combined stress field is now *larger* than those of a single dislocation. The energy of the configuration than increases and the dislocations will *repulse* each other. That will happen if regions of compressive (or tensile) stress from one dislocation overlaps with regions of compressive (or tensile) stress from the other dislocation.
  - **2.** If the combined stress field is *lower* than that of the single dislocation, they will *attract* each other. That will happen if regions of compressive stress from one dislocation overlaps with regions of tensile stress from the other dislocation
- This leads to some simple cases (look at the stress / strain pictures if you don't see it directly)
  - **1.** Arbitrarily curved dislocations with *identical* ***b*** on the *same* glide plane will *always* repel each other.



  - **2.** Arbitrary dislocations with *opposite* ***b*** vectors on the *same* glide plane will *attract and annihilate* each other



  - *Edge dislocations* with *identical* or *opposite* Burgers vector ***b*** on *neighboring* glide planes may *attract* or *repulse* each other, depending on the precise geometry. The blue double arrows in the picture below thus may signify repulsion or attraction.

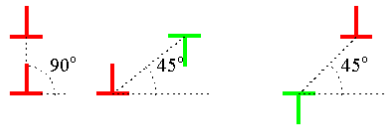


- The general formula for the forces between edge dislocations in the geometry shown above is

$$F_x = \frac{Gb^2}{2\pi(1-\nu)} \cdot \frac{x \cdot (x^2 - y^2)}{(x^2 + y^2)^2}$$

$$F_y = \frac{Gb^2}{2\pi(1-\nu)} \cdot \frac{y \cdot (3x^2 + y^2)}{(x^2 + y^2)^2}$$

- For $y = 0$, i.e. the same glide plane, we have a $1/x$ or, more generally a $1/r$ dependence of the force on the distance r between the dislocations.
- For $y < 0$ **or** $y > 0$ we find zones of repulsion and attraction. At some specific positions the force is zero - this would be the equilibrium configurations; it is shown below.
- The formula for F_y is just given for the sake of completeness. Since the dislocations can not move in y-direction, it is of little relevance so far.

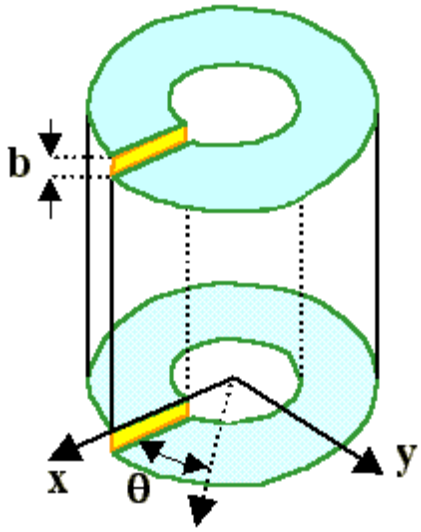


Possible equilibrium configurations

The illustration in the link gives a quantitative picture of the forces acting on one dislocation on its glide plane as a function of the distance to another dislocation.

5.2.6 Essentials to 5.2: Dislocations - Elasticity Theory, Energy , and Forces

- Around a dislacotion is a *displacement field* (= vector field) , which defines a *strain field* (= tensor field), which gives cause to a *stress field* (= tensor field) via *elastic* relations. Stress times strain give the (potential) *energy* contained in these fields and thus the *energy of a dislocation*; derivatives of energy with respect to coordinates give *forces* acting on dislocations
  - The *displacement field* $\underline{\boldsymbol{u}}(x,y,z)$ can be obtained by just looking hard at the dislocation - then write it down.
  - The rest is just Math - not all that easy, but not reyll difficult either.
- In cylinder coordinates (r, θ, z) rather simple expressions for the stress and the strain result but with the two major problems emerging as soon as we look at the energy *per unit length* of., e.g. a screw dislocation:



$$u_z = \frac{b \cdot \theta}{2\pi} = \frac{b}{2\pi} \cdot \tan^{-1}(y/x) = \frac{b}{2\pi} \cdot \arctan(y/x)$$

$$E_{el}(\text{screw}) = \frac{G \cdot b^2}{4\pi} \cdot \int_0^{\infty} \frac{dr}{r}$$

  - Both boundaries lead to infinte energy values!

- The first problem comes from overextending elastic theory, only good at small deformations, to the core region of the dislocation, the second one because the strain decreases so slowly that it is still felt far away form the dislcotion.
  - The problem gets repaired by defining an inner and outer cut-off radius r_o and R, respectively, adding some core energy E_{core}, worrying a lot if you are given to it, and finally coming out with an extermely simple, usually good enough, and very important approximation for the energy per length unit $|\underline{b}|$

$$E_{el} = \frac{G \cdot b^2}{4\pi} \cdot \int_{r_o}^{R} \frac{dr}{r} + E_{core} \approx \frac{G \cdot b^2}{4\pi(1-\nu)} \cdot \left(\ln \frac{e \cdot R}{b} \right)$$

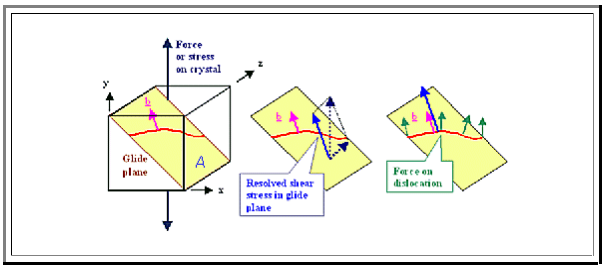
$$E_{disl} \approx Gb^2$$

  - Putting numbers into the equation gives several **eV** per unit length $|\underline{b}|$ and thus tells us that dislocations tend to be straight lines (shortest possible length!).

A dislocation moves if forces are acting on it, causing plastic deformation. In other words: work $W = F \cdot A_S$ is done if a dislocation sweeps over an area A_S.

The procedure for calculating the force is simple:

- Take the forces F acting on your crystal.
- Determine the component F_g in the glide plane of your dislocation that points in the direction of the Burgers vector $\underline{b}$.
- Calculate the resolved shear stress τ_{res} in the glide plane from the force component (= F_g/A).
- The force acting on a unit length of the disclotion is $F_{dis} = \tau \cdot b$ and is always perpendicular to the line direction $\underline{t}$

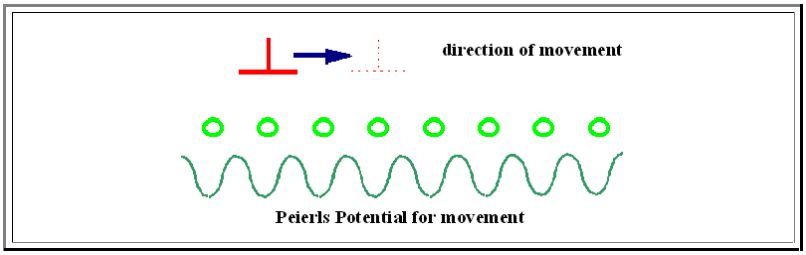


5.3 Movement and Generation of Dislocations

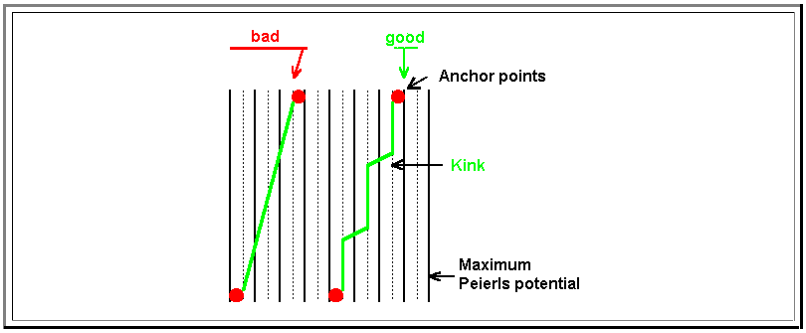
5.3.1 Kinks and Jogs

Kinks in Dislocations

- In most *crystals* and under most circumstances *there is no such thing as a straight dislocation*. Real dislocations contain **kinks** and **jogs** - sudden deviations from a straight line on atomic dimensions.
- These *defects within a defect* may strongly influence the mobility of dislocations and are thus of importance.
  - They owe their existence first of all to the fact that dislocations always "live" in a crystals - in a periodic arrangement of atoms. We have not used that fact so far, except in a rather abstract way for the very definition of dislocations with the Volterra cut. Now it is time to appreciate the effects of the crystal on the fine structure of dislocatons.
  - Kinks (and jogs) may be produced by several mechanisms, in particular they may be formed by the *movement* of the dislocation.
- First, let's look at ***kinks***. For that we first have to consider the concept of the **Peierls potential** of a dislocation.
  - Consider the movement of an edge dislocation as shown below. The green circles symbolize the last atom on the inserted half- plane of an edge dislocation. In local equilibrium its distance to the atoms to the left or right will be same for basic reasons of symmetry, cf. the perspective drawing of an edge dislocation.

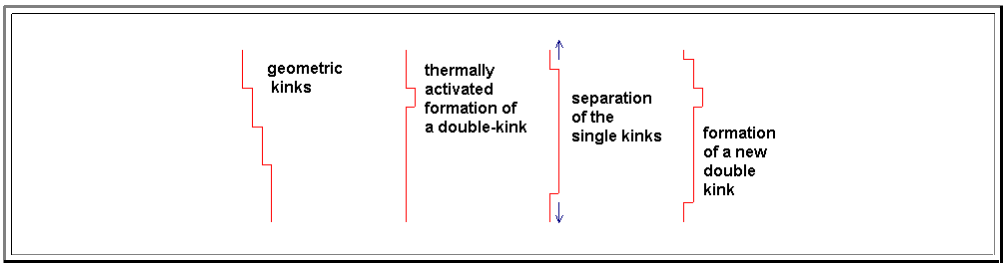


  - If the dislocation is to move, the last atom (and the ones above to some extent) has to press against the neighboring atom on one side and move away from the atom on the other side. That is clearly a situation with a higher energy which can be cast into a potential energy curve as shown in the illustration. At some point when both lattice plans are most affected, there is a maximum and a new minimum as soon as the dislocation has moved by one Burgers vector. The minima and maxima of this **Peierls potential** are along directions of high symmetry.
  - To overcome the maximum of the Peierls potential, the stress has to be larger than some intrinsic **critical shear stress** τ_{crit}.
- The Peierls potential defines special low-energy directions in which the dislocation prefers to lie. This is the *third* rule for directions that dislocations like to assume! (Try to remember the first and second rule, or use the links).
  - In other words, the inserted half-plane for the easy-to-imagine case of an edge dislocation should be clearly defined and should be in a *symmetric* position between its neighbour planes - exactly as we always have drawn it.
  - A dislocation that is almost, but not quite an edge dislocation, thus would prefer to be a pure edge dislocation over long distances and concentrate the "non-edginess" in small parts of its length as shown below. The same is true for screw dislocations, even so it is not quite as easy to contemplate.



  - The dislocation runs in the minima of the Peierls potential as long as possible and then crosses over briskly when it has to be. The transition from one Peierls minimum to the next one is called a *kink* as shown in the picture above.
- The kinks that come into existence in this way are called **geometric kinks**. But there is also a second kind, the **thermal equilibrium (double)-kink**, i.e. a crossing over to a neighboring Peierls valley followed by a "jump" back.
  - A kink, or better a double-kink, is simply a defect in an otherwise straight dislocation line, adding some energy *and* entropy. Since the **formation energy of a double kink** is not too large, they will be present in **thermal equilibrium** with concentrations following a standard Boltzmann distribution.

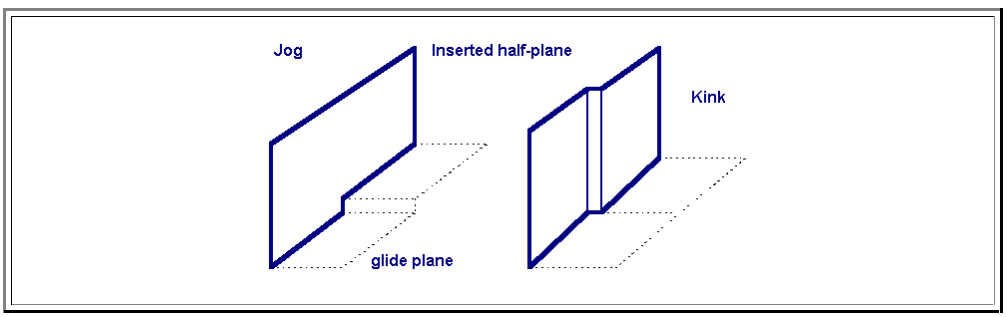
- To make that perfectly clear: While a dislocation by itself is never in thermal equilibrium, i.e. will never form spontaneously by thermal activation, this is not true for the defects it may contain. Double-kinks, seen as defects in a dislocation line, form and disappear spontaneously, if sufficient thermal energy is available; their number or density thus will follow a Boltzmann distribution.
- Once a thermal double-kink has been formed, the two single kinks may move apart; if the process is repeated, we have a *new mode of dislocation movement* for an otherwise perhaps immobile dislocation. This is shown below.



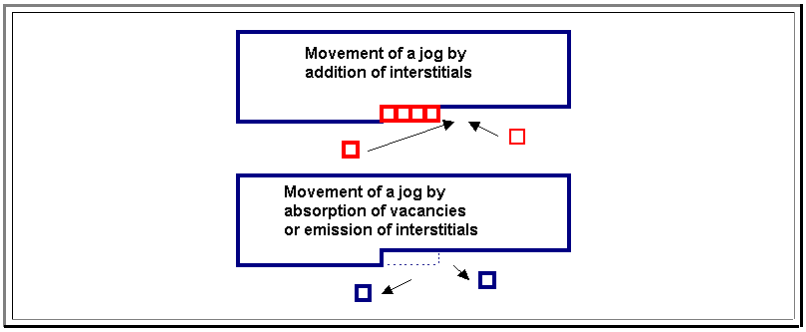
- Kinks then are *steps of atomic dimension* in the dislocation line that are fully contained in the *glide plane* of the dislocation
  - With this general definition, we can consider kinks in all dislocations, not just edge dislocations.
- Screw dislocations have a Peierls potential, too, and thus they may contain kinks. The kink, per definition, is then a very short a piece of dislocation with *edge character*.
  - This has far reaching consequences: A screw dislocation with a kink now either has a *specific glide plane* - the glide plane of the kink - or the kink is an anchor point for the screw dislocation.
  - Kinks can do more: As indicated above, at not too low temperatures when the generation of thermal double kinks becomes possible, the applied stress may be below the critical shear stress needed to move the dislocation in toto (i.e. move it across the Peierls potential), but might be large enough to separate double kinks and thus promote dislocation movement and plastic deformation. We have one of several effects here that make crystals "softer" at high temperatures.
- The best way to investigate kinks are **internal friction** experiments.
  - An oscillating deformation is chosen, e.g. by vibrating a thin specimen driven by an electromagnetic field. The amplitude and thus the internal stress and strain are easily measured. As long as the stress is not too large, deformation proceeds by the generation and the movement of kinks. This is a fully reversible process and the response to an external stress thus is purely elastic even though a dislocation moved!
  - However, in contrast to elasticity just coming from stretching the bonds between the atoms, the generation and movement of double kinks takes time and is strongly temperature dependent. Specific time constants are involved and a peculiar frequency dependence of the elastic response will be observed which contains information about the kinks.

Jogs in Dislocations

- The term "**Jogs**" is sometimes considered to be the term for all "breaks" or steps in a dislocation line with atomic dimensions. Kinks then would be a subclass of jogs with the speciality of being in the glide plane.
  - However, it is customary to use the term "jogs" for all *steps that are not contained in the glide plane*. Looking just at the inserted half-plane of an edge dislocation, jogs and kinks would look like this:



  - But remember: Jogs and kinks can occur in any dislocation, not just edge dislocations - they are just not as easily drawn!
- Jogs in edge dislocations are obviously prime places for the emission or absorption of point defects as is shown in the next illustration which looks at the inserted half-plane of an edge dislocation.



- The movement of jogs by emission or absorption of point defects means that the dislocation moves. This particular process of dislocation movement is called **climb of dislocations**. It is a movement that does *not* take place in the glide plane of the dislocation.

Generally speaking, we define:

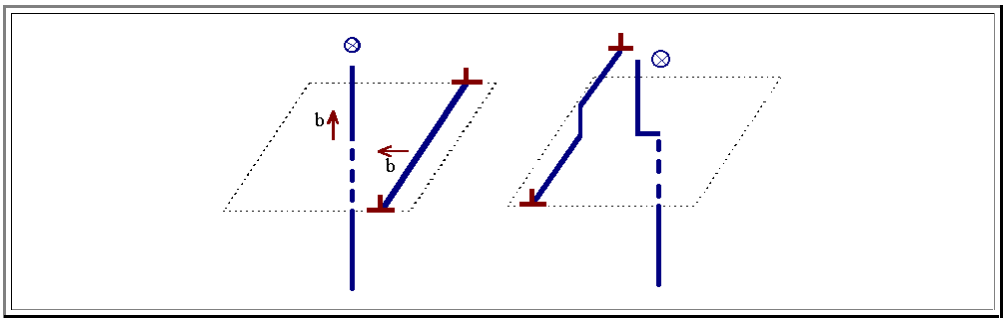
- **Conservative movement** of dislocations = movement in the glide plane = **glide** (for short) = movement without assistance of point defects.
- **Non conservative movement of dislocations** = movement not in the glide plane = **climb** (for short) = movement needing the assistance of point defects.

Generation of Kinks and Jogs

How do kinks and jogs come into existence? Three mechanisms can be identified.

- **1.** Thermally activated generation of double kinks as discussed above.
- **2.** Thermally *induced* generation of jogs by *absorption or emission of point defects*. This mechanism is thermally induced (and not "activated") because it responds to a super- or undersaturation of point defects. At large under- or supersaturations, the process becomes more likely. Here we have one of the source/sink processes needed for point defect equilibrium.
- **3.** Intersection of dislocations

The last process is new and needs some explanation. Lets look at the movement of an edge dislocation in the following geometry:



- The intersection of the edge dislocation with the screw dislocation produces one jog each per dislocation. (Consider the cut-and-move procedure and you will see why). It is clear that the same thing happens for the intersection of arbitrary mixed dislocations - a jog characterized by the Burgers vector of the dislocation that moved across will be generated.

This gives us a general relation and explains to some extent why plastic deformation is an extremely non-linear process:

- Movement of dislocations *generates* jogs.
- Jogs influence severely the *movement* of dislocations - so there is some *feedback* in the process of plastic deformation, and feedback of any kind is the hallmark of non-linear processes.

Considering jogs and kinks (together with knots), we start to consider *real* dislocations - and its getting complicated.

- And don't forget: All those great electron microscope pictures showing all kinds of dislocations, *never* show the jogs and kinks! They are simply too small. So even dislocation that look like perfect straight lines in a **TEM** picture, may be full of jogs and kinks.

There is one last property induced by these defects in a defect:

- Jogs, kinks and their combinations may produce "**debris**" left behind by a moving dislocation, because it is often "better" for dislocations to tear away from immobile parts like jogs, leaving behind a trail of point defects which in turn may agglomerate.

- If the jog is large extending over several lattice planes, a whole trail of small dislocation loops may form. The formation of a trail of vacancies in the wake of a jogged moving screw dislocation is illustrated in the link

---

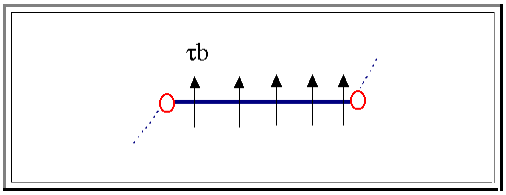
Some more text to come - but try the exercise anyway!

Exercise 5.3-1
Forces on a dislocation

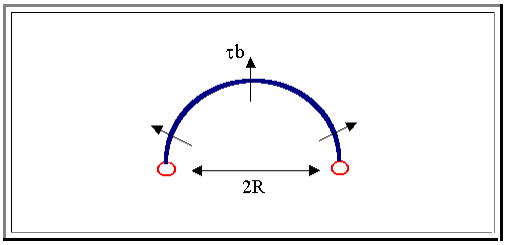
5.3.2 Generation of Dislocations

Whereas we now learned a little bit about the complications that may occur when dislocations move, we first must *have* some dislocations before plastic deformation can happen. In other words: We need **mechanisms** that **generate dislocations** in the first place!

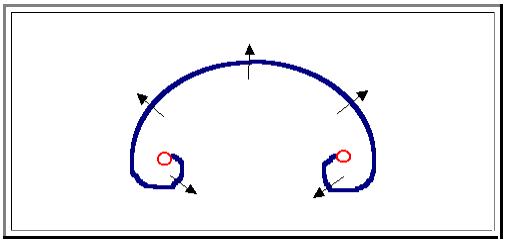
- Of course, dislocations can just be generated at the surface of the crystal; the simple pictures showing plastic deformation by an (edge) dislocation mechanism give an idea how this may happen. But more important are mechanisms that generate dislocations in the bulk of a crystal. The most important mechanism is the **Frank-Read mechanism** shown below.



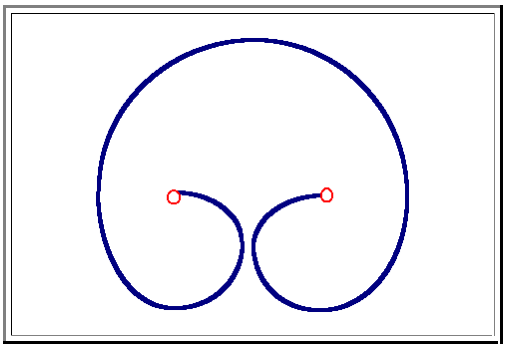
- We have a segment of dislocation firmly anchored at two points (red circles). The force $\boldsymbol{F = b \cdot \tau_{res}}$ is shown by a sequence of arrows



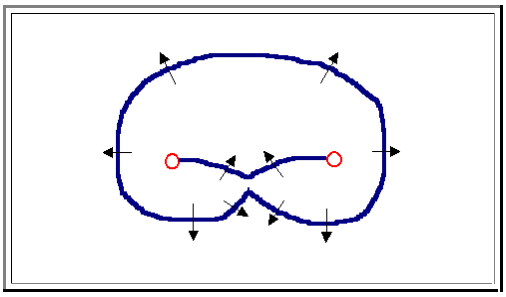
- The dislocation segment responds to the force by bowing out. If the force is large enough, the critical configuration of a semicircle may be reached. This requires a maximum shear stress of
  $\boldsymbol{\tau_{max} = Gb/R}$



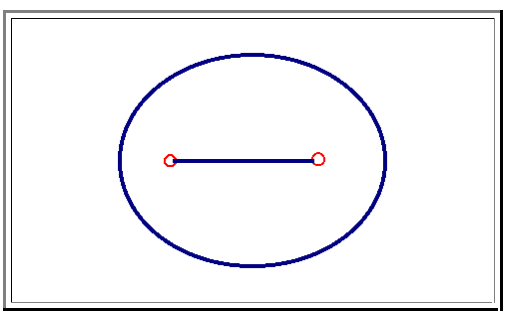
- If the shear stress is higher than $\boldsymbol{Gb/R}$, the radius of curvature is too small to stop further bowing out. The dislocation is unstable and the following process now proceeds automatically and quickly.



- The two segments shortly before they touch. Since the two line vectors at the point of contact have opposite signs (or, if you only look at the two parts almost touching: the Burgers vectors have different signs for the same line vectors), the segments in contact will annihilate each other.



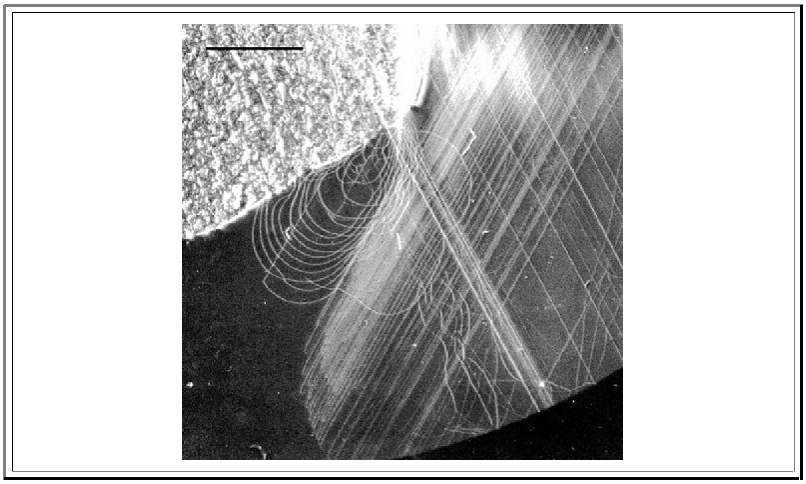
- The configuration shown is what you have immediately after contact; it is totally unstable (think of the rubber band model!). It will immediately form a straight segment and a "nice" dislocation loop which will expand under the influence of the resolved shear stress.
- The regained old segment will immediately start to go through the whole process again, and again, and again, ... - as long as the force exists. A whole sequence of nested dislocation loops will be produced.



- Stable configuration after the process. The loop is free to move, i.e. grow much larger under the applied stress. It will encounter other dislocations, form knots and become part of a network. The next loop will follow and so on - as long as there is enough shear stress.

The Frank-Read process, although looking a bit odd, will occur many times under sufficient load. It can produce any density of dislocations in short times, because the newly formed dislocations will move, become anchored at some points, and start to generate Frank-Read loops, too.

- Of course, Frank-Read dislocation sources can also be stopped - e.g. by cutting through the generating dislocation by another dislocation. We thus will have a certain finite dislocation density under certain external conditions. It may, however, depend on many parameters, including the history of the material.
- Some kind of Frank-Read mechanism may also operate from irregularities on the surface (external or internal), an example of such a source is shown in the X-ray topography below.



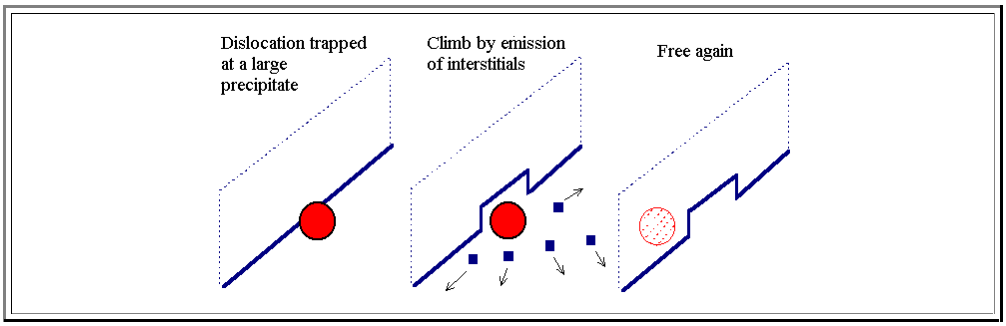
- This picture comes from the work of K.B. Kostin (a former student in Kiel) together with many others in St. Petersburg. It is a result of investigations into "wafer bonding", where to **Si** wafers are placed on top of each other and "bonded", so that a single piece of **Si** results - with a grain boundary in between. The mottled area in the upper left hand corner shows such a bonded structure, whereas the dark area containing the dislocations as white lines, remained unbonded.
- Dislocations were introduced into one of the wafers and one point on the edge of the bonded area acted as a Frank-Read source. The nested series of dislocation loops is splendidly visible. There are also lots of straight dislocations which have moved considerable distances from their point of origin.

How else can we make dislocations? Suffice it to mention that there are variants of the basic Frank-Read mechanism, too and some more exotic mechanisms. We will not go into details; the important part is that it is generally an easy process to generate many dislocations provided you already have a few to start with.

- Last but not least: "*Frank*" is not the first name of Mr. *Read* - as ever so often, two independent persons figured out this mechanism at practically the same time (in **1950**) - look up the link for details.

5.3.3 Climb of Dislocations

- As we have already seen, **dislocation climb** couples point defects and dislocations in a very direct way. This has the immediate consequence that climb processes will depend on *temperature*, because:
  - The types and concentration of equilibrium point defects are temperature dependent.
  - The supersaturation, which is the driving force for point defect reactions including ***climb***, is temperature dependent.
  - The mobility of point defects, i.e. their diffusion coefficient, is temperature dependent.
- Unfortunately for most applications, climb makes immobile dislocations mobile again (albeit they may move v e r y s l o w l y).
  - Coupled to the slow dislocation movement by climb is a slow plastic deformation with a strong temperature dependence, which would not occur without point defects - we have an **ageing mechanism**. If screws lose their tension, cables start to bow, and metals suddenly fracture after years of dutiful service, you are probably looking at the results of climb processes.
  - The major mechanism by which climb processes enable dislocations to move, is the circumvention of otherwise insurmountable obstacles, as shown below.



- Screw dislocations can climb, too, turning into a helix shape. Examples of climbed screw dislocations are provided in chapter 6

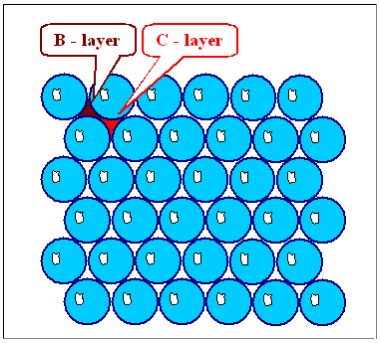
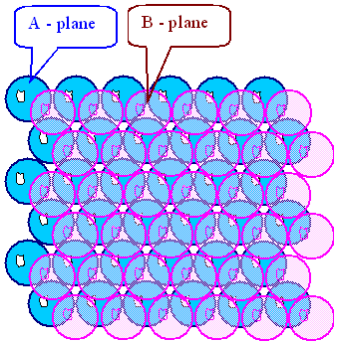
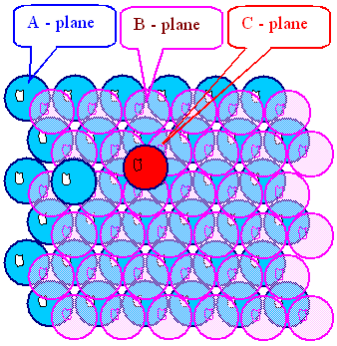
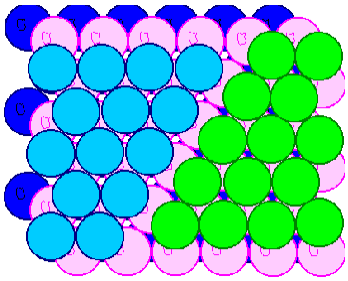
5.3.4 Essentials to 5.3: Movement and Generation of Dislocations

5.4 Partial Dislocations and Stacking Faults

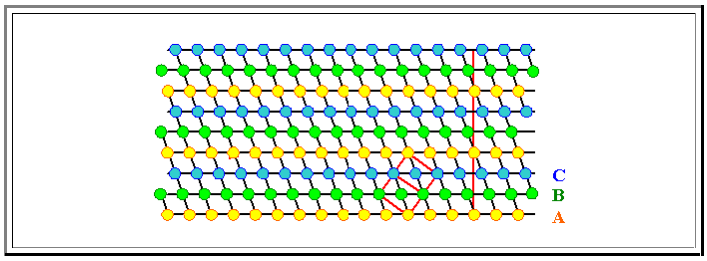
5.4.1 Stacking Faults and Close Packed Lattices

Stacking Faults and Frank Dislocations

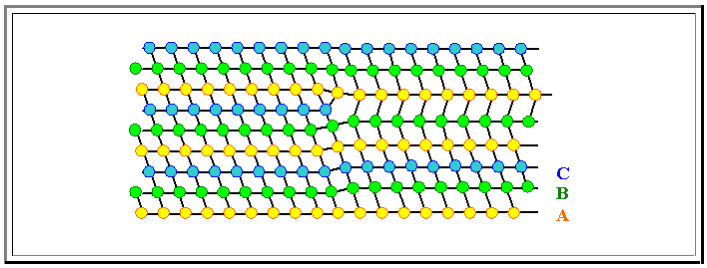
- Let's consider a close packed lattice, and look at the close packed planes.
  - In a simple model using perfect spheres we have the following situation:

<table>
<tr><td rowspan="2"> 
</td><td></td><td>We take the blue atoms as the base plane for what we are going to built on it, we will call it the "A - plane".</td></tr>
<tr><td rowspan="2"> 
</td><td>The next layer will have the center of the atoms right over the depressions of the A - plane; it could be either the B - or C - configuration.</td></tr>
<tr><td></td><td>Here the pink layer is in the "B" position</td></tr>
<tr><td rowspan="2"> 
</td><td></td><td>If you pick the B - configuration (and whatever you pick at this stage, we can always call it the B - configuration), the third layer can either be directly over the A - plane and then is also an A - plane (shown for one atom), or in the C - configuration.</td></tr>
<tr><td></td><td>If you chose "A"; you obtain the hexagonal close packed lattice (hcp), if you chose "C", you get the face centered cubic lattice (fcc)</td></tr>
<tr><td rowspan="2">  </td><td></td><td>You can't have it both ways. If you start in the C position somewhere (in the picture the green atoms) and on the A position somewhere else (light blue), you will get a problem as soon as the two layers meet.</td></tr>
<tr><td></td><td>For varieties sake, and to be able to distinguish the layers better, the bottom A layer here is in dark blue.</td></tr>
</table>

- The stacking sequences of the two close-packed lattices therefore are
  - **fcc**: **ABCABCABCA...**
  - **hcp**: **ABABABA...**
- Looking at this sequences in cross-section is a bit more involved; it is best done in a **<110>** projection of the **fcc** lattice

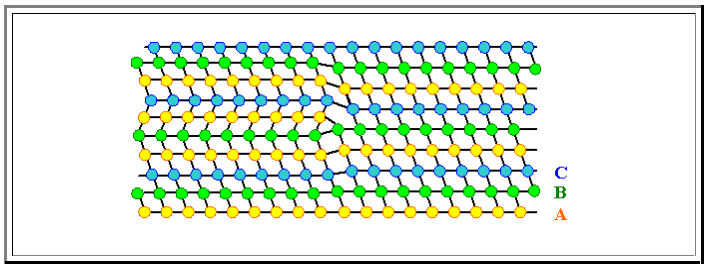


- Planes with the same letter are on lines perpendicular to the **{111}** planes, as indicated by thin black lines.
- The projection of the elementary cell is shown with red lines.
- We now remove *parts* of a horizontal **{111}** plane - e.g. by agglomeration of vacancies on that plane - it shall be a **C**-plane here.



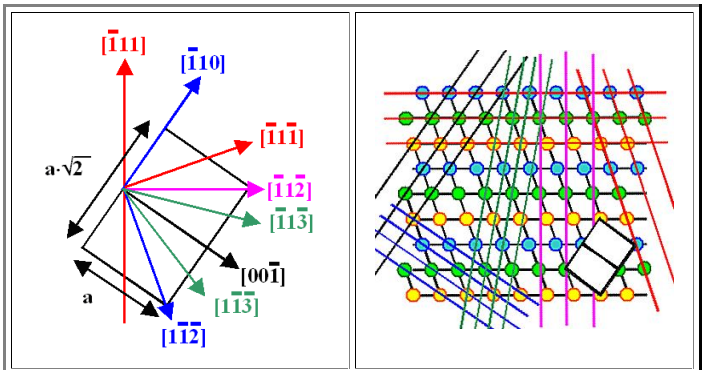
- Now **A** and **C**- planes become neighbors and relax into the configuration shown.
- We produced a **stacking fault** because the stacking sequence **ABCABCA..** has been changed to the faulty sequence **ABCABABCA...** The stacking fault is between the large letters.
- Stacking faults by themselves are simple two-dimensional defects. They carry a certain **stacking fault energy** γ; very roughly around a few **100 mJ/m^2**.
- The disc of vacancies obviously is bordered by an *edge dislocation*. What is the Burgers vector of this dislocation? We shall see farther down.

If we do not condense *vacancies* on a plane, but fill in a disc of agglomerated *interstitials*, we obtain the following structure



- The stacking sequence **ABCABCA...** again is faulty; it is now **ABCABACABCA...** . The stacking fault is between the large letters.
- This is a *different kind of stacking fault* than the one from above.
- For historical reasons, we call the stacking fault produced by vacancy agglomeration "**intrinsic stacking fault**" and the stacking fault produced by interstitial agglomeration "**extrinsic stacking fault**".
- The extrinsic stacking fault also seems to be bordered by an edge dislocation. Again, what is the Burgers vector?

In order to determine the Burgers vector of the apparent dislocations bordering the stacking faults, we must do a Burgers circuit *or* use the Volterra definition. For this we must first be clear about the directions in the chosen projection. This is shown below.

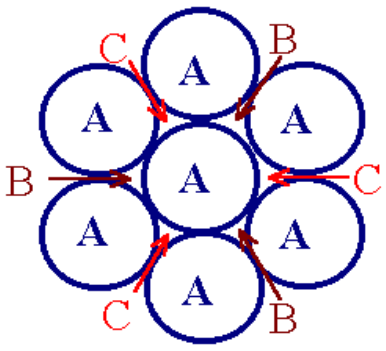
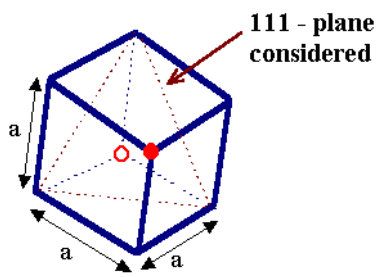
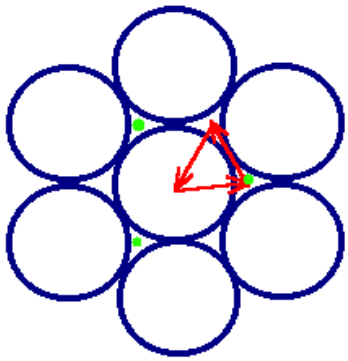


Directions in the **<110>** projection shown for the elementary cell traced out on the right or above	Traces of the (color-coded) *planes* (right angle to direction) in the **<110>** projection and the elementary cell.

- From a Burgers circuit or from a Voltaterra cut, we obtain the same result (Try it! It is easier in this case to hop from atom to atom (instead from lattice point to lattice point); start at the stacking fault).
  - The Burgers vector of these dislocations is ***b* = ± a/3 <111>** - *and this is not a translation vector of the* ***fcc*** *- lattice*! Do not, at this point, forget the distinction between lattice and crystal!
  - Dislocations with Burgers vectors of this type are called **partial dislocations**, or more correctly, **Frank** partial dislocations, or simply **Frank dislocations**.
- This brings us to a general **definition**: Dislocations with Burgers vector that are *not* translation vectors of the lattice are called *partial dislocations*. They must by necessity border a two-dimensional defect, usually a stacking fault.
  - This can be verified with the **Volterra construction** if we add one element: Make a cut in a **{111}** plane and shift by **a/3<111>** perpendicular to the plane. The element added is that we now include shift vectors that are *not* translation vectors of the *lattice*, but vectors between *equivalent positions* of the *atoms*.
  - Partial Burgers vectors and stacking faults thus may exist if the packing of atoms defining the crystal has additional symmetries not found in the lattice. Check this advanced module for an elaboration.
  - As stated in the definition of the Volterra cut-shift-weld procedure, you now must add or remove material. The total effect is the creation of a Frank partial along the cut line and, by necessity, a stacking fault on the cut part of the **{111}** plane.
- We also see now that the primary defects which are generated by the agglomeration of intrinsic point defects in **fcc** lattices are small "**stacking fault loops**".

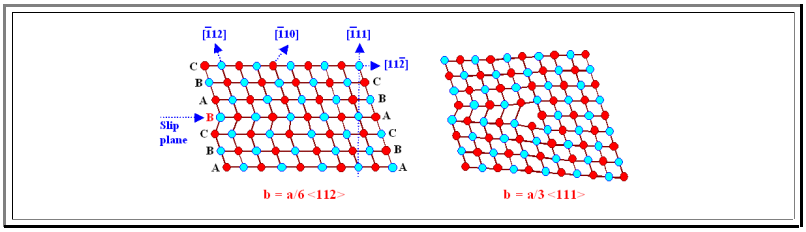
Shockley Dislocations

- Now we may ask a question: Can we produce stacking faults *without the participation of point defects*? Indeed, we may - use the Volterra definition to see how:
  - Make a cut on a **{111}** plane, e.g. between the **A**- and **B**-plane.
  - Move the **B**-plane so it is now in a **C**-position. No material must be removed or added.
  - Weld together: You now have the stacking sequence **ABCACABCA...** instead of **ABCABCA..**, i.e. you produced the stacking sequence of an *intrinsic* stacking fault.
- The vector of the shift must be the Burgers vector of the partial dislocation resulting from this operation as the boundary of the intrinsic stacking fault. This shift vector can be seen by projecting the elementary cell on the close packed **{111}** plane where we did the cut.

		
The displacement vectors for producing stacking faults with the Volterra construction. We have all vectors pointing from one "dent" to a neighboring one.	The directions in the **{111}** plane. If you superimpose the two red circles, you have the projection shown on the left.	Each one of the red vectors would move a **{111}** plane from an A-position to a **B** position (marked by a green dot).

- The relevant displacement vectors are of the type ***b* = a/6<112>**. (Check it! It's good exercise for getting used to lattice projections). Dislocations with this kind of Burgers vector are called **Shockley partial dislocations**, Shockley dislocations, or simply **Shockley partials**.

- In our **<110>** projection, Shockley and Frank partials look like this (after a picture from "Hull and Bacon"). The pictures are drawn in a slightly different style, to make things a bit more complicated (get used to it!)



- You can't quite see the Shockley dislocation? Well, neither can I. But it is time to get used to the fact that not all dislocations are edge dislocation, clearly visible in schematic drawings. We will encounter dislocations that are far weirder and almost impossible to "see" in a drawing, or hard to draw at all. But nevertheless they exist, possess a stress- and strain field described by the formulas from before, and are just the real world inside crystals.

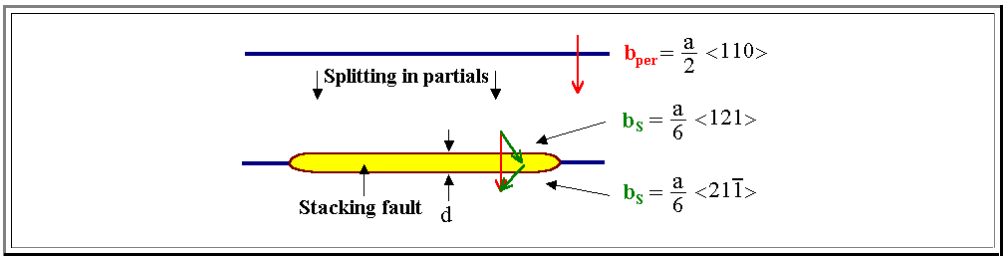
By now you are wondering if these partial dislocations are an invention of bored professors? *Well, they are not!* They are more or less the *only* kind of dislocations that really exist in **fcc** crystals (and some others)!

- The reason for this is that perfects dislocations (with a Burgers vector of the type **a/2<110>**, i.e. a lattice translation vector) will *dissociate to form partial dislocations*. This is one kind of a possible reaction involving partial dislocations, which we are going to study in the next subchapter.

5.4.2 Dislocation Reactions Involving Partial Dislocations

Splitting of Perfect Dislocations into Partial Dislocations

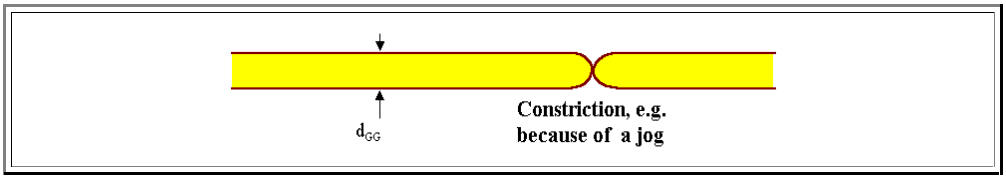
- A perfect dislocation may dissociate into two partial dislocations because this lowers the total energy.
  - The Burgers vector ***b* = *a*/2[110]** may, e.g., decompose into the two **Shockley partials *a*/6[121]** and ***a*/6[2,1,–1]** as shown below.
  - Of necessity, a *stacking fault* between the two partial dislocations must also be generated.



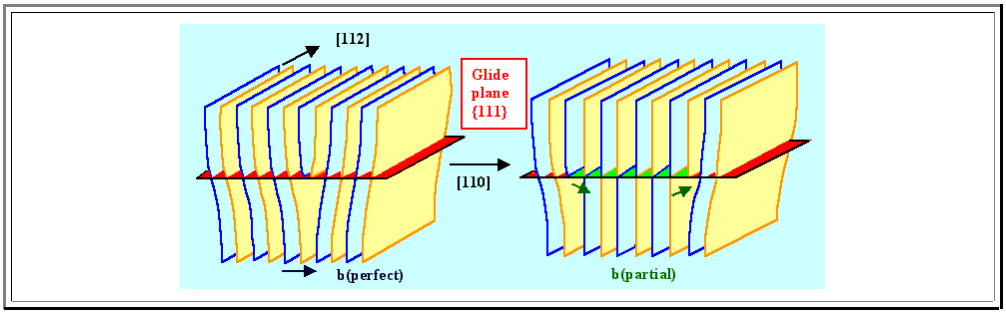
  - You can think of this as doing *two* Volterra cuts in the same plane, each on with the Burgers vector of one of the Shockley partials, but keeping the cut line apart by the distance ***d***. Each cut by itself makes a stacking fault, but the superposition of both creates a perfect lattice.
- Lets balance the energy of this reaction:

Energy of the perfect dislocation	$= G \cdot b^2 = G \cdot (a/2<110>)^2$	$= \frac{G \cdot a^2}{2}$
Energy of the two partial dislocations	$= 2G \cdot (a/6<112>)^2 = 2G \cdot a^2/36 \cdot (1^2 + 1^2 + 2^2)$	$= \frac{G \cdot a^2}{3}$

  - We thus have a clear *energy gain* $-E_{split} = G \cdot a^2$ by having smaller Burgers vectors. This energy gain does not depend on the distance ***d*** between the dislocations.
- But we are not done yet; we have two more energy terms to consider:
  - **1.** The *energy of interaction* $+E_{inter}$; it will be large at short distances. The dislocations repulse each other and the energy going with this interaction is proportional to **1/*d***. Based on this *alone*, the partial dislocations thus would tend to maximize ***d***.
  - **2.** The *energy of the stacking fault* $+E_{SF}$ stretched out by necessity between the two partial dislocations. This stacking fault energy is always $E_{SF} = \gamma \cdot$ **area**, or, taken per *per unit of length* as for the dislocations, $E'_{SF} = \gamma \cdot d$. Based on this *alone*, the partial dislocations thus would tend to minimize ***d***.
- In total we have some energy *gain* by just forming partial dislocations in the first place, but energy *losses* if we keep them too close together, or if we move them too far apart.
- We thus must expect that there is an **equilibrium distance d_{eq}** which gives a minimum energy for the total defect which consists of a **split dislocation** and a **stacking fault**. This equilibrium distance d_{eq} will depend mostly on the stacking fault energy γ; for small γ's we expect a larger distance between the partials.
  - In principle, we can calculate d_{eq} by writing down the total energy, i.e. the sum of the energy gain by forming partial dislocations plus the energy of the interaction plus the stacking fault energy, then find the minimum with respect to ***d*** by differentiation. This is a basic execise, what you will get is $d_{eq} \propto \gamma^{-½}$
- Instead of a a pure one-dimensional defect - our perfect dislocation - we have now something complicated, some kind of ribbon stretching through the crystal. Moreover, this stacking fault ribbon may be constricted at some knots or jogs, and may look like this:



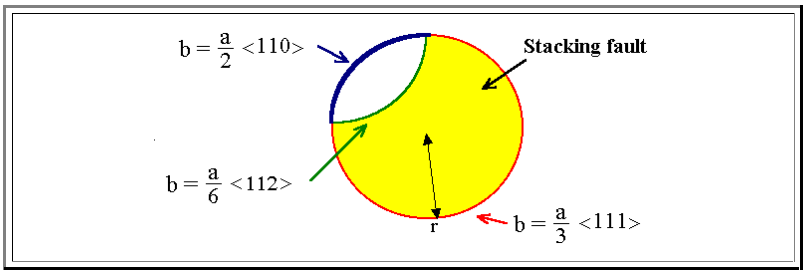
- How would this look in cross-section? We take a picture after "Hull and Bacon"



- It is clear that a dislocation split into *Shockley partials* is still able to glide on the same glide plane as the perfect dislocation; the stacking fault just moves along. It can also change its length without any problems.
- For *Frank type partials* this is *not* true. The loop it usually bounds could only move on its glide cylinder. Changing the length would involve the absorption or emission of point defects.

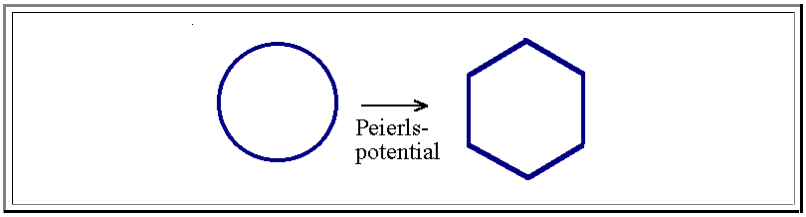
Reactions between dislocation now tend to become messy. You must consider the reaction between the partials and taking into account the stacking fault. However, processes now become possible that could not have occurred before. Lets look at some examples.

- A small dislocation loop formed by the agglomeration of vacancies, that in its pure form cannot add much to plastic deformation, may transmutate into a dislocation loop bounded by a perfect Burgers vector (which in turn may split into Shockley partials) - it is now glissile and can increase its length ad libitum. How does that happen?
- As shown below, the Frank partial bounding the vacancy disc defining the stacking fault has a Burgers vector of the type ***b* = *a*/3<111>**. It then may split into a perfect dislocation with **b = *a*/2<110>** and a Shockley partial with ***b* = *a*/6<112>** (which must lie in the loop plane). The Shockley partial moves across the loop, removing the stacking fault - we have an "**unfaulting**" process. A loop bounded by a perfect dislocation, free to move, is left. The glide plane of the perfect dislocation is not the plane of the loop; the Burgers vector of the perfect dislocation, after all, must have a sizeable component perpendicular to the loop plane in order for the sum of the Burgers vectors to be zero.

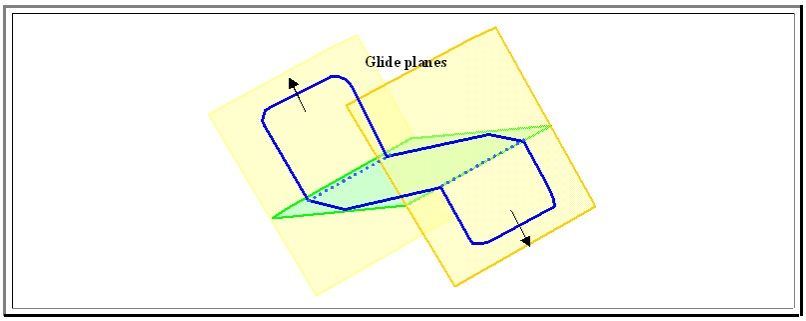


- The Shockley dislocation, once formed, will move quickly over the loop - pulled by the stacking fault like by a tense rubber sheet. The driving force for the reaction is the stacking fault energy: As the loop increases in size because more and more vacancies are added and the radius ***r*** grows, the energy of the loop increases with r^2 due to the stacking fault. However, the line energy of the dislocation only increases with ***r*** no matter what kind of dislocation is bounding the loop.
- There is therefore always a critical radius r_{crit} where a perfect loop becomes energetically favorable.

The perfect loop now feels the Peierls potential, it may try to align the dislocation into the **<110>** directions, always favorable in **fcc** lattices the loop then assumes a hexagonal shape.



- Now all segments are able to glide. If the resolved shear stress for some segments is large enough, they are going to move, pulling out long dislocation dipoles in the direction of the movement. The beginning of this process may look look this:



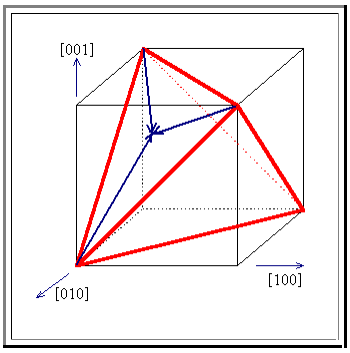
What we have, in summary, is one of the problems of **Si** materials technology:

- We have an efficient source for dislocation generation by vacancy (or, in **Si**, interstitial) agglomeration in formerly dislocation free crystals! And this is not a theoretical possibility, but reality if you are not very careful in growing your crystals. Many examples are shown in the link.

The Thompson Tetrahedron

As we have seen, there are now many possible dislocation reactions. In writing down reaction equations, you must use the specific Burgers vector (e.g. ***a*/6[1, -2, 1]**) and not the general type (***a*/6<112>** for the example). This can be cumbersome and is prone to produce errors.

- Fortunately there is a extremely useful tool for **fcc** lattices to keep the vectors in line: The *Thompson tetrahedron*.
- The Thompson tetrahedron is simply the tetrahedron formed by the **{111}** planes with consistently indexed planes and edges.



- If we look at the **{111}**-planes tetrahedron, we see the following connections
  - The *edges* are **<110>** directions, they may be used to represent the Burgers vectors of the perfect dislocations and the preferred direction for the line vectors because of the Peierls potential (red lines).
  - The *faces* are **{111}** planes, they show the positions of potential stacking faults.
  - The Burgers vector of the Shockley partials that may bound a stacking fault of the given **{111}** plane are the vectors running from the *center* of the triangular faces to the corners (blue lines)
  - The Frank dislocations that also can bound a stacking fault, run from the center of the triangular faces to the center of the tetrahedron (not shown).

- For a "short-hand" description, it is conventional, to enumerate the edges by **A,B,C,D** and the centers of their triangles by **α, β, γ** and **δ**. The relevant vectors than become, e.g., **AB** or **Aγ**.
- It is a good idea (*really!*) to really build a Thompson tetrahedron - maybe from some stiff cardboard; the link gives the detailed net.

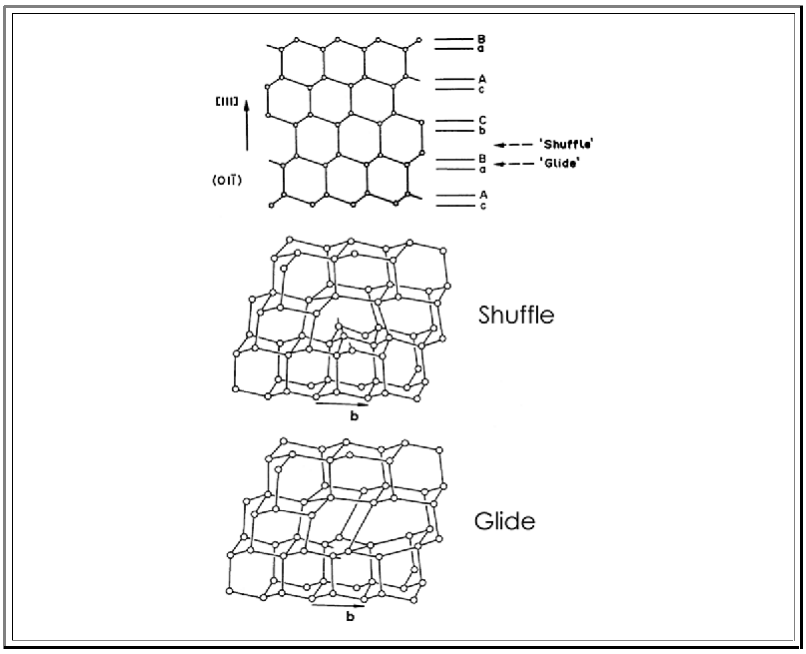
5.4.3 Some Dislocation Details for Specific Lattices

Some Specialities in fcc Lattices

- **Lomer–Cotrell** and **stair-rod** dislocations
  - Lets look at the reaction between two perfect dislocations on different glide planes which are split into Shockley partials, e.g. with the (perfect) Burgers vectors
    $\boldsymbol{b}_1 = \boldsymbol{a}/2[-1,1,0]$ on the **(111)** plane
    $\boldsymbol{b}_2 = \boldsymbol{a}/2[101]$ on the **(–1,1,1)** plane
  - If you have not yet produced your personal Thompson tetrahedra - now is the time you need it!
  - The two Shockley partials meeting first will always react to form a dislocation with the Burgers vector

$$\boldsymbol{b}_{LC} = \frac{\boldsymbol{a}}{6}[011]$$

  - Use your Thompson tetrahedron to verify this!
- This is a *new type of Burgers vector*. A dislocation with this Burges vector is called a *Lomer-Cotrell dislocation*.
  - A Lomer-Cotrell dislocation now borders *two* stacking faults on two different **{111}** planes, *it is utterly immobile*.
  - The total structure resulting from the reaction - a Lomer-Cotrell dislocation at the tip of two stacking fault ribbons bordered on the other side by Shockley partials - is called a **stair-rod dislocation** because it is reminiscent of the "stair-rod" that keeps the carpet ribbons in place that are coming down a stair. What it looks like is shown in the link.
- It is clear that this is a reaction that must and will occur during plastic deformation. Since it makes dislocations completely immobile, it acts as a **hardening mechanism**; it makes plastic deformation more difficult.
- Another speciality in **fcc**-crystals, which would never occur to you by hard thinking alone, are **stacking fault tetrahedra**.
  - Stacking fault tetrahedra are special forms of point defect agglomerates. Lets see what the are and how they form by again looking at low energy configurations:
- Frank partials bonding a vacancy disc have a rather high energy ($\mathbf{b} = \boldsymbol{a}/3\,[111]$, $\boldsymbol{b}^2 = \boldsymbol{a}^2/3$) compared to a Shockley partial ($\boldsymbol{b}^2 = \boldsymbol{a}^2/6$) or Lomer-Cotrell dislocation ($\boldsymbol{b}^2 = \mathbf{a}^2/18$), which also can bound stacking faults. Is there a possibility to change the dislocation type?
  - There is! Imagine the primary stacking fault to be triangular. Let the Frank partial dissociate into a Lomer-Cotrell dislocation and a Shockley partial which can move on one of the other **{111}**-planes intersecting the edge of the triangular primary stacking faults. (If you do not have a Thompson tetrahedra by now, it serves you right!)
  - Let the Shockley partials move; wherever they meet they form another Lomer-Cotrell dislocation. If you keep them on other triangular areas, they will finally meet at one point - you have a tetrahedron formed by stacking faults and bound by Lomer-Cotrell dislocations; the whole process is shown in the link.
  - If this seems somewhat outlandish, look at the electron microscopy pictures in the link!
- Next, lets look at slightly more complicated **fcc**-crystals: the **diamond structure** typical not only for diamond, but especially for **Si**, **Ge**, **GaAs**, **GaP**, **InP**, ... Now we have *two* atoms in the base of the crystal, which makes things a bit more complicated.
  - First of all, the extra lattice plane defining an edge dislocation may now come in *two* modifications called "**glide**"- and "**shuffle" set**, because the inserted half-plane may end in two distinct atomic positions as shown below. The properties of dislocations in semiconductors - not only their mobility but especially their possible states in the bandgap - must depend on the configuration chosen.



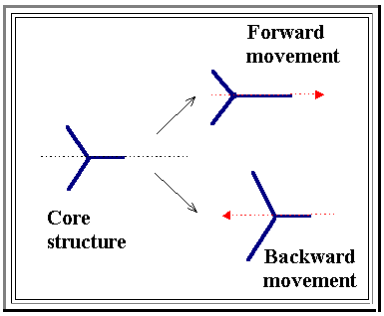
- Which configuration is the one chosen by the crystal? It is still not really clear and a matter of current research.

Some Specialities in bcc Lattices

- The basic geometry in **bcc** lattices is more complicated, because it is not a close-packed lattice.
  - The smallest possible *perfect* Burgers vector is

$$b_{bcc} = \frac{a}{2}\langle 111 \rangle$$

  - Glide planes are usually the most densely packed planes, but in contrast to the **fcc** lattice, where the **{111}** planes are by far most densely packed, we have several planes with very similar packing density in **bcc** crystals, namely **{111}**, {**112**} and **{123}**.
  - This offers many possibilities for **glide systems**, i.e. the combinations of possible Burgers vectors and glide planes. Segments of dislocations, if trapped on one plane may simply change the plane (after re-aligning the line vector in the planes).
  - Stacking faults (and split dislocations) are not observed because the stacking fault energies are too large.
  - But the core of the dislocations, especially for screw dislocations, can now be extended and rather complicated. Screw dislocations in **<111>** directions, e.g., have a core with a threefold symmetry. This leads to a basic asymmetry between the forward and backward movement of a dislocation:



  - Imagine an oscillating force acting on a **bcc** metal - **Fe** for that matter. The screw dislocation will follow the stress and oscillate between two bowed out positions. As long as the maximum stresses are small compared to the critical stress needed to induce large scale movement, the process should be completely reversible.
  - However, due to the asymmetry between forwards and backwards movement, there is a certain probability that once in a while the screw dislocation switches glide planes. It then may move for a large distance, inducing some deformation, In due time, things change irreversibly leading to a sudden failure called "**fatigue**".
  - This is only *one* mechanism for fatigue and only serves to demonstrate the basic concept of long-time changes in materials under load due to details in the dislocation structure of materials.

5.4.4 Essentials to 5.4 Partial Dislocations and Stacking Faults

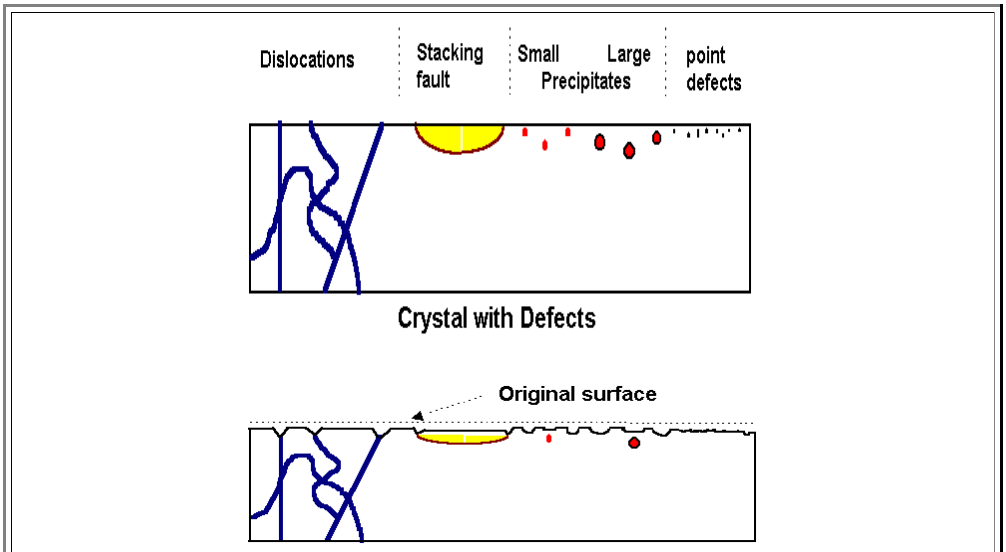
6. Observing Dislocations and Other Defects

6.1 Decoration and Conventional Microscopy

6.1.1 Preferential Etching

Basics of Preferential Etching

- The basic idea behind **preferential etching** is to mark defects intersecting the surface by a small pit or groove, so they become visible in a microscope.
  - Start with a well polished surface that does not show any structures in a **light microscope** (including high magnifications and sensitive modes, e.g. phase or interference contrast
  - Find an etching solution that dissolves your material much more quickly around defects than in perfect regions (that is the tricky part).
  - Expose (= etch) your sample in this solution for an appropriate amount of time. What happens will be something like this:



Model crystal with several kinds of defects intersecting the (polished) surface on top, and surface structure after preferential etching of defects.

- After preferential etching you obtain well developed **etch pits** (actually something looking more like pointed etch cones) at the intersection points of dislocations (including partial dislocations) and the surface and **etch grooves** at the intersection line of grain boundaries and stacking faults with the surface. Precipitates will be shown as shallow pits with varying size, depending on the size of the precipitate and its location in the removed surface layer. Areas with high densities of very small precipitates may just appear rough. Two-dimensional defects as grain boundaries and stacking faults may be delineated as grooves.
  - There is a certain problem with grain boundaries, however: They may also be *delineated*, i.e. rendered visible, with chemicals that do not preferentially etch defects, but simply dissolve the material with a dissolution velocity that depends on the grain orientation (this is the rule and not the exception for most chemicals).
  - In this case grain boundaries show up as *steps* and not as *grooves*. Small steps and grooves, however, look very similar in a light microscope and may easily be mixed up.
- You may think: So what! - in any case I see the grain boundary. Well, almost right, but not quite - there are problems:
  - Grain boundaries separating two grains with similar orientation with respect to the surface would not be revealed.
  - The delineation of grain boundaries obtained under uncertain etching conditions suggests that you delineated *all defects* - but in fact you did not. Delineation of grain boundaries thus must not be taken as an indication that the etching procedure works and there are no defects, because you don't see any!
- Before we look at examples and case studies, two important points must be made:
- **1.** Defect etching for many scientists is a paradigm for "**black art**" in science. There are good reasons for this view:
  - Nobody knows how to mix a preferential etching solution for some material from theoretical concepts. Of course you must look for chemicals or mixtures of chemicals that react with your material, but not too strongly. But after this bit of scientific advice you are on your own in trying to find a suitable preferential etch for your material.

- Well-established preferential etching solutions usually have unknown and poorly understood properties. They sometimes work only on specific crystallographic orientations; their detection limits for small precipitates are usually unknown; they may also depend on other parameters like the doping level in semiconductors; and so on.

2. Defect etching *in practice* is more *art* then science.

- Beginners, even under close supervision by a master of the art, will invariably produce etched samples with rich structures that have nothing to do with defects - they produced so-called **etch artifacts**. It takes some practice to produce reliable results.
- **But:** Defect etching still is by far the most important and often most sensitive technique for observing and detecting defects!

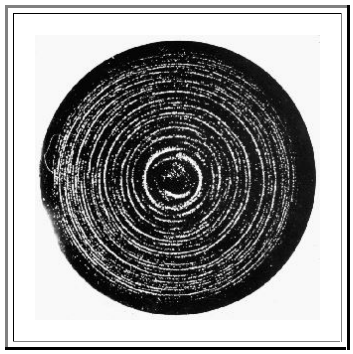
There are many routine procedures for delineating the defects structure of metals by etching. Here we will focus on defects etching in Silicon; which is still the major technique for defect investigations in **Si** technology. Some details and peculiarities of defect etching in **Si** can be found in the link. In what follows we look at the power and possible mechanisms of preferential etching in the context of examples from recent research.

Defect Etching Applied to Swirl Defects in Silicon:

The name "Swirl defects" was used for grown-in defects in large **Si** crystals obtained by the float-zone technique in the seventies.

- Swirl defects are a subspecies of what now is known as "**bulk micro defects**" (***BMD***); they are nothing but agglomerates of the point defects present in thermal equilibrium near the melting point with possible influences of supersaturated impurities still present in ultra clean **Si** (only oxygen and on occasion carbon).
- Whereas the relatively large swirl defects are no longer present in state-of-the-art **Si** crystals, point defect agglomerates and oxygen precipitates still are - there is no way to eliminate the equilibrium defects! **BMDs** are a major concern in the **Si** industry because they cause malfunctions of integrated circuits. The link leads to some recent papers on point defects and **BMDs** in **Si** crystals.

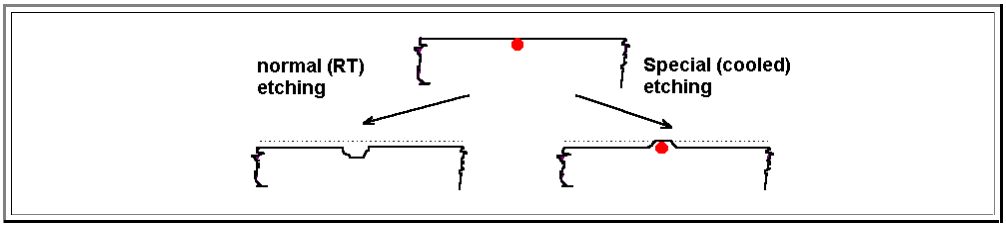
Most of the examples relating to **Si** are taken from the work of **B.O. Kolbesen** (formerly at Siemens; now (2001) at the University of Frankfurt).



- The name "swirl" comes from the spiral "swirl-like" pattern observed in many cases by *preferential etching* as shown on the right.
- Close inspection revealed two types of etch features which must have been caused by different kinds of defects. Lacking any information about the precise nature of the defects (which etching can not give), they were termed "**A**-" and "**B**-swirl defects". More pictures and information in the link

Understanding the precise nature of swirl defects was deemed to be very important for developing crystal growth techniques that could avoid these detrimental defects.

- But etching alone can not give structural data, and other techniques as, e.g., transmission electron microscopy, could not be applied directly because the densities of swirl defects was too small (the likelihood of having a defect in a typical **TEM** sample was practically zero). A combination of a special etching technique and **TEM**, however, could give the desired results.
- The power and the "black art" component of defect etching is nicely demonstrated by the following development: A "special etch" which was simply the old solution, but cooled to about freezing temperatures, did not produce etch pits (and thus remove the defect) for A-swirls, but hillocks (still containing the defect).

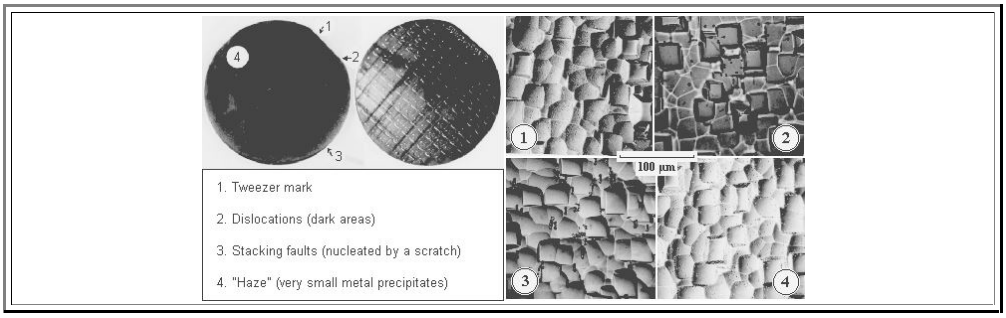


The hillocks identified the precise location of the **A**-swirl defect. A special preparation technique rendered the areas containing hillocks transparent for **TEM** investigations, and the structure of **A**-swirls defects could be identified. They consisted of dislocation loop arrangements that were generated by the agglomeration of interstitials. This gave the first direct evidence that self-interstitials are important in **Si**.

- **B**-swirl defects could not be identified with this technique - their nature is still not clear.
- More about swirl defects and the application of preferential etching can be found in an original paper (in German) in the link.

Process Control by Etching Defects during the Manufacture of Integrated Circuits

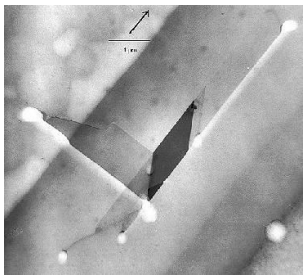
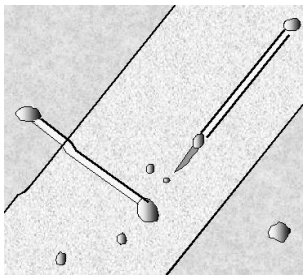
- The manufacture of **integrated circuits** (***IC***) involves many processes prone to introduce defects in the more or less perfect starting crystal.
  - All high temperature processes induce temperature gradients which lead to stress and thus to a driving force for plastic deformation. Since the starting material is dislocation free, the decisive process is the generation of the first dislocations which is much easier if small precipitates or dislocation lops are already present.
  - Thermal oxidation introduces **Si** interstitials with a strong tendency to agglomerate into stacking fault loops, so-called **oxidation induced stacking faults** (***OSF***).
  - All processes tend to induce trace amount of metals which will diffuse into the **Si** and eventually precipitate.
  - Ion implantation destroys the lattice to a large degree up to complete amorphization. Even upon careful annealing some defects may be left over.
- As a general rule, all defects in the electronically active part of an **IC** (roughly the the first **5 µm - 10 µm** of the wafer) are deadly for the device. They have to be avoided and that means that they have to be monitored first. The method of choice is *preferential etching*.
- Lets look at an example
  - The pictures show a **Si** wafer with several defect types introduced during very early stages of processing. Details are provided in the link.



- A few more example are provided in the links. They might be a bit unconvincing, but be aware that looking into an actual microscope gives you much more information than what can be captured in a few pictures.
  - Development of stacking faults in bipolar transistors
  - Precipitates and other defects
- We are now able to compare weaknesses and strength of preferential etching for defect detection:

Strength	Weaknesses
• Simple and cheap • Rather sensitive • Applicable to large areas • Needs no special knowledge (as e.g. **TEM**)	• Black art • Detection limit unclear • What you see must be interpreted • Problems with artifacts • Mechanism not clear • No systematic developments of etches

- One last example serves to illustrate the "*what you see must be interpreted*" point. Shown is a complex defect composed of stacking faults, dislocations and possibly a microtwin in full splendor in a **TEM** micrograph (left), and a schematic outline of what the preferential etching would look like in an optical microscope.

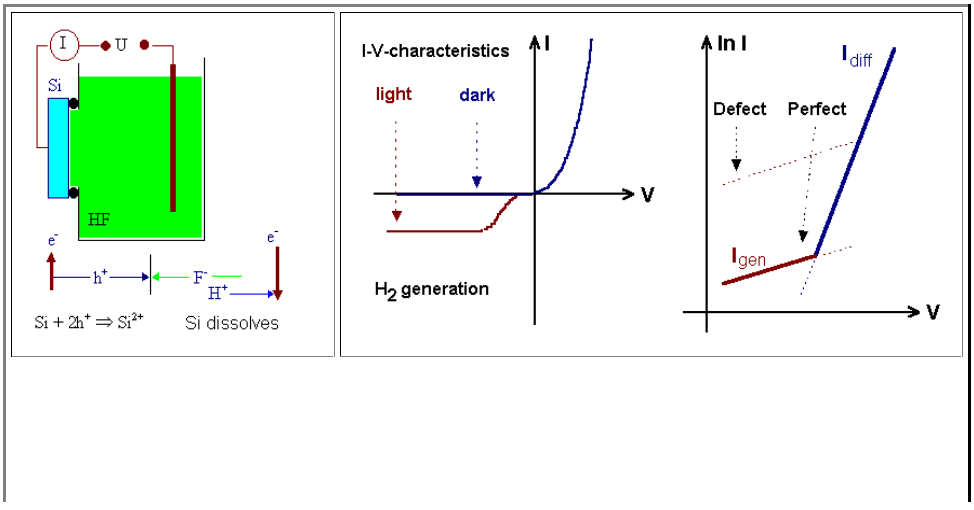
	
TEM micrograph	What you would see with preferential etching Since the etch pits are smaller than **1 µm**, they only would appear as blurred black-white structures

- The planar defects are inclined in a thin foil; what one sees is the projection. One surface was preferentially etched; at the intersection of the defect with this surface the etch features can be seen as bright areas (the sample thickness is smaller at etched parts). The stacking fault lines will be clearly visible in an etch picture, but the various dislocations involved are etched with different strengths.
- It will not be possible to conclude from the etch pattern alone on the complexity of the actual defect. This stacking fault assembly corresponds to some extent to the etch pattern shown in the development of stacking faults in bipolar patterns given in the link.

- Chemical etching on occasion is driven to extremes - simply because there is no alternative. The link leads to an advanced module, where a particular tricky case study is presented

Anodic Etching of Defects and EBIC

- Chemical etching, as any chemical dissolution process, is an oxidation-reduction process expressed in chemical terms. Carriers are transferred from the substrate to the chemicals, new compounds form and go into solution. The paradigmatical model for these processes is **anodic dissolution** under applied bias, where the carriers are supplied by a controlled external power source. Maybe a way towards the understanding of preferential etching comes from the electrochemistry of the specimen?
- Anodic etching has been studied to some extent in Silicon. It leads to a rather unexpected wealth of effects that are at the focus of some current resarch projects. The experiment is simple:
  - Bias the (**p**-type) **Si** sample positively in some electrolyte that contains hydrofluoric acid (**HF**). The **HF** itself is "contacted" by some inert electrode, e.g. a **Pt** wire, which establishes a closed circuit.
  - The **Si-HF**- junction behaves to some extent like a **Schottky** junction; current flow, however, is always accompanied by a chemical reaction. The current density first increases steeply with the applied bias, then reaches a maximum (called j_{PSL}; **PSL** stands for "porous **Si** layer") and decreases again (that is when the analogy with a Schottky junction fails), goes through a second maximum (called j_{ox}) and finally starts to oscillate .
  - In the "forward" regime of the junction, the reaction is the dissolution of **Si** (in reverse condition it is H_2 evolution).
- If a polished specimen that was subjected to a current density considerably smaller than the first peak value is inspected after some etching time, its defect will be revealed in a way reminiscent of purely chemical etching. This can be understood (in parts) by considering current flow in terms of diffusion current and generation currents as introduced in basic **pn-** (or Schottky)-junction theory. The major ingredients for anodic etching are shown below.



Basic experimental set-up, current flow and chemical reaction	Measured ***I-V-***characteristic and theoretical plot of **ln** ***I*** **vs.*** ***V*** with diffusion and generation currents. Around a defect the generation current is larger than in perfect **Si**.

- Preferential defect etching thus can be understood in terms of current flow: At small current densities the generation currents are larger than the diffusion current, the area around **electronically active** defects (i.e. defects that generate carriers) should be etched more deeply and etch pits should appear. At larger current densities the differential etch rate should disappear. The experiments support this view to some extent; the link contains some results
  - General results of anodic etching
- The consideration of the influence of defects on a Schottky junction suggests a different approach to the detection of electronically active defects: Measure the local leakage current or radiation induced current of a junction. This can be done by injecting current locally by an electron beam through a thin Schottky barrier while measuring the induced current. Electronically active defects will recombine more carriers than the defect-free regions, the current will be locally reduced.
  - This method exists and is called "**electron beam induced current**" technique (***EBIC***) if a scanning electron microscope is used as the basic instrument. If a scanned light beam is used, we have the "**light beam induced current**" technique or ***LBIC***; the mainstay of solar cell development with poly crystalline **Si**.
  - The principle of **EBIC** is shown in the link.
  - If one compares anodic etching, chemical etching and **EBIC**, much can be learned about defects and the detection methods, but many questions remain open. Some examples are given in the link
- Anodic etching is still a virulent research issue within the context of the general electrochemistry of semiconductors.

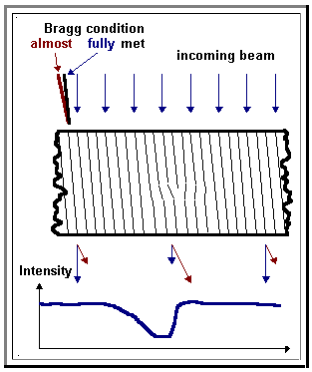
6.1.2 Infrared Microscopy

- Materials that are transparent to visible or - more important - **infra red light** (***IR***) may be investigated in transmission. This usually requires that the sample is optically polished on both sides. Especially semiconductors are transparent in **IR** light and **IR** microscopy is often used to investigate defects; particularly in **III-V** compounds. Defects may be rendered visible by:
  - **Polarization microscopy**. Elastic strain fields may rotate the polarization angle of polarized light to some (small) degree. The strain fields around defects can thus be made visible; an example is shown in the link.
  - **Absorption contrast**. Precipitates, for example, consist of some other material with different optical properties - it may not be transparent to **IR** light. In this case they would be directly visible as dark spots.
- If the primary defects are not precipitates but e.g. small dislocation loops resulting from vacancy agglomeration, they may be turned into a precipitate by a technique called **defect decoration**. This is usually done as follows:
  - Diffuse a fast moving element into the sample (e.g. **Li** or **Cu** for **Si**) at relatively high temperatures (however, without changing the primary defect configuration).
  - Cool down sufficiently fast to nucleate the precipitation of the decorating element *only at defects*, but not so fast that not enough diffusion jumps are possible and you do not get any precipitation. If you cool too slowly, homogeneous nucleation may produce precipitates everywhere and the technique is useless.
  - The primary defects are now heavily decorated with impurity precipitates and visible in **IR** microscopy (or other techniques). However, the dimensions have been enlarged, the primary defect structure may have changed, and you must keep in mind that you are now looking at a different defect from what you wanted to study in the first place!
- Nevertheless, **IR**-microscopy with or without decoration, has made important contributions to the study of defects in crystals. Its weaknesses and strengths can be summarized as follows.

Strength	Weaknesses
• Relatively cheap • Partially quantitative (strain fields) • Large and small areas can be investigated at medium resolution (ca. **1 µm**).	• Well polished surfaces on both sides required • Involved specimen preparation if decoration is used • Often not very specific as to the nature of defects • Only applicable to "medium" defect densities • Not overly sensitive • Interpretation uncertain if decoration techniques are used.

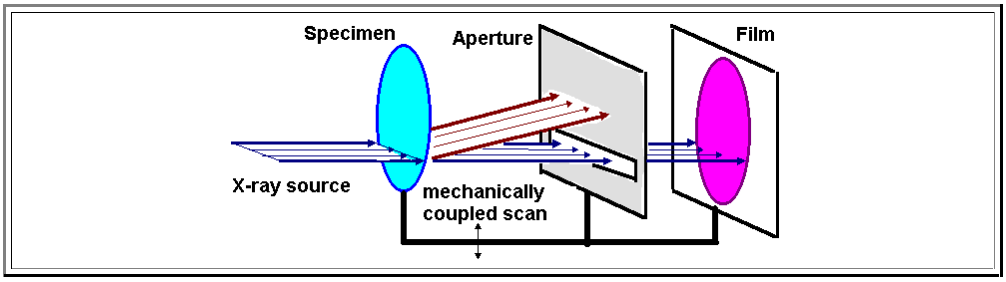
6.2 X-Ray Topography

- There are no efficient lenses for **X**-rays and therefore no **X**-ray microscopes. Still, there are ways to image defects with **X**-rays.
  - The essential part for imaging defects in crystals is the **diffraction** of the **X**-rays in the crystal lattice. This is in contrast to the conventional **X**-ray imaging technique in medical applications were the differential *absorption* of **X**-rays in differently dense tissue is used.
- The basic principle (which is also valid for imaging with electron beams in the transmission electron microscope) is shown below:
  - The specimen is oriented with respect to the incoming wave in such a way that the **Bragg-condition** for diffraction is only met (or nearly met) for just *one* set of lattice planes.
  - All defects with strain fields will locally deform the lattice and thus change the Bragg condition locally. The intensity of the diffracted beam will react to this and vary around defects. This is schematically shown below



- In the example, the specimen is oriented in such a way that the Bragg condition in the perfect part of the crystal is almost, but not quite met. There will be no diffraction or, more quantitatively speaking, a rather low intensity of the diffracted beam. The primary beam thus is transmitted almost without any losses.
- To the left-hand side of the edge dislocation, the strain field bends the lattice plane locally into the Bragg position. In this area the primary beam is strongly diffracted and loses intensity.
- The intensity of the diffracted beam is mirror symmetric to the primary beam.

- For the imaging of defects (typically in **Si**-wafers, with or without processing) the following basic set-up is used.



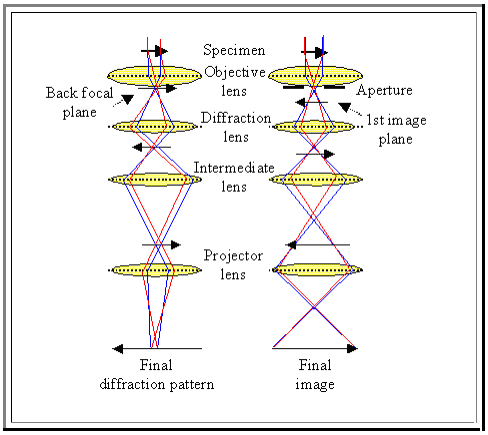
- An **X**-ray source with a thin "one-dimensional" beam cross-section illuminates a line of the wafer. Only the primary beam (or, for dark-field imaging, the diffracted beam) is admitted through an aperture on the film. Wafer, aperture, and film are scanned through the beam.
- Some examples of **X**-ray topography are given in the following links; another one we have already encountered before.
  - Total view and resolution limit
  - Case study in bipolar technique
- The strengths and weaknesses of **X**-ray topography are quite apparent:

Strength	Weaknesses
• Imaging of large wafers with good resolution (ca. **5 µm**) possible • Detailed analysis (e.g. Burgers vectors) possible within limits • No specimen preparation necessary	• Very expensive • rather long exposure times even with powerful (typically **50 kW**) **X**-ray tubes • Resolution/sensitivity not good enough for single/small defects

6.3 Transmission Electron Microscopy

6.3.1 Basics of TEM and the Contrast of Dislocations

- **Transmission electron microscopy** (***TEM***) is by far the most important technique for studying defects in great detail. Much of what was stated before about defects would be speculative theory, or would never have been conceived without **TEM**.
- Using **TEM**, we look through a piece of material with electron "waves," usually at high magnification.
  - In contrast to **X**-ray imaging, lenses for electron beams exist: Magnetic fields (and, in principle, electric fields, too) can be made with gradients that act as convex lenses for the electron waves. For very general reasons it is not possible to construct electromagnetic concave lenses and that means that imaging systems are not very good because lens aberrations cannot be corrected as in conventional optics.
  - Still, the intensity distribution of the electron waves leaving the specimen can be magnified by an electron optical system and resolutions of ≈ **0,1 nm** are attainable.
  - The electrons interact with the material in two ways: inelastic and elastic scattering. Inelastic scattering (leading eventually to absorption) must be avoided since it contains no local information. The electron beam then will be only elastically scattered, i.e. diffracted; the lattice and the defects present modulate amplitude and phase of the primary beam and the diffracted beams locally.
  - The energy of the monochromatic electron beam is somewhere between **(100 - 400) keV**, special instruments go up to **1,5 MeV** (at a price of ca. **8 M€**). Keeping inelastic scattering of the electrons small has supremacy, this demands specimen thicknesses between **10 nm** to ca. **1 µm**. The resolution depends on the thickness; high-resolution TEM (***HRTEM***) demands specimens thicknesses in the **nm** region.
- This has a major consequence: The total volume of the material investigated by **TEM** since it started in the fifties, is less than **1 cm^3**!
  - Taking and interpreting **TEM** images is a high art; it takes several years of practice. The major part of any **TEM** investigation is the specimen preparation. Obtaining specimens thin enough and containing the defects to be investigated in the right geometry (e.g. in cross-section) is a science in itself.
  - Still, practically all detailed information about extended defects comes from **TEM** investigations which do not only show the defects but, using proper theory, provide quantitative information about e.g. strain fields.
- The key is the electron-optical system. It not only serves to magnify the intensity (and, in **HRTEM**, the phase) distribution of the electron waves of the electron waves leaving the specimen, but, at the throw of a switch, provides electron diffraction patterns. The picture shows the basic electron-optical design of a **TEM**



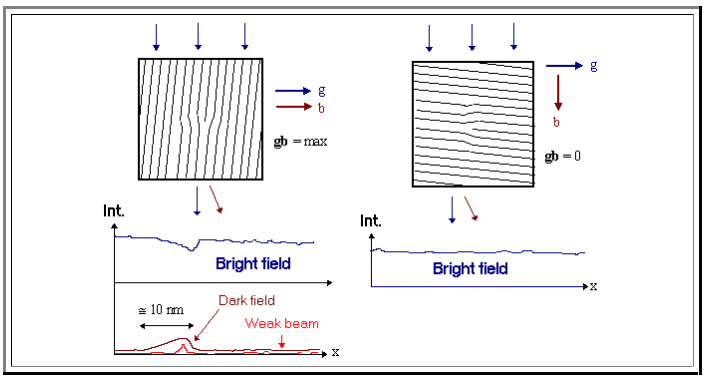
- At least four (usually five) imaging lenses are needed in addition to two condenser lenses (not shown). For most imaging modes an aperture right after the objective lens must be provided.
- The beam paths for the diffraction mode and the imaging mode are shown on the left.
- The most important lens is the *objective lens*. Its resolution limit defines the resolution of the whole microscope.
- The aperture after the objective lens is essential for the conventional imaging modes. It is usually set to only admit the primary beam, or one of the diffracted beams into the optical system.

- The image, or better, the **contrast** of a dislocation depends on several parameters. Most important are:
  - The **diffraction conditions**. Is the Bragg condition fulfilled for many reciprocal lattice vectors ***g***, for none, or just for two? All cases are easily adjusted by tilting the specimen relative to the electron beam while watching the diffraction pattern. The preferred condition for regular imaging is the "two-beam" case with only one "reflex" excited; i.e. the Bragg condition is only met for one point in the reciprocal lattice or one **diffraction vector *g*** (usually with small Miller indices, e.g. **{111}** or **{220}**.
  - The **excitation error**: Is the Bragg condition met exactly (excitation error **= 0**; dynamical case) or only approximately (excitation error **< 0** or **> 0**; kinematical case).
  - The magnitude of the scalar product between the reciprocal lattice vector **g** and the Burgers vector **b**, $\underline{\mathbf{g}} \cdot \underline{\mathbf{b}}$. If it is zero or very small, the contrast is weak, i.e. the dislocation is invisible.

- The **imaging mode**. Is the primary beam admitted through the aperture and used for imaging (**bright field** condition), or a diffracted beam (**dark field** condition)? In other word, is it the intensity distribution of the primary beam or of a diffracted beam that constitutes the image? Or are several beams used whose interference produces a high-resolution image?

How is the proper diffraction condition selected experimentally? Fortunately, a little bit of inelastic scattering produces so-called **Kikuchi** lines which provide a precise and easily interpretable guide to the exact diffraction condition obtained by tilting the specimen. The link shows examples.

- The following picture illustrates some imaging conditions for dislocations with maximum and minimum **gb** product.

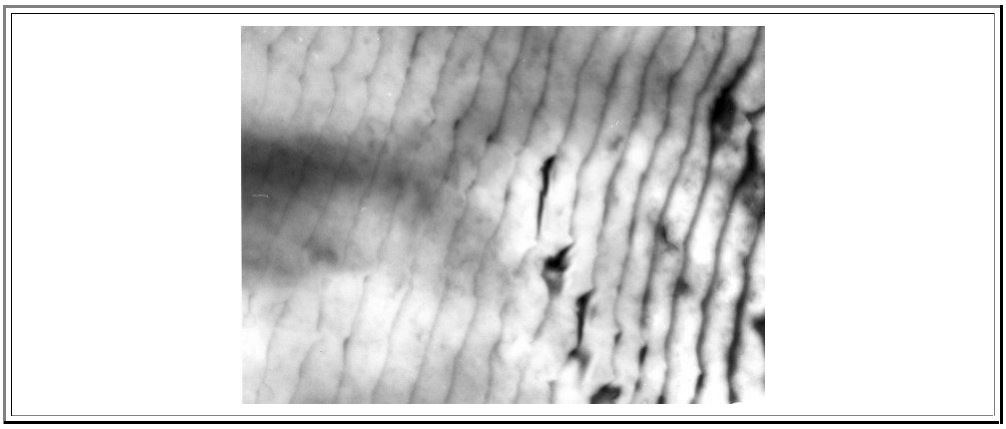


We may draw the following conclusions; they are justified by the full theory of **TEM** contrast.

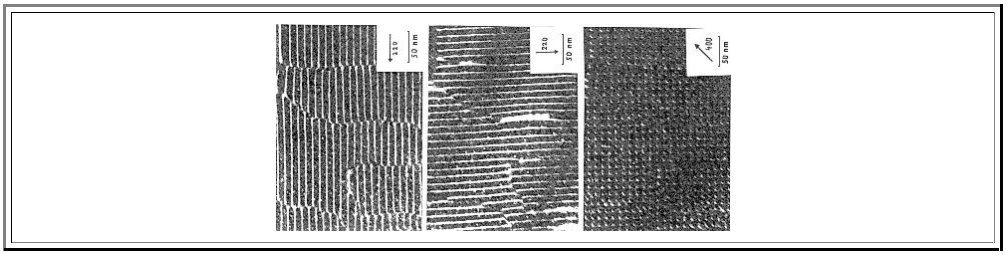
- Dislocations are invisible or exhibit only weak contrast if $\boldsymbol{g} \cdot \boldsymbol{b} = 0$. This can be used for a **Burgers vector analysis** by imaging the same dislocation with different diffraction vectors and observing the contrast.
- Under kinematic bright field conditions (Bragg condition met almost, but not quite), the dislocation is imaged as a dark line on a bright background. The width of the line corresponds to the width of the region next to one side of the dislocation where the Bragg condition is now met; which is usually several **nm**.
- Under dark field conditions the dislocation appears bright on a dark background.
- Under dark field conditions with large excitation errors the Bragg condition is only met in a small region close to the core of the dislocation. The image consists of a thin white line on a pitch black background. This is the so-called "**weak-beam**" condition; it has the highest resolution of conventional imaging modes. It is hard to use, however, because almost nothing is seen on the screen (making adjustments difficult) and long exposure times are needed which are only practical with a very stable instrument.

6.3.2 Examples and Case Studies for Dislocations

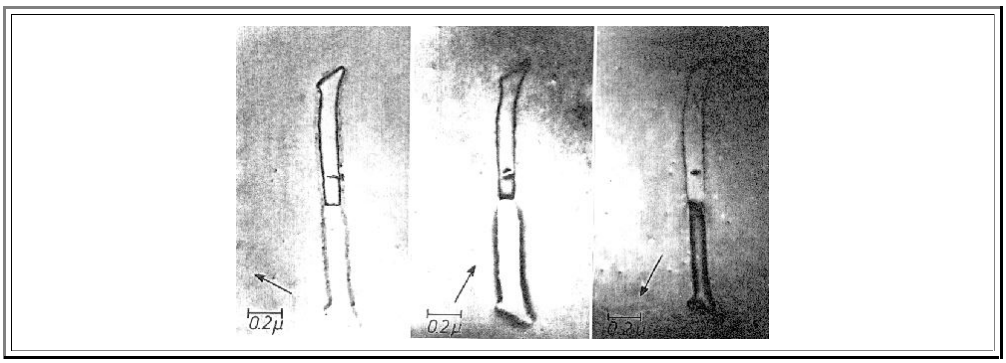
- In what follows a few examples for imaging dislocations with **TEM** are shown. Where possible, examples have been selected that have been used before (e.g. in the context of dislocation loop formations) or will be used later (e.g. in the context of defects in boundaries).
- The first example demonstrates the contrast of dislocations as a function of the excitation error



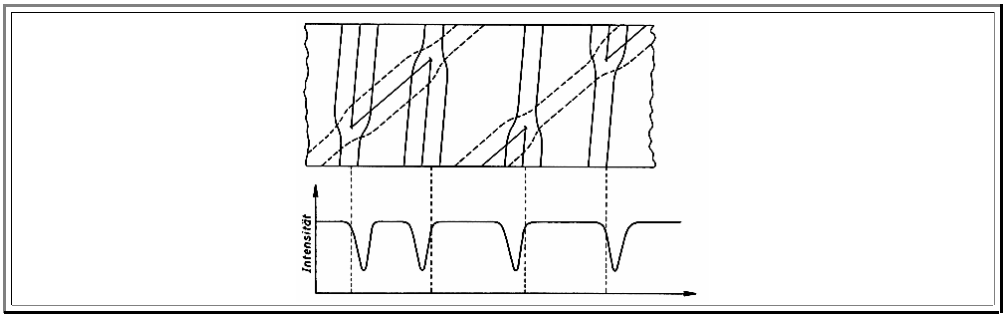
  - The specimen was bent a little; so the excitation error changes from left to right. On the left hand side, the excitation error is relatively large; on the right hand side it is small. The contrast on the left is weak, but the resolution is good; on the right hand side the dislocation appears a as strong, but blurred black line.
- The second example demonstrates the contrast disappearance for $\boldsymbol{g} \cdot \boldsymbol{b} = \mathbf{0}$.
  - Shown is a network of pure screw dislocations in **Si** which we will encounter again in the context of grain boundaries.



  - Only one set of dislocations shows up in the dark field conditions employed for the **g = {220}** type of diffraction vector which is parallel to one Burgers vector and perpendicular to the other one. With a **g = {400}** diffraction vector both sets of dislocation are imaged, but there is a loss of clarity.
- Next we will see how a dislocation loop can be analyzed.
  - Shown are dislocation loops of the "A-swirl defects" imaged with two different diffraction vectors (drawn in as arrows) and a **+g/–g** pair. In the first image, the contrast of the lower dislocation loop has disappeared (the fuzzy line is due to the precipitates along the dislocation line). The two pictures on the right show the lower loop, the image is wide or narrow, depending on the sign of the diffraction vector **g**.

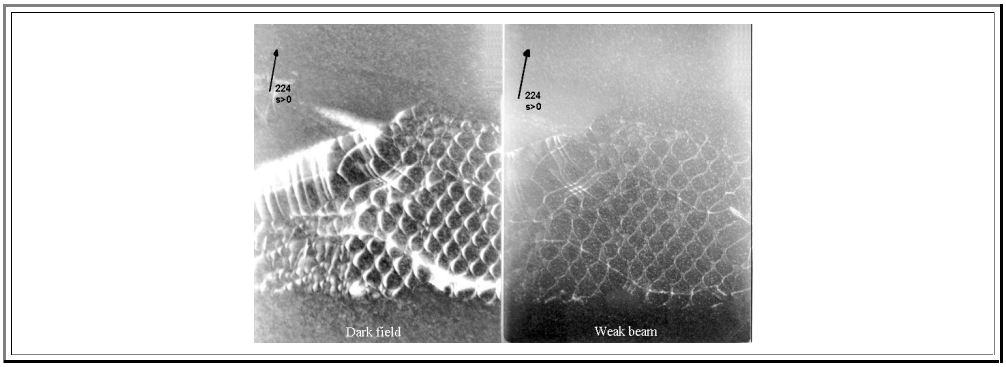


  - This is an important effect because it allows to analyze the nature of a dislocation loop as schematically illustrated below. The image of the loop lies inside or outside the geometric projection; upon reversing the sign of **g** or **b** (and this means switching from vacancy to interstitial type), the image switches between the two extremes.



- For a given geometry it is possible to predict if the contrast is "inside" or "outside"; the nature of a loop may thus be determined. But beware! There are many possibilities of committing a sign error! Printing the negative with emulsion side up or down, e.g., will exchange the signs and turn a vacancy loop into an interstitial loop or vice versa.

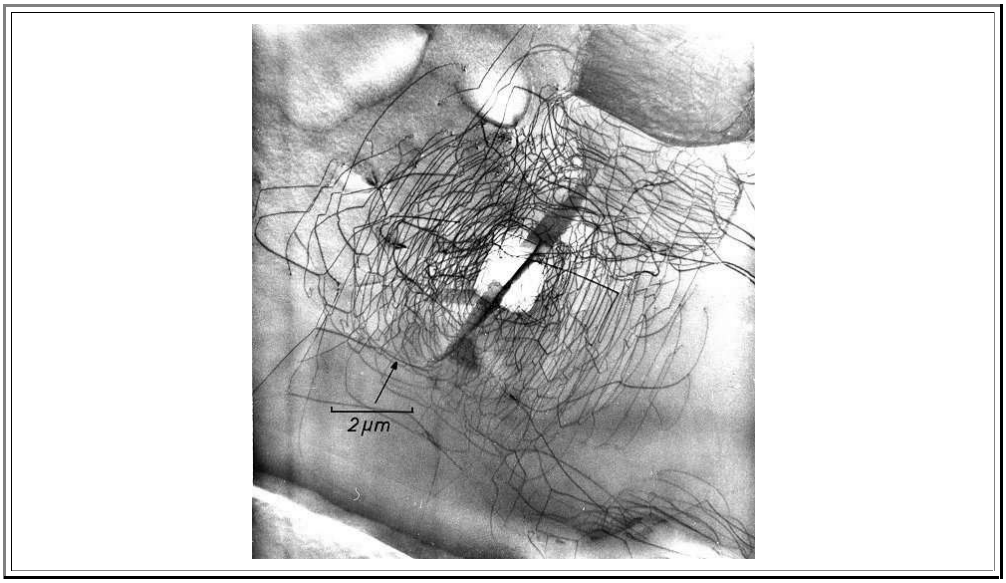
The next example compares regular dark field and weak-beam conditions. The object is a very dense dislocation network with a complicated structure which we will encounter again in the context of grain boundaries.



- The weak-beam image on the left shows a lot more detail, but the signal to noise ratio is rather bad. This is about the limit of the resolution obtainable under weak beam conditions.
- In the link, a comparison between weak-beam conditions and bright field can be found.

The last picture shows conventional bright field imaging.

- The tip of a probe produced some mechanical damage in the emitter area of a transistor in an integrated circuit (the bright square area in the center of the tangle). A microcrack was generated (the elongated black shape); upon heating in the next processing cycle the dislocation tangle was formed to relieve the stress.



- By tilting the specimen while keeping the imaging conditions constant (this involves a rotation around the diffraction vector), a second picture can be obtained under a somewhat different imaging direction. The two pictures can be viewed in a stereo viewer and will produce the full three-dimensional glory of the structure.

More examples can be found in the links

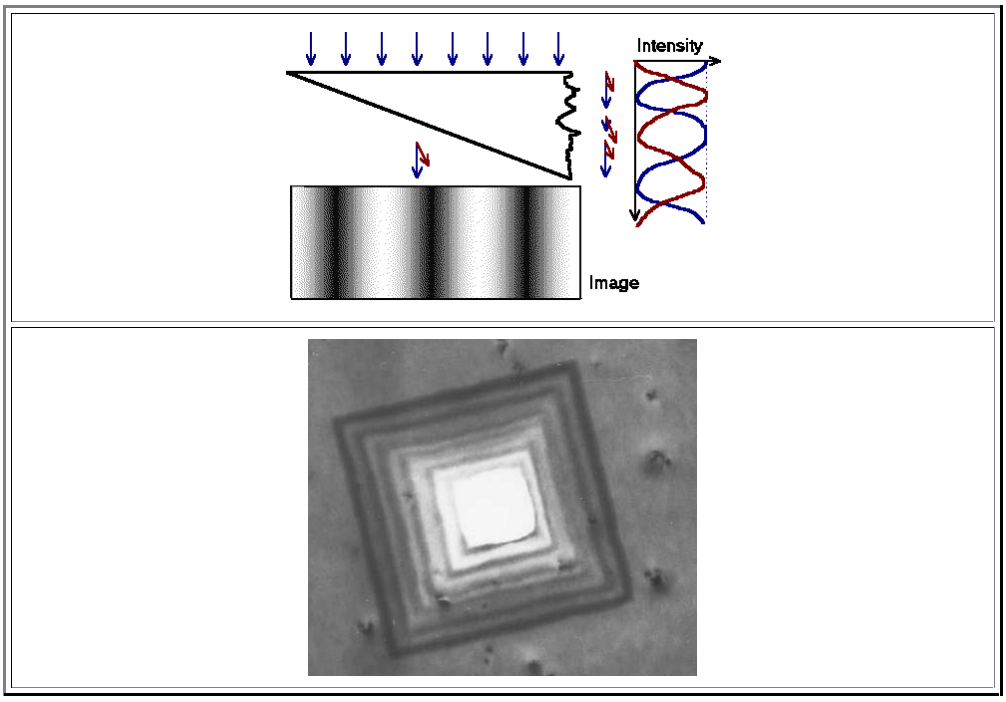
- Weak-beam image of a dislocation network involving partial dislocations under different diffraction conditions
  Unknown ribbon defect in **Si**, showing difficulties in interpretation
  Radiation damage in **Co** showing possibilities of interpretation
  FeSi precipitates in **Si**
  Prismatic punching of dislocation loops from precipitates
  Helix dislocations produced by climb of screw dislocations

Dislocations in **TiAl** being pinned by debris
PtSi on **Si** showing the power of **diffractions patterns** (for another example use the link)
Comparison weak beam - bright field

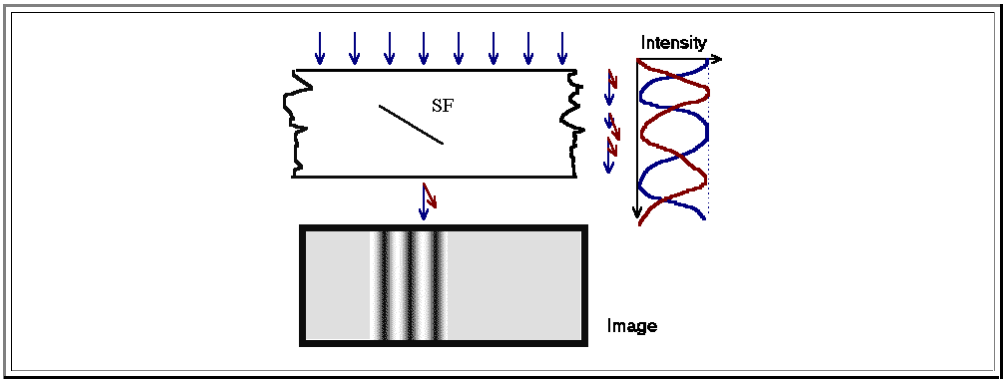
6.3.3 Stacking Faults and Other Two Dimensional Defects

Stacking faults

- Two-dimensional defects like stacking faults, but, to some extent also grain- and phase boundaries, give rise to some special contrast features.
  - Stacking faults are best seen and identified under dynamical two-beam condition; i.e. the Bragg condition is exactly met for one point in the reciprocal lattice.
  - This automatically implies that the diffracted beam, if seen as the primary beam, also meets the Bragg condition; it is diffracted back into the primary beam wave field.
  - This leads to an oscillation of the intensity between the primary and the diffracted beam as a function of depth in the sample; the "wave length" of this periodic intensity variations is called the **extinction length** ξ.
- For a wedge-shaped specimen, the intensity of the primary or diffracted wave thus changes with the local thickness; it goes through maxima and minima.
  - The illustration shows the resulting image: a system of black and white fringes, called thickness fringes or **thickness contours** is seen on the screen.On top a schematicdrawing, on the bottom the real thing. In this case it is an etch pit in a **Ge** sample which is the usual inverted pyramid with **{111}** planes.

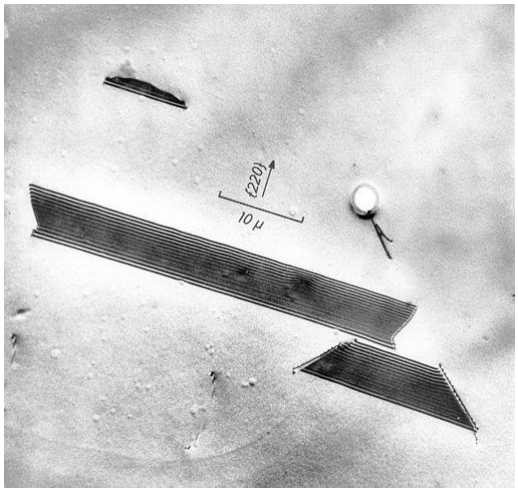


- A stacking fault can be seen as the boundary between two wedge shaped crystals which are in direct contact, but with a displacement ***R*** along the wedge.
  - As a result, the two fringe systems resulting from the two wedges do not fit together anymore. A new fringe system develops delineating the stacking fault; we see the typical **stacking fault fringes**

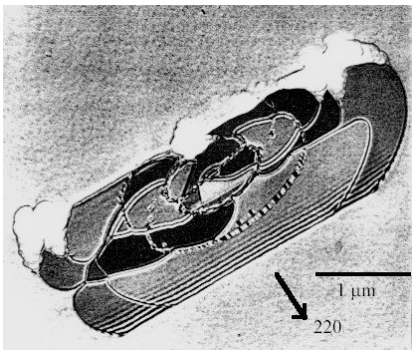


  - Again, getting all the signs right, the nature of the stacking fault can be determined. If intrinsic stacking faults under some imaging conditions would start with a white fringe, extrinsic stacking faults would start with a black one. Reversing the sign of the diffraction vector ***g*** or the displacement vector ***R*** changes white to black and vice versa.
- If more kinematical conditions are chosen, the amplitude of the intensity oscillation decreases; the stacking fault contrast assumes an average intensity that is usually different from the normal background intensity - stacking faults appear in grey.

A few examples: The picture below shows three defects that behave as predicted and could be stacking faults. Indeed, the small defect in the top half and the very large defect are stacking faults. The smaller defect in the bottom part, however is a **micro twin**. This is not evident from one picture, but can be concluded from **contrast analysis**.



The next picture shows a complicated arrangement of several stacking faults:



- A whole system of overlapping **oxidation induced stacking faults** in **Si**. The biggest loop was truncated by the specimen preparation; the fringe system where the stacking fault intersects with one surface is clearly visible.
- The other surface was preferentially etched; the etch pits down the (Frank) dislocation lines are clearly visible.
- The overlap of several stacking faults leads to changing background contrasts - from black to no contrast (whenever multiples of three stacking faults overlap) to almost white.
- Similar if less complicated contrast effects were already encountered in illustrations given before in the context of point defect agglomeration.
- More examples of a typical **oxidation induced stacking faults** in **Si** (***OSF***) are given in the link

But there are limits to **TEM** analysis: Sometimes defects are observed which resist analysis. One example is shown in the link; another one we will encounter in the next subchapter.

Other Defects

The strain-induced contrast of dislocations due to local intensity variations in the primary and diffracted beams and the fringe contrast of stacking faults due to local phase shifts of the electron waves, if taken together, are sufficient to explain (quantitatively) the contrast of any defect.

- It may get involved, and not everything seen in **TEM** micrographs will be easily explained, but in general, contrast analysis is possible and the detailed structure of the defect seen can be revealed within the limits of the resolution (you cannot, e.g., find a kink in a dislocation (size ca. **0,3 nm**) with a typical kinematical bright field resolution of **5 nm**).

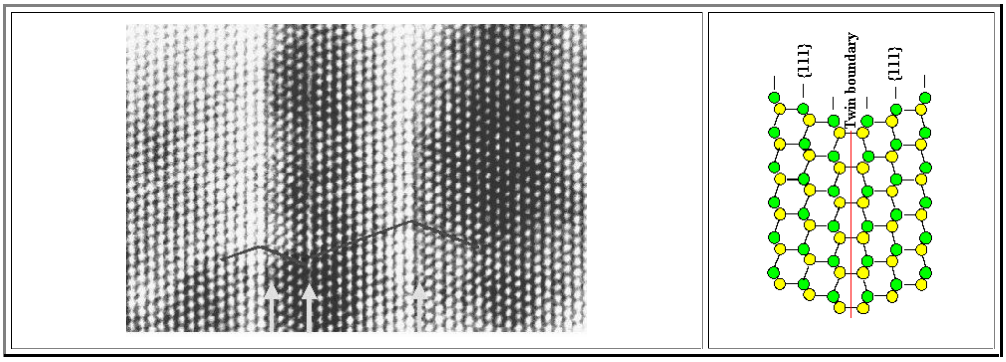
In the links a gallery of micrographs is provided with a wide spectrum of defects. Bear in mind that most examples are from single crystalline and relatively defect free Silicon. The images of regular poly-crystalline materials would be totally dominated by their grain boundaries (see the examples at the end of the list).

- Small dislocation loops in Cobalt produced by ion-implantation.
- Precipitates in Silicon with dislocation structures.
- Needle shaped **$FeSi_2$** precipitates in **Si**; the bane of early IC technology.

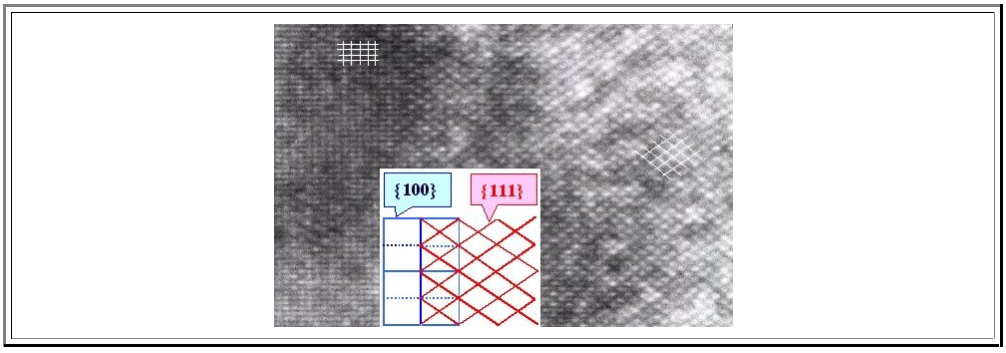
- Helical dislocations resulting from the climb of screw dislocations.
- Bowed-out dislocations in a **TiAl** alloy; kept in place by point defects and small precipitates.
- A thin film of **PtSi** on **Si** as an example of the "real" world of fine-grained materials.
- Overview of **TiAl** as an example of a specimen with a high defect density.

6.3.4 High Resolution TEM

- High-Resolution **TEM** (***HRTEM***) is the ultimate tool in imaging defects. In favorable cases it shows directly a two-dimensional projection of the crystal with defects and all.
  - Of course, this only makes sense if the two-dimensional projection is down some low-index direction, so atoms are exactly on top of each other.
- The basic principle of **HRTEM** is easy to grasp:
  - Consider a very thin slice of crystal that has been tilted so that a low-index direction is exactly perpendicular to the electron beam. All lattice planes about parallel to the electron beam will be close enough to the Bragg position and will diffract the primary beam.
  - The diffraction pattern is the **Fourier transform** of the periodic potential for the electrons in two dimensions. In the objective lens all diffracted beams and the primary beam are brought together again; their interference provides a back-transformation and leads to an enlarged picture of the periodic potential.
  - This picture is magnified by the following electron-optical system and finally seen on the screen at magnifications of typically $\mathbf{10^6}$.
- The practice of **HRTEM**, however, is more difficult them the simple theory. A first illustration serves to make a few points:



  - The image shows one of the first **HRTEM** images taken around **1979**; it is the **<110>** projection of the **Si**-lattice; a schematic drawing is provided for comparison. It also contains a few special grain boundaries, called twin boundaries.
- We notice a few obvious features:
  - Instead of two atoms we only see a dark "blob."
  - Or does the dark blob signal the open channels in the lattice projection? There is actually no way of telling from just one picture.
  - The twin boundaries look fine in comparison to the drawing at a first glance. Looking more closely, one realizes that there are a few unclear points: The yellow arrow points to "fuzzy" lattice planes to the right (or left) of the boundary. Following a fringe across the boundary seems to result in an offset - what does it mean? But what should we expect defects (in this case the twin boundaries) to look like? After all, they destroy the periodicity of the lattice and it is not obvious what Fourier transforms of defects will produce in general cases.
- The last point is easy to solve: Just do a simulation of a defect (i.e. calculate the image for an assumed slice of a crystal with all atoms at the proper positions), but mind the points mentioned below! These are the limitations to **HRTEM** stemming from the non-ideality of the instrument and the specimen:
  - The specimen is not arbitrarily thin! If the thickness is in the order of the extinction length, some reflexes may have very small intensities because they were diffracted back into the primary beam. The objective lens then will not be able to reconstruct the spatial frequencies contained in these reflections; the image looks like a different lattice.
  - This can be nicely seen in a **HRTEM** image of **Si** where the thickness of the sample increases continuously:



- The inset shows the lattice in **<110>** projection; an elementary cell is given by the large rectangles formed by solid blue lines.
- On the right hand side of the picture, all reflections are excited; the very strong **{111}** reflections dominate the image and the **{111}** lattice planes (indicated by white lines) are most prominent. On the left hand side the thickness happens to be in a region where the **{111}** reflections are weak; the **{400}** type reflections dominate (**{100}** etc. are "forbidden" in the diamond lattice). The lattice appears rectangular.

In principle, this can be calculated, too, without much problems. What is much more problematic is the "**contrast transfer function**" of the objective lens.

- If we consider the objective lens to be some kind of amplifier that is supposed to amplifies (spatial) frequencies in the input with constant amplification and without phase distortion, the objective lens is a *very bad* amplifier. It has a frequency response that is highly nonlinear, the amplification drops off sharply for high (spatial) frequencies (meaning short distances). In other word, the resolution is limited (to roughly **0,1 nm** in good **TEMs**); you cannot see smaller details.
- But worse yet, around the resolution limit, the objective lens induces strong phase shifts as a function of several parameters (the most important one being the focus setting); this influences the interference pattern which will define the image.

Both effects together can be expressed numerically in the contrast transfer function of the lens. If you know that function (for every picture you take) you may than calculate what the image would look like for a "perfect" lens with a certain resolution limit; or somewhat easier, you calculate what a crystal with the defects you assume to be present would look like in your particular microscope with the contrast transfer function that it has.

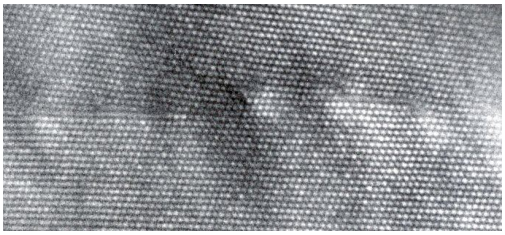
Neither approach is very easy; the amount of computing needed can be rather large. Worse, you must determine parts of the contrast transfer function experimentally; and that involves taking several images at different focus settings. Still, **HRTEM** images provide the ultimate tool for defect studies. They are perfectly safe to use without calculations if you obey two simple rules:

- Only look at pictures where the perfect part of the crystal looks as it should. After all, you usually know what kind of material you are investigating. So if the image of a diamond structure looks like the left part of the illustration above; throw it away (or at least use with care). It it looks like a diamond strucure you can't go totally wrong in interpreting the picture.
- Only draw qualitative conclusions (e.g. there is a dislocation in this **GaAs** specimen!); never draw quantitative conclusion (e.g. it ends at a Ga atom!) without calculating the image.

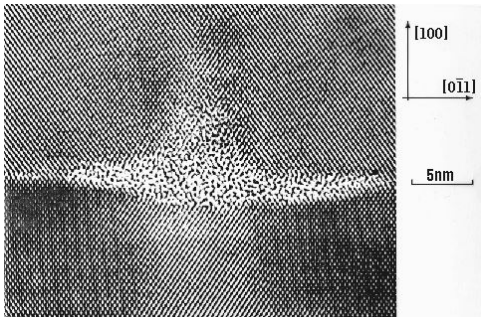
Some more details to **HRTEM** imaging can be found in the (German) article in the link

Three examples may serve to illustrate **HRTEM** here; more will be found in the upcoming chapters.

- The first picture shows a small angle grain boundary in **Si**. This was the first picture of this kind; it only can be interpreted qualitatively; the contrast transfer function was not known. What we see beyond doubt are several lined-up dislocations which constitute a boundary - the top half of the lattice is tilted with respect to the bottom half.

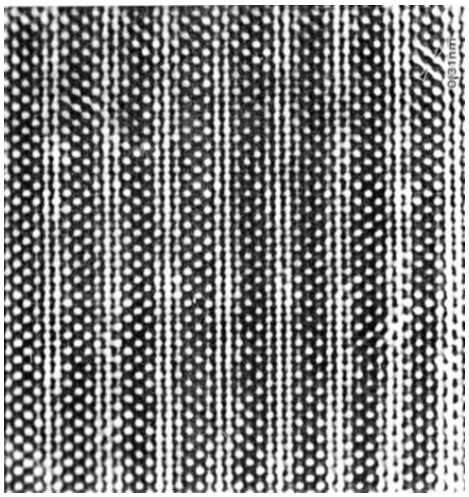


- The next picture (from W. Bergholz) shows an **SiO_2** precipitate in **Si**. Again, a qualitative interpretation is neither possible nor necessary. It is clear that the precipitate, albeit very small, is not spherical



Courtesy of W. Bergholz

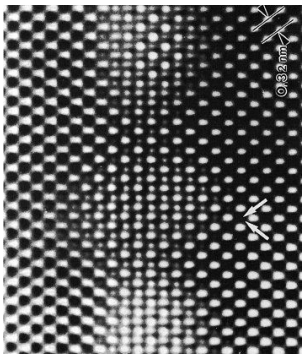
The last example shows quantitative **HRTEM** (from W. **Jäger**) Careful imaging under various conditions, extraction of the contrast transfer function and prodigious computing allowed not only to image a sequence of **Si - Ge** multilayers produced by molecular beam epitaxy (***MBE***), but to identify the positions of the **Si** and **Ge** atoms. The first picture shows an overview. The brighter regions indicate the **Ge** layers, but it is not clear exactly how the lattice changes from **Si** to **Ge**.



Courtesy of W. Jäger

This image also demonstrates the progress made in building electron microscopes. The "old" pictures shown above were taken with a the best general-purpose **TEM** available around **1980** (Siemens Elmiskop **102**). The last pictures were taken with a **TEM** optimized for high resolution around **1995**.

Next a comparison between an enlarged part of the **Ge/Si** stack is shown together with a quantitative evaluation of this and other pictures obtained at different focus settings from W. Jaeger and his group. The color codes defined **Ge** concentrations and a very clear representation of the multilayer sequence is obtained.



Courtesy of W. Jäger

7. Grain Boundaries

7.1 Coincidence Lattices

7.1.1 Twin Boundaries

7.1.2 The Coincidence Site Lattice

7.1.3 The DSC Lattice and Defects in Grain Boundaries:

7.2 Grain Boundary Dislocations

7.2.1 Small Angle Grain Boundaries and Beyond

7.2.2 Case Studies: Small Angle Grain Boundaries in Silicon I

7.2.3 Case Studies: Small Angle Grain Boundaries in Silicon II

7.2.4 Generalization

7.3 O-Lattice Theory

7.3.1 The Basic Concept

7.3.2 Working with the O-Lattice

7.3.3 The Significance of the O-Lattice

7.3.4 Periodic O-Lattices and Pattern Elements

7.3.5 Pattern Shift and DSC Lattice

7.3.6 Large Angle Grain Boundaries and Final Points

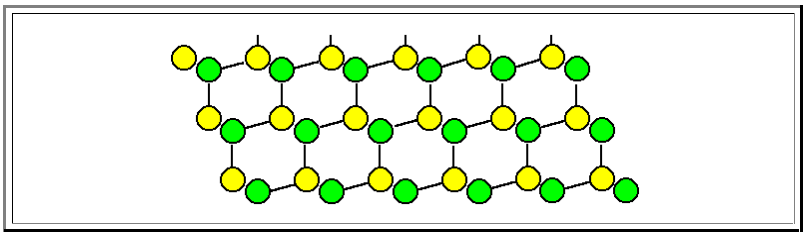
7. Grain Boundaries

7.1 Coincidence Lattices

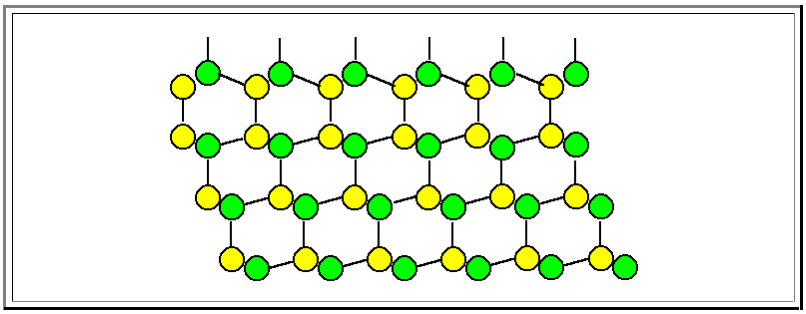
7.1.1 Twin Boundaries

General Remarks

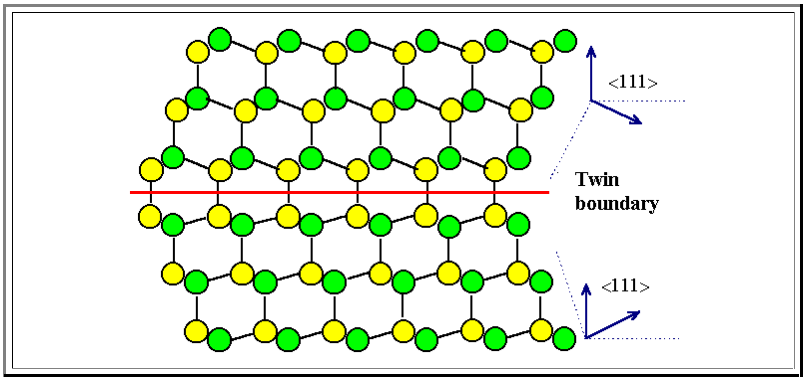
- So-called "**twin boundaries**" are the most frequently encountered grain boundaries in Silicon, but also in many other **fcc** crystals. This must be correlated to an especially low value of the interface energy or **grain boundary energy** always associated with a grain boundary. This becomes immediately understandable if we construct a (coherent) twin boundary. The qualifier "coherent" is needed at this point, we will learn about its meaning below.
  - Lets look at the familiar **<110>** projection of the diamond lattice:



  - Now we introduce a stacking fault, e.g. by adding the next layer in mirror-symmetry (structural chemists would call this a "*cis*" instead of a "*trans*" relation):



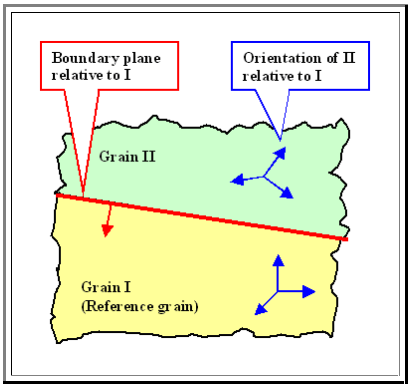
  - If we were to continue in the old stacking sequence, we would have produced a stacking fault. However, if we continue with mirror-symmetric layers, we obtain the following structure *without changing any bond lengths or bond angles*:



- We generated a (coherent) **twin boundary!** This is obviously a *special* grain boundary with a high degree of symmetry.

Now let's try to describe what we did in *general* geometric terms. To describe the twin boundary from above or just *any* boundary geometrically, we look at the general case illustrated below:

- We have two arbitrarily oriented grains joined together at the boundary plane. We thus need to define the "arbitrary" orientation and separately the boundary plane.



Choosing grain **I** as reference, we now can always "produce" a arbitrary orientation of the second grain by "cutting" a part of grain **I** off - along the boundary plane - and then rotating it by arbitrary angles α, β and γ around the ***x-***, ***y-*** and ***z-*** axis (always defined in the reference crystal, here grain **I**).

- Of course, the grain **II** thus produced would not fit together anymore with grain **I**. So we simply remove or add grain **II** material, until a fit is produced.
- Alternatively, we could simply rotate the second crystal around *one* angle α if we pick a suitable polar vector for that.
- Specifying this polar (unit) vector will need *two* numbers or its direction - e.g. two angles relative to the boundary plane or any other reference plane, and *one* number for its length specifying the angle of rotation.
- In either way: *three* numbers then for the orientation part.

The boundary then is defined by **5** parameters: The *three* rotation angles needed to "produce" grain **II**, and *two* parameters to define the boundary plane by its Miller indices **{hkl}** in the coordinate system of the reference grain **I**.

- Why do we need only *two* parameters to define the boundary plane? After all, we usually need *three* Miller indices **{hkl}** to indicate a plane in a crystal?
- Good question! We need only *two* numbers in this case because here a *unit vector* with the right orientation given by **{hkl}** is sufficient - the length, likewise encoded in **{hkl}**, does not matter. Since you need only *two* angles between a unit vector and the coordinate axis' to describe it unambiguously, *two* numbers are enough, even so they cannot be given straightforward in **{hkl}** terms.
- The third angle is then always given by the **Euler relation**

$$\sin^2\alpha + \sin^2\beta + \sin^2\gamma = 1$$

- Again, Miller indices do not only give the direction of a vector perpendicular to a crystallographic plane, but also a specified length which contains the distance between the indexed crystallographic planes - and that's why they need *three* numbers in contrast to a unit vector.

Thus, constructing a simple (coherent) twin boundary, looking at it and generalizing somewhat, we learned a simple truth:

A (simple) grain boundary needs (at least)
5 parameters
for its geometric description

That was some basic geometry for a *simple* grain boundary! For *real* grain boundaries we must add the complications that may prevail, e.g.:

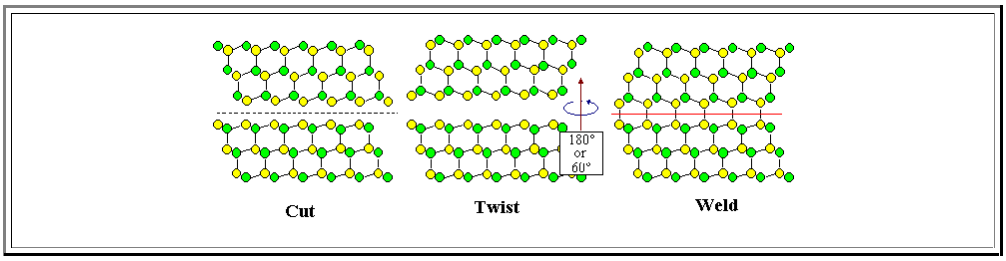
- The grain boundary is not flat (not on one plane), but arbitrarily bent.
- The grain boundary contains (atomic) steps and other local "grain boundary defects".
- The grain boundary contains foreign atoms or even precipitates.
- The grain boundary is not crystalline but consists of a thin amorphous layer between the grains.

Since all those (and more) complications *are* actually observed, we must (albeit reluctantly) conclude:

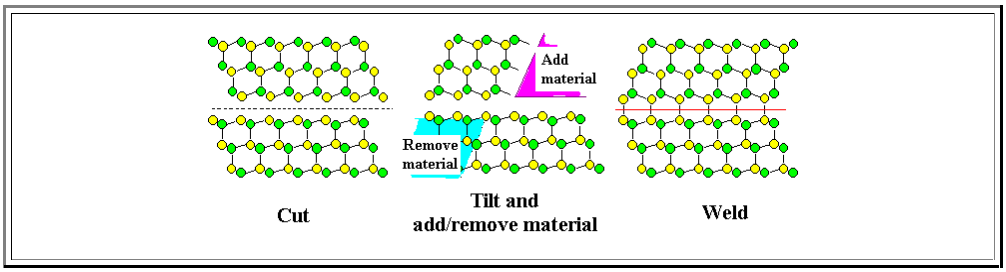
Grain boundaries are rather complicated defects!

Detailed Consideration of the Coherent Twin Boundary

- We can learn more about grain boundaries by analyzing our (coherent) twin somewhat more. While we generated this defect by adding **{111}** layers in a mirror-symmetric way, there are *other ways of doing it*, too:
- We first consider the twin boundary as a pure **twist boundary**. The recipe for creating a *twist* boundary following the general recipe from above is as follows:
  - Cut the crystal along a **{111}** plane (using Volterras knife, of course).
  - Rotate the upper part by **180°** or **60°** around an axis *perpendicular to the cut plane* (= "*twist*")
  - Weld the two crystals together. There will be no problem, everything fits and all bonds find partners.
- This procedure is shown below. We are forced to conclude that we obtain exactly the same twin boundary that we produced above!



- Now lets look at the other extreme: We construct our twin boundary as a pure **tilt boundary**. The recipe for creating a *tilt* boundary following the general recipe from above is as follows:



  - Cut the crystal along a **{111}** plane (using Volterras knife, of course)
  - Rotate the upper part by **70.53°** around an axis *perpendicular to the drawing plane* (= "*tilt*" the grain); i.e. use a **<110>** direction for rotating.
  - Now, however, we must fill in or remove material as necessary.
  - Weld together.
- This procedure is shown above; again we obtain exactly the same twin boundary we had before!
  - This is not overly surprising, after all the symmetries of the crystal should be found in the construction of grain boundaries, too.
  - Well this is conceptually easy to grasp, it generates a major problem in the mathematical analysis of the grain boundary structure. What you want to do then is to generate a grain boundary by some *coordinate transformation* of one grain, and than analyze its properties with respect to the necessary *transformation matrix*. If there are several (usually infinitely many) possible matrices, all producing the same final result, you have a problem in picking the "right" one. We will run into that problem in chapter 7.3.

Grain Boundary Orientation and Energy

- Now lets look at the energy of the twin boundary and see if we can generalize the findings. We are interested in answering the following questions:
  - Is the grain boundary energy (= the energy needed to generate **1 cm^2** of grain boundary) a function of the **5** parameters needed to describe the boundary? The answer, of course, will be *yes*, so now we ask:
  - For a *given orientation*, could a small change in the three angles describing that orientation induce large changes in the energy? Asking a bit more pointedly: Are there orientations leading to boundaries with especially small or large energies? *Are there favorable and unfavorable orientations*?
  - For a *given orientation*, are there possibilities to minimize the energy, e.g. by changing the boundary plane? *Are there favorable and unfavorable planes*?

We will be able to answer these question to some extent by using our twin boundary. First lets look at the energy of the (coherent) twin boundary as shown above.

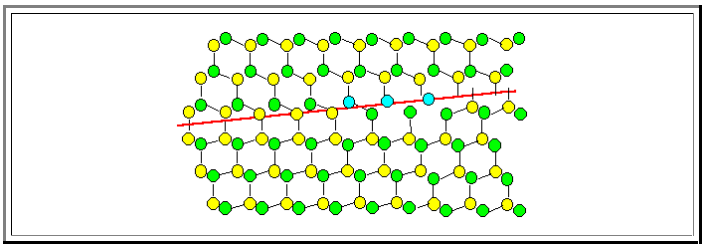
- We would expect a rather small energy per unit area, because we did not have to change bond lengths or angles. We should expect that the energy of a (coherent) twin boundary should be comparable to that of a stacking fault.
- It is hard to imagine a boundary with lower energy and this explains why one always finds a lot of twin boundaries in cubic (and hexagonally) close-packed crystals

Now lets generate a twin boundary with the "twist" or "tilt" recipe, but with *twist or tilt angles slightly off the proper values*. Lets assume a twist angle of e.g. **58°** instead of **60°**. We then make a boundary with a similar, but distinctly different orientation.

- Try it! Can't be done. Nothing will fit any more; a lot of bonds must be stretched or shortened and bent to make them fit. No doubt, the energy will increase dramatically. In other words:

The energy of a grain boundary may dependent *very much* on the precise orientation relationship

Next lets imagine the generation of a twin boundary where the twist or tilt angle is exactly right, but where the *cut-plane is slightly off* ***{111}***. The result looks like the schematic drawing below:



- Again, we have a hard time with the welding procedure. Some atoms will find partners with a slight adjustment of bonds, but other are in awkward positions, e.g. the atoms colored blue in the above illustration. *The energy will be much higher than for a* ***{111}*** *plane* - for sure. In other words:

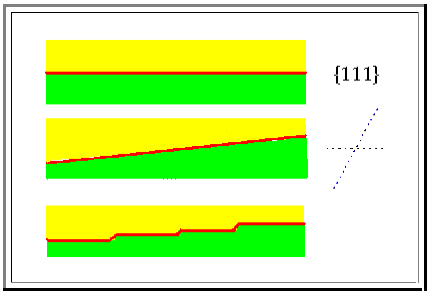
The energy of a grain boundary may dependent *very much* on the Miller Indices of its plane

- We now can understand the meaning of the qualifier "**coherent**" in connection with a twin boundary: Only twin boundaries on **{111}** planes are simple boundaries; they are called coherent to distinguish them from the many possible *incoherent* twin boundaries with planes other than **{111}**.

Optimization of Grain Boundary Energies

From the last observation we can easily deduce a first recipe for optimizing, i.e. lowering, the energy of a given grain boundary:

- Decompose the grain boundary plane into planes with low energies. If that cannot be done, form at least large areas on low-energy planes and small areas of connecting high-energy planes. In other words, approximate the plane by a zig-zag configuration of planes.
- This process is called facet forming or **facetting**, the boundary plane forms **facets**. This is illustrated below.



- Grain boundary on low-energy plane (i.e. a **{111}** plane for a twin boundary). The **{111**} planes are indicated by the dashed lines
- Grain boundary on high -energy plane
- Energy optimization by facetting on {**111}**. The total area increases somewhat, but the energy decreases.

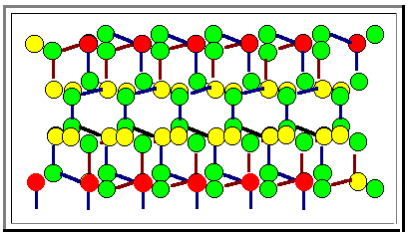
This is an important insight with far-reaching consequences: We need no longer worry very much about the grain boundary plane! It is always possible to optimize the energy by facetting.

- Facetting involves, of course, the movement of atoms. However, only small movements or movements over small distances are needed, so facetting is not too difficult if the temperatures are not too low. So *the crystal has an option* - it can change the boundary plane by moving a few atoms around.
- Experience, too, seem to show that boundary planes are not very important: Grain boundaries, as revealed by etching or other methods, are usually rather curved and do not seem to "favor" particular planes - with the exception of coherent twins. This, however, is simply an illusion because the facetting takes place on such a small scale that it is not visible at optical resolution.

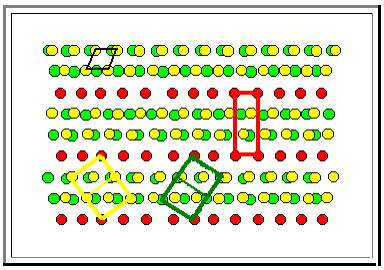
- We now must deal with the *relation between the relative orientation and the grain boundary energy*. Two questions come to mind:
  - Are there *any other low-energy orientations* besides the rotations around **<111>** or **<110>** that produces the low-energy twin boundary?
  - Is there a way to minimize the energy of a grain boundary that is close to, but not exactly in a low energy orientation (some analogon to the facetting of the planes)?
  - This is not an easy question, because the crystal *does not have an option of changing the orientation relationsship*. In principle it would be possible, but it would imply moving a lot of atoms - all the atoms inside a grain - and that is rather unlikely to occur.
- Answering these questions will lead us to an important theory for the structure of grain boundaries (and phase boundaries) which will be the subject of the next sub-chapter.

7.1.2 The Coincidence Site Lattice

- If we look at the infinity of possible orientations of two grains relative to each other, we will find some *special orientations*. In two dimensions this is easy to see if we rotate two lattices on top of each other.
  - You can watch what will happen for a hexagonal lattice lattice by activating the link.
- A so-called **Moirée pattern** develops, and for certain angles *some* lattice points of lattice **1** *coincide* exactly with *some* lattice points of lattice **2**. A kind of superstructure, a **coincidence site lattice** (***CSL***), develops. A question comes to mind: Do these special coincidence orientations and the related **CSL** have any significance for grain boundaries?
- Lets look at our paragon of grain boundaries, the twin boundary:



- Shown are the two grains of the preceding twin boundary, but superimposed. Coinciding *atoms* (in the projection) are marked red. However, this might be coincidental (excuse the pun), because the *atoms* in this drawing are not all in the drawing plane. Note that it is *not* relevant if the boundary itself is coherent or not - only the orientation of the grains counts.
- And once more, note that the lattice is not the crystal! We are looking for coinciding *lattice points* - not for coinciding atom positions (but this may be almost the same thing with simple crystals).

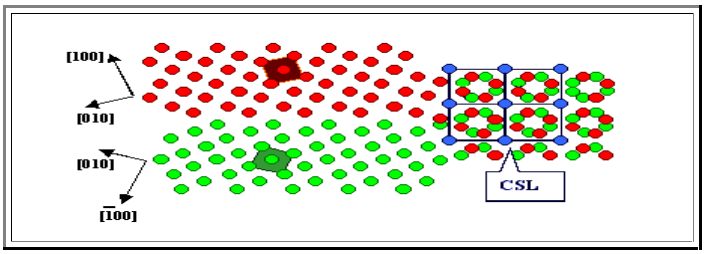


- So in this picture the same situation is shown for the **fcc** *lattice* belonging to the grain boundary. Again, coinciding *lattice points* are marked red and a (two-dimensional) elementary cell of the **CSL** is also shown in red. The two (three-dimensional) elementary cells of the **fcc** lattices are also indicated.
- It is definite from this picture that the the twin boundary belongs to the class of boundaries with a coincidence relation between the two lattices involved.

- From the animation in the link above it was clear that many coincidence relations exist for two identical two-dimensional lattices. In order to be able to extend the **CSL** consideration to three dimensions and to generalize it, we have to classify the various possibilities. We do that by the following definition:

> ***Definition:***
> The relation between the number of lattice points in the unit cell of a **CSL** and the number of lattice points in a unit cell of the generating lattice is called Σ (Sigma); it is the unit cell volume of the **CSL** in units of the unit cell volume of the elementary cells of the crystals.

  - A given Σ specifies the relation between the two grains unambiguously - although this is not easy to see for, let's say, two orthorhombic or even triclinic lattices.
- If we look at the twin boundary situation above, we see that $\Sigma_{twin} = 3$ (you must relate the two-dimensional lattices her; one is pointed above out in black!). For the three-dimensional case we still obtain **Σ = 3** for the twin boundary, so we will call twin boundaries from now on: ***Σ3 boundaries***.
  - A **Σ1** boundary thus would denote a perfect (or nearly perfect) crystal; i.e. no boundary at all. However. boundaries relatively close to the **Σ1** orientation are all boundaries with only *small* misorientations called "**small-angle grain boundaries**" - and they will be subsumed under the term **Σ1** boundaries for reason explained shortly.
  - Since the numerical value of Σ is always odd, the twin boundary is the grain boundary with the most special coincidence orientation there is, i.e. with the *largest number of coinciding lattice points*.
- Next in line would be the **Σ5** relation defining the **Σ5** *boundary*. It is (for the two-dimensional case) most easily seen by rotating two square lattices on top of each other.



- This also looks like a pretty "fitting" kind of boundary, i.e. a low energy configuration.

A suspicion arises: Could it be that grain boundaries between grains in a **CSL** orientation, especially if the Σ values are low, have particularly small grain boundary energies?

The answer is: *Yes,* but... . And the "but" includes several problems:

- Most important: How do we get an answer? Calculating grain boundary energies is still very hard to do from first principles (Remember, that we can't calculate melting points either, even though its all in the bonds). First principles means that you get the exact positions of the atoms (i.e. the atomic structure of the boundary and the energy). Even if you guess at the positions (which looks pretty easy for a coherent twin boundary, but your guess would still be wrong in many cases because of so-called "rigid body translations"), it is hard to calculate reliable energies.

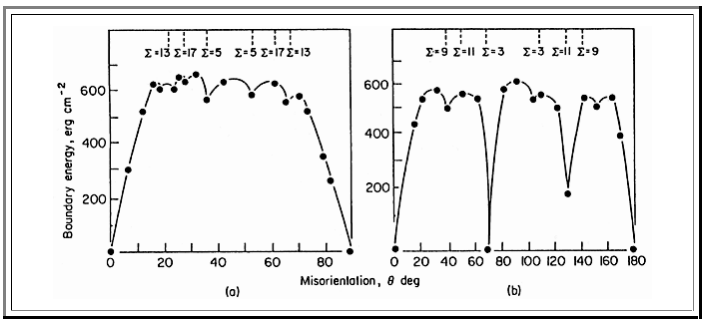
So we are left with experiments. This involves other problems:

- How do you measure grain boundary energies?
- How do you get the orientation relationship?
- How do you account for the part of the energy that comes from the habit plane of the boundary - after all, a coherent twin (habit plane = **{111}**) has a much smaller energy than an incoherent one?

Getting experimental results appears to be rather difficult or at least rather time consuming - and so it is!

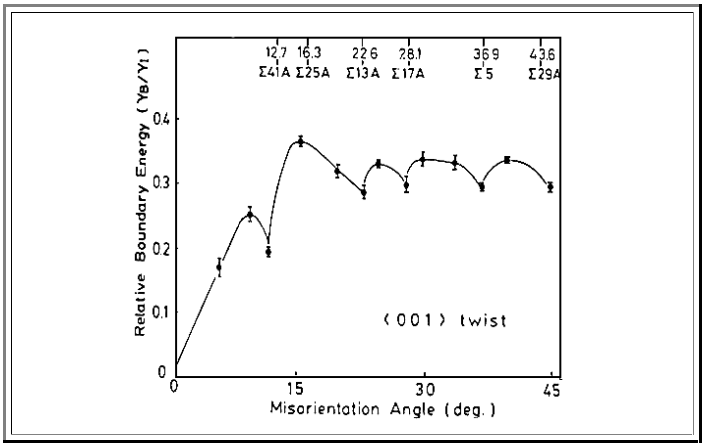
- Nevertheless, results have been obtained and, yes, low Σ boundaries tend to have lower energies than average.
- However, the energy does not correlate in an easy way with Σ; it does not e.g. increase monotonously with increasing Σ. There might be some Σ values with especially low energy values, whereas others are not very special if compared to a random orientation.

The result of (simple) calculations for special cubic geometries are shown in the picture:



- Shown is the calculated (**0°K**) energy for symmetric tilt boundaries in Al produced by rotating around a **<100>** axis (left) or a **<110>** axis (right). We see that the energies are lower, indeed, in low Σ orientations, but that it is hard to assign precise numbers or trends. Identical Σ values with different energies correspond to identical grain orientation relationships, but different habit planes of the grain boundary.

The next figure shows grain boundary energies for twist boundaries in **Cu** that were actually measured by Miura et al. in **1990** (with an elegant and ingenious, but still quite tedious method).

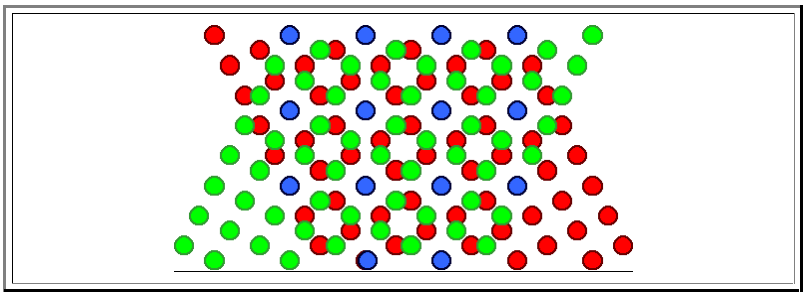


- Clearly, some Σ boundaries have low energies, but not necessarily all.

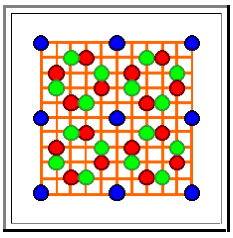
- Nevertheless, in practice, you tend to find low Σ boundaries, because (probably) all low energy grain boundaries are boundaries with a defined Σ value. And these boundaries may have special properties in different contexts, too.
  - The link shows the critical current density (at which the superconducting state will be destroyed) in the high-temperature superconductor **$YBa_2Cu_3O_7$** with intentionally introduced grain boundaries of various orientations and **HRTEM** image of one (facetted) boundary. It is clearly seen that the critical current density has a pronounced maxima which corresponds to a low Σ orientation in this (Perovskite type) lattice.
- However, despite this or other direct evidence for the special role of low Σ boundaries, the most clear indication that low Σ boundaries are preferred comes from innumerable observations of a different nature altogether - the observation that grain boundaries very often contain **secondary defects** with a specific role: They correct for small deviations from a special low Σ orientation.
  - In other words: Low Σ orientations must be preferred, because otherwise the crystal would not "spend" some energy to create defects to compensate for deviations.
- If we accept that rule, we also have an immediate rule for preferred habit planes of the boundary:
  - Obviously, the best match can be made if as many **CSL** points as possible are contained in the plane of the boundary. This simply means:
  - Preferred grain boundary planes are the closest packed planes of the corresponding **CSL** lattice.
- We will look at those grain boundary defects in the next sub-chapter.

7.1.3 The DSC Lattice and Defects in Grain Boundaries:

- Grain boundaries may contain special defects that *only* exist in grain boundaries; the most prominent ones are **grain boundary dislocations**. Grain boundary dislocations are linear defects with all the characteristics of lattice dislocations, but with very specific Burgers vectors that can *only* occur in grain boundaries.
  - To construct grain boundary dislocations, we can use the universal Volterra definition.We start with a "low Σ" boundary and make a cut in the habit plane of the boundary. The cut line, as before, will define the dislocation line vector ***l*** which by definition will be contained in the boundary.
  - Now we displace one grain with respect to the other grain by the Burgers vector ***b*** so as to preserve the *structure of the boundary* everywhere except around the dislocation line. In other words: the structure of the boundary after the shift looks exactly as before the shift.
- What does that mean? What is the "structure of the boundary" and how do we preserve it?
  - Well, we have a **CSL** on both sides of the boundary. We certainly will preserve the structure of the boundary if we shift by a translation vector of the **CSL**, i.e. by a rather large Burgers vector. We than would preserve the coincidence site *lattice* - which is fine, but far too limited. We already preserve the structure of the boundary if we simply *preserve the coincidence*!
  - It is best to illustrate what this means with a *simple animation*: Two superimposed lattices form a ***CSL** marked in blue*. The red lattice moves to the left, and at first there are no more coincidences of lattice points - the **CSL** has disappeared and we have a different structure. However, after a short distance of shifting - far smaller than a lattice vector of the **CSL**, *coincidence points* appears and we have a **CSL** again - but with the coincidence *points* now in different positions.

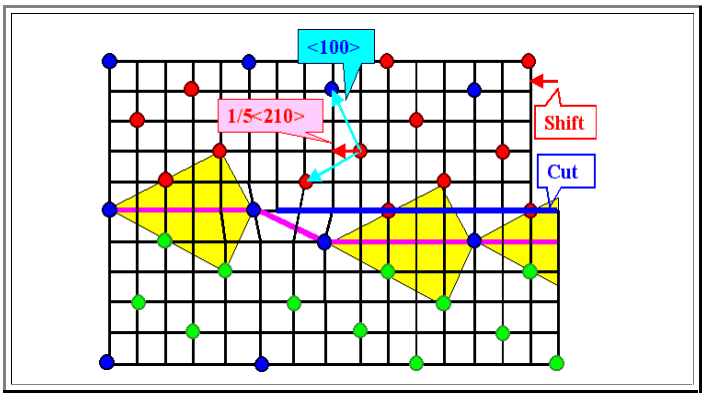


  - We found a displacement vector that preserved the structure of the boundary - sort of experimentally. There are others, too, and the possible displacements vectors that conserve the **CSL** obviously are not limited to vectors of the crystal lattice; *they can be much smaller*. This we can generalize:
- The set of all possible displacement vectors which preserve the **CSL** defines a *new kind of lattice*, the so-called ***DSC*-lattice**. The abbreviation "**DSC**" stands for "*Displacement Shift Complete*", not the best of possible names, but time-honored by now.
  - A better way of thinking about it would be to interprete the abbreviation as "*Displacements which are Symmetry Conserving*". Displacing one grain of a grain boundary with a **CSL** by a vector of the corresponding DSC lattice thus preserves the structure of the boundary because it preserves the symmetries of the **CSL**. We now conclude:
  - *Translation vectors of the **DSC** lattice are possible Burgers vectors $\boldsymbol{b}_{GB}$ for grain boundary dislocations*. As for lattice dislocations, only the smallest possible values will be encountered for energetic reasons.
  - Grain boundary dislocations constructed in this way *by (Volterra) definition*, have most of the properties of real dislocations - just with the added restriction that they are confined to the boundary. Strain- and stress field, line energy, interactions, forming of networks - everything follows the same equations and rules that we found for lattice dislocations.
  - It remains to be seen how the **DSC**-lattice can be constructed. From the illustration it is clear that every vector that moves a lattice site of grain **1** to a lattice site of grain **2** is a **DSC**-lattice vector. This leads to a simple "*working*" definition:
- The **DSC**-lattice is the *coarsest sub-lattice* of the **CSL** that has *all* atoms of *both lattices* on its lattice points. Most lattice points of the **DSC**-lattice, however, will be empty



  - This is the **DSC**-lattice for the animation above. Its easy enough to obtain, *but*:
  - A formal and general definition of the **DSC** lattice (including near **CSL** orientations) is one of the most difficult undertakings in grain boundary theory. If you love tough nuts, turn to chapter 7.3 and proceed.

- Any translation of one of the two crystals along a vector of the orange DSC-lattice will keep the **CSL**, but will generally shift its origin. Only if a **DSC** vector is chosen that is also a vector of the **CSL**, will the origin of the **CSL** remain in place.
- Looking back at the **Σ5** boundary from before, we now can enact the cut and the displacement procedure and generate a picture of the dislocations that must result. The result contains a little surprise and is shown in cross-section below:



- The cut was made from the right. The top crystal (red lattice points) was shifted by a unit vector of the **DSC** lattice, which is a **1/5<210>** vector in both crystal lattices in this case. The second crystal (green lattice points) was left completely unchanged. The coincidence points are blue. We observe two somewhat surprising effects:
- The boundary plane (as indicate by the pink line) after the shift is not identical with the plane of the cut
- The **CSL** has an interruption in both grains - it doesn't fit anymore. Disturbing - *but totally unimportant*. The **CSL**, after all, is *totally meaningless* for real crystals - the (mathematical) coincidence points *in the grains* have no significance for the grains. The *only* significance of a coincidence orientation is that it provides an especially good fit of the two grains at a boundary, i.e. it allows for a particularly favorable boundary *structure*. And the *structure* of the grain boundary is unchanged by the introduction of the grain boundary dislocation, except around its core region. This is indicated by the characteristic diamond shapes (yellow) in the picture above that can be taken as the hallmark of this **Σ5** structure.
- Think about it! Finding the yellow diamonds is the practical way of finding the position of the boundary. However you define the position - you will find the preserved structure as expressed in the yellow diamonds here.
- Introducing the grain boundary dislocation thus had the unexpected additional effect of introducing a **step** in the grain boundary. Some atoms *had to be changed* from green to red to obtain the structure, but that again is an *artifact of the representation*. Real atoms are all the same; they do not come in green and red and do not care to which crystal they belong.

We see that the recipe works: Dislocations in the **DSC** lattice preserve the structure of the boundary; they leave the coincidence relation unchanged. *However*, they also *may* introduce steps in the plane of the boundary -we cannot yet be sure that this always the case.

- Note that is not directly obvious how the step relates to the dislocation, i.e. how the vector describing the step can be deduced from the **DSC** lattice vector used as Burgers vector. (If you see an obvious relationship - please tell me. I'm not aware of a simple formula applicable in all cases).
- Note also: While many (if not all) grain boundary dislocations are linked with a step, the reverse is not true: There are many possible steps in a boundary that do not have *any* dislocation character. More to that in chapter 8.3

The extension to *three dimensions* is obvious, but also a bit mind-boggling. Still, some general rules can be given

- The larger the elementary cell of the **CSL**, the smaller the elementary cell of the **DSC**-lattice!
- If you suspected it by now: The **DSC** lattice is the reciprocal lattice (in space) of the **CSL**.
- The volumes of **CSL**, crystal lattice and **DSC** lattice relate as $\Sigma : 1 : \Sigma^{-1}$ for cubic crystals.

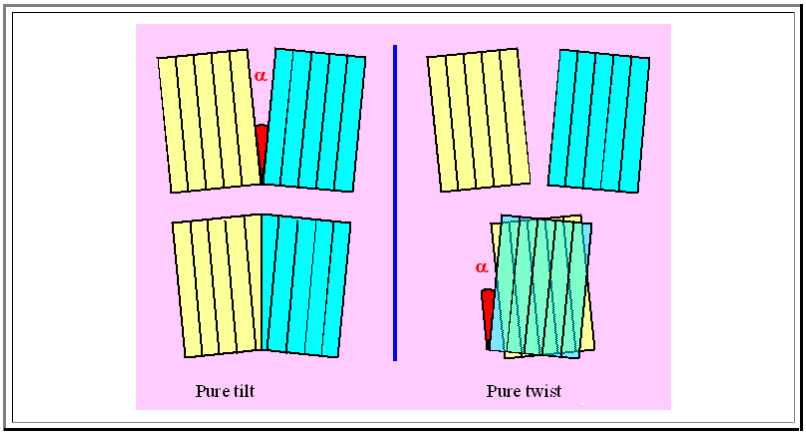
What are all these lattices good for? The main import is:

- A grain boundary between two grains that is *close to, but not exactly at* a low-energy (=low Σ) orientation may decrease its energy if grain boundary dislocations with a Burgers vector of the **DSC** lattice belonging to the low-Σ orientation are introduced so that the dislocation free parts are now in the *precise **CSL** orientation* and all the misalignment is taken up by the grain boundary dislocations.
- We will see how this works in the next sub-chapter.

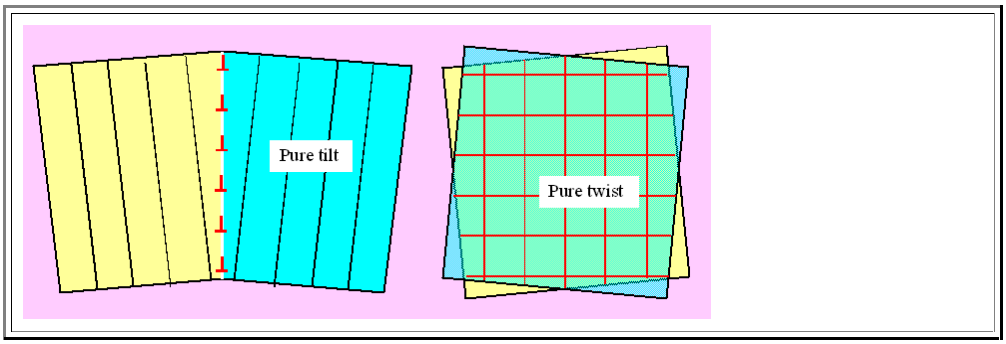
7.2 Grain Boundary Dislocations

7.2.1 Small Angle Grain Boundaries and Beyond

- The determination of the precise dislocation structure needed to transform a *near-coincidence* boundary into a *true* coincidence boundary with some superimposed grain boundary dislocation network can be exceedingly difficult (to you, not to the crystal), especially when the steps possibly associated with the grain boundary dislocations must be accounted for, too.
  - Nevertheless, the structure thus obtained is what you will see in a **TEM** picture - the crystal has no problem whatsoever to "solve" this problem!
- In order to get familiar with the concept, it is easiest, to consider the environment of the **Σ= 1** grain boundary, i.e. the boundary between two crystals with *almost* identical orientation.
  - This kind of boundary is known as "**small-angle grain boundary**" (***SAGB***) , or, as already used above as "**Σ1** boundary".
  - We can easily imagine the two extreme cases: A pure tilt and a pure twist boundary; they are shown below.



  - Obviously, we are somewhat off the **Σ 1** position. Introducing grain boundary dislocations now will establish the exact **Σ1** relation between the dislocations (and something heavily disturbed at the dislocation cores). The **DSC**-lattice as well the **CSL** are identical with the crystal lattice in this case, so the grain-boundary dislocations are simple lattice dislocations.
  - Introducing a sequence of edge dislocations in the tilt case and a network (not necessarily square) of screw dislocations in the twist case will do the necessary transformation; this is schematically shown below



- This may not be directly obvious, but we will be looking at those structures in great detail in the next paragraph. Here we note the important points again:
  - Between the dislocation lines we now have a *perfect **Σ1** relation* (apart from some elastic bending).
  - All of the misfit relative to a perfect Σ orientation is concentrated in the grain boundary dislocations.
  - We thus *lowered* the grain boundary energy in the area between the dislocations and *raised* it along the dislocations - there is the possibility of optimizing the grain boundary energy. The outcome quite generally is:
- Grain boundaries containing grain boundary dislocations which account for small misfits relative to a preferred (low) Σ orientation, are in general preferable to dislocation-free boundaries.
  - The Burgers vectors of the grain boundary dislocations could be translation vectors of one of the crystals, but that is energetically not favorable because the Burgers vectors are large and the energy of a dislocation scales with Gb^2 and there is a much better alternative:
  - The dislocations accounting for small deviations from a low Σ orientation are *dislocations in the **DSC** lattice* belonging to the *CSL lattice* that the grain boundary Σ endeavors to assume. Why should that be so? There are several reasons:

1. Dislocations in the **DSC** lattice belong to *both* crystals since the **DSC** lattice is defined in both crystals.
2. Burgers vectors of the **DSC** lattice are smaller than Burgers vectors of the crystal lattice, the energy of several **DSC** lattice dislocations with a Burgers vector sum equal to that of a crystal lattice dislocations thus is always much smaller. With
   Σ_i ***b***$_i$**(DSC) = *b*(Lattice)** , we always have
   Σ_i ***b***$_i^2$**(DSC) << b^2(Lattice)**.
   This is exactly the same consideration as in the case of lattice dislocations split into partial dislocations.
3. A dislocation arrangement with the same "*Burgers vector count*" along some arbitrary vector ***r*** produces exactly the same displacement (remember the basic Volterra definition and the double cut procedure).

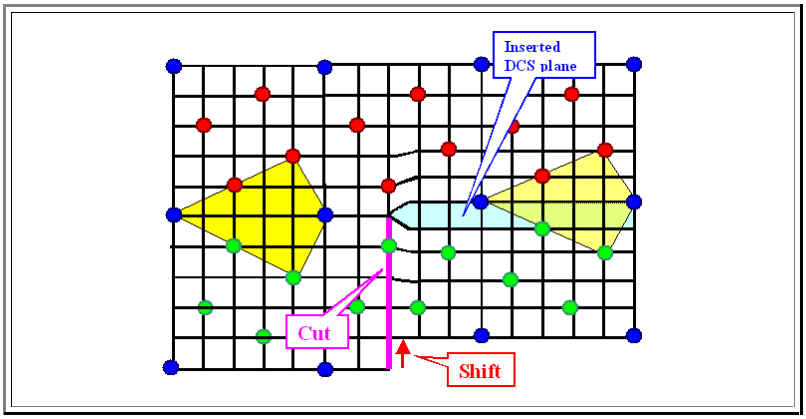
In other words: We can always imagine a low angle boundary of crystal lattice dislocations that produces exactly the small misorientation needed to turn an arbitrary boundary to the nearest low Σ position and superimpose it on this boundary.

Next, we decompose the crystal lattice dislocations into dislocations of the **DSC** lattice belonging to the low Σ orientation.

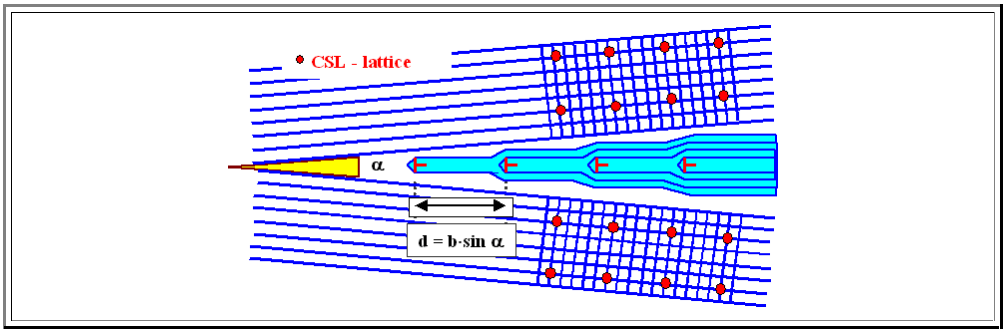
This will be the dislocation network that we are going to find in the real boundary!

Lets illustrate this. First we construct another kind of **DSC** lattice dislocation, very similar, but different to the one we had before. The coincidence points are marked in blue, atoms of the two crystal lattices in green and red.

The plane of the cut now is perpendicular to the boundary and extends, by necessity, all the way to the boundary. We produced a **DSC** edge dislocation with a Burgers vector perpendicular to the boundary plane (and a step of the boundary plane).



If we were to repeat this procedure at regular intervals along the boundary, we obtain the structure schematically outlined below.



In essence, we superimposed a tilt component with a tilt angle α that for small angles is given by is given by

$$\alpha = \frac{d}{b}$$

with ***d*** = spacing of the **DSC** lattice dislocations and ***b*** = Burgers vector of the **DSC** lattice dislocations.

In short, we can do everything with **DSC** lattice dislocations in a grain boundary that we can do with crystal lattice dislocations. This leads to the crucial question alluded to before:

How do we *calculate* the **DSC**-lattice? As an example for the most general case of grain boundaries in triclinic lattices? Or even worse: For *phase boundaries* between two different lattices (with different lattice constants)?

The answer is: Use the "**Bollmann theory** or " **O-lattice theory**" - it covers (almost) everything.

However, unless you are willing to devote a few months of your time in learning the concept and the math of the **O**-lattice theory, you will encounter problems - it is not an easy concept to grasp.

We will deal with the O-lattice theory in a backbone II section, here we note that the most important cases have been tabulated. Some solutions *for fcc crystals* are given in the table:

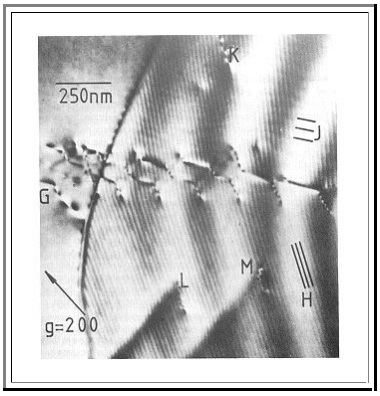
Σ	Generation	***b*** **from DSC-lattice**
1	"Small-angle GB"	a/2 <110>, possibly split into partials
3	Twin	a/6 <112>, a/3 <111>
5	37° around <100>	a/10 <130>
9	39,9° around <110>	1/18<114>, 1/9<122>, 1/18<127>,
19	26,5° around <110>	a/38 <116>, a/19<133>, a/19 <10,9,3>
41	12,7° around <100>	a/82<41,5,4>, a/82<910>, ...

Interestingly (and very satisfyingly), the **DSC** lattice vectors belonging to the **Σ = 3** boundary are our old acquaintances, the partial Burgers vectors associated with stacking faults in the crystal lattice.

This is but natural - a **Σ = 3** twin boundary is after all a very close relative to stacking faults.

Now a question might come up: **Σ = 41** is not exactly a "**low Σ**" value; and Burgers vectors of **a/82<41,5,4>** appear to be a bit odd, too. So does this still make sense? Are boundaries close to a **Σ41** orientation still special and bound to have grain boundary dislocations?

Only the experiment can tell . The following **TEM** picture shows a **Σ 41** grain boundary (from Dingley and Pond, Acta Met. **27**, 667, 1979)



A network of grain boundary dislocations with Burgers vectors **b = a/82 <41,5,4>** and an average distance of **20 nm** is visible. The two sets of dislocations run parallel to the lines indicated by **H** and **J**.

Sorry, but it is there, even at **Σ = 41**. Why - we do not really know, although Bollmann theory does provide an answer on occasion.

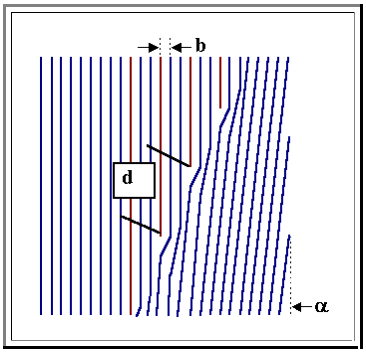
Obviously, if you want to understand the structure of grain boundaries, you must accept the concept of grain boundary dislocations even at rather large values of Σ and correspondingly low values of the Burgers vectors.

In the next paragraph we will study some cases in more detail.

7.2.2 Case Studies: Small Angle Grain Boundaries in Silicon I

First we will take a close look at some small angle grain boundaries in Silicon. Whereas they are the most simple boundaries imaginable, they are still complex enough to merit some attention. They are also suitable to demonstrate a few more essential properties of grain boundary dislocations.

- Lets first look at a pure tilt boundary as outlined in the preceding paragraph. Below is shown how edge dislocations can accommodate the misfit relative to the **Σ = 1** orientation (for a boundary plane that contains the dislocations lines).

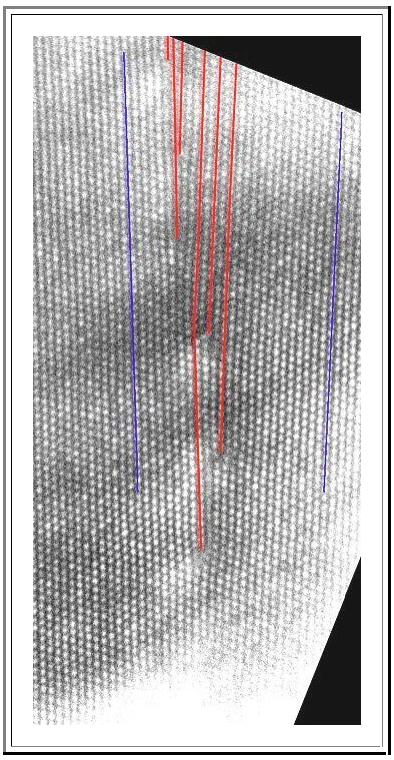


- The distance ***d*** between the dislocations is for small tilt angles α as before given by

$$\alpha = \frac{d}{b}$$

- This is a simple version of a general relation betwen Burgers vectors and misorientation in small angle grain boundaries called **Franks formula** (more correctly Frank-**Bilby** formula).

- In real life it looks slightly more complicated - but not much:

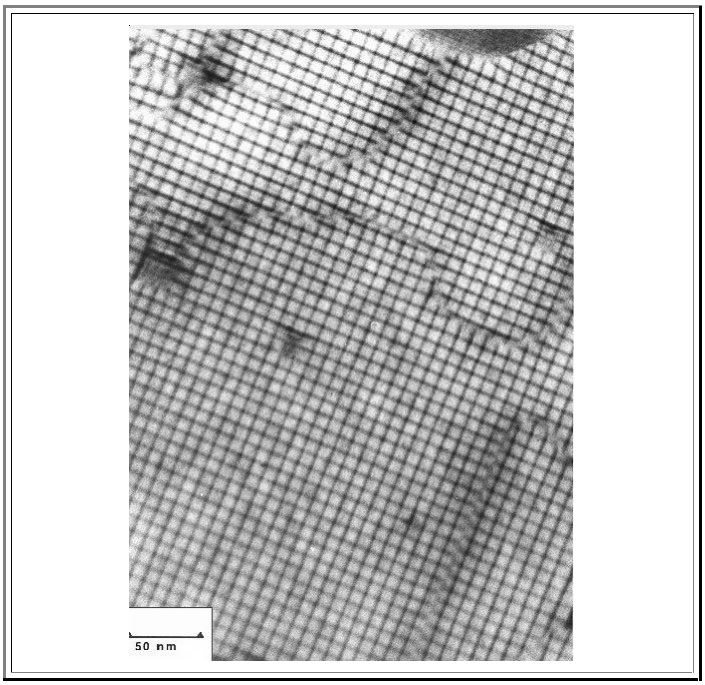


- This is an early **HRTEM** of a small angle tilt boundary in **Si**. The red lines mark the edge dislocations, the blue lines indicate the tilt angle.

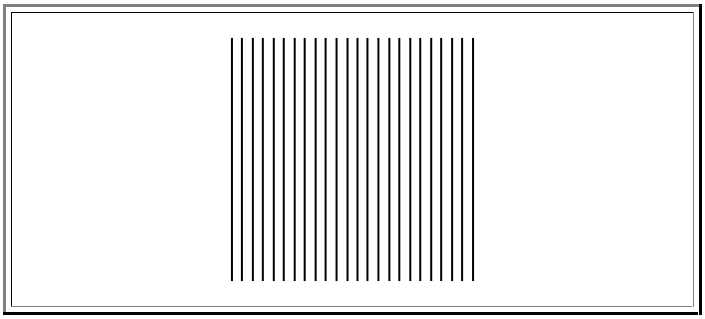
- This picture nicely illustrates that we have indeed a **Σ1** relationship in the area between the dislocations, i.e. a perfect crystal. The dislocations are not all in a row, but that does not really matter.

Next, we look at twist boundaries.

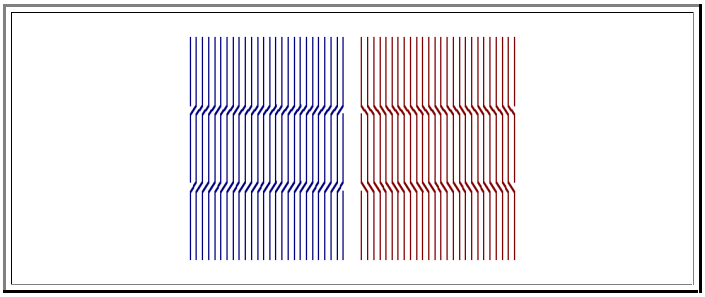
- These and some of the other boundaries were artificially produced to study the structure. Two specimens of **Si** with a desired orientation relationship were placed on top of each other and "sintered" or "welded" together at high temperatures. This process, first called "sintering" is now known as "**waferbonding**" and used for technical applications.
- The result for a slight twist between **{100}** planes is shown in the next picture:



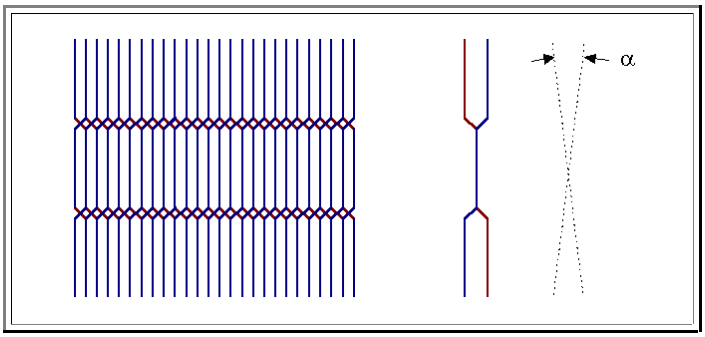
- This is a remarkable picture. As ascertained by contrast analysis, it shows a square network of *pure screw dislocations*. The picture is remarkable not only because it shows a rather perfect square network of screw dislocations, but because it is obviously a **bright field** **TEM** micrograph, however with a resolution akin to **weak-beam** conditions.
  - Pictures like this one are obtained by orienting the specimen close to a **{100}** (or, in other cases, **{111}**) orientation thus exciting many reflections weakly. All dislocations are then imaged, but the detailed contrast mechanism causing the superb resolution is not too clear.
- Lets first find out why a network like this can produce the required twist. We do this in reverse order, i.e. we will construct a screw dislocation network in a perfect crystal and see what it does.
  - We start by looking at **{100}** lattice planes below and above the (future) grain boundary plane. They are exactly on top of each other, we obtain a (trivial) picture of a formal low-angle twist boundary with twist angle α **= 0°**.



  - Now we introduce two screw dislocations running from left to right. Referring to the same kind of picture in chapter 5, we see that the lattice planes below the screw dislocations are bent to the right (blue lines), above the screw dislocations to the left:



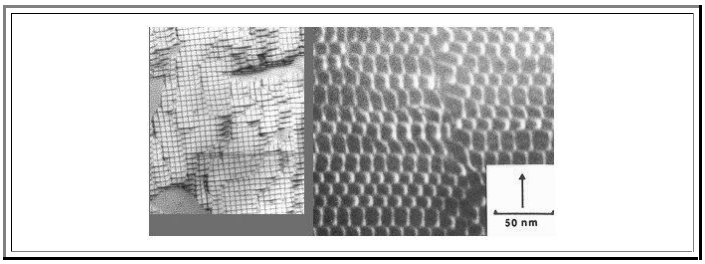
With many dislocations, the average orientation of the lattice planes below the small angle grain boundary will rotate to the right, above to the left. The combined effect is shown below.



- If we want to rotate not just one set of lattice planes, but all of the top part of the crystal, we need at least a second set of screw dislocations. This produces a screw dislocation network of the kind shown in the **TEM** micrograph above.
  - The relation between the twist angle α and the dislocation spacing ***d*** is again a simple version of the general case given in Franks formula:

$$\alpha = \frac{b}{d}$$

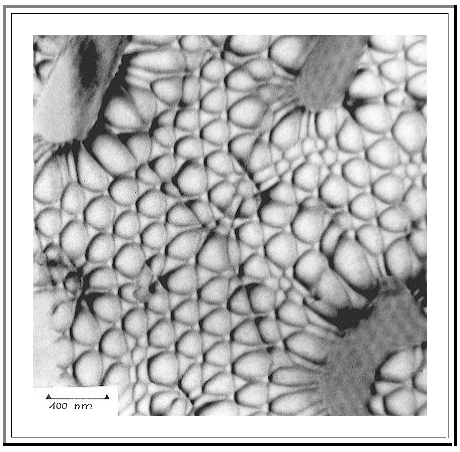
  - A detailed drawing of this dislocation network structure can be viewed in the link.
  - With luck, it is possible to image the lattice plane in a **HRTEM** micrograph. The link shows examples - the only **HRTEM** image of screw dislocations obtained so far.
- The exact geometry of the network for the same twist angle α in an *arbitrary* twist boundary depends on many factors:
  - The *twist angle α* which determines the spacing between the dislocations. For **Σ = 1** boundaries it simply the twist angle itself, for arbitrary boundaries it is the twist angle needed to bring the boundary to the closest low **Σ** orientation
  - The *Burgers vectors* of the dislocations. Even in small-angle grain boundaries they could be perfect, or split into partials. In arbitrary boundaries they must be grain boundary dislocations with a Burgers vector of the proper **DSC** lattice. Note that a network of *grain boundary* screw dislocations simply superimposes some twist to whatever orientation the boundary has without those dislocations.
  - The *type of the dislocations*. For an arbitrary twist plane, the Burgers vectors of the possible dislocations are not necessarily contained in the grain boundary plane; the required pure screw dislocations do not exist. In this case mixed dislocations must be used with a component of the Burgers vector in the grain boundary plane.
  - The *symmetries* of the two crystal planes in contact at the grain boundary - even low-angle twist boundaries with a twist around the **<100>** axis can be joined on planes other than **{100}**.
  - The complications that may arise because the (perfect) *dislocations split into partials*. Obviously, that has not happened in the case shown above. The reason most likely is that the splitting would have to be on two different **{111}** planes inclined to the boundary plane (look at your Thompson tetrahedron!) which leads to very unfavorable knot configurations. Since the distance between dislocations is of the same order of magnitude as the typical distance between partials, we do not observe splitting into partials or dissociation of the knots.
- We now can understand the very regular square network shown in the picture above - it is really about the most simple structure imaginable.
  - But we still need to explain the interruptions in the network; the lines along which the net is shifted. In fact, to see a very regular network like this you must be pretty lucky; more often than not often (artificially) made twist grain boundaries look like this:



  - Both pictures show the result of the attempt to make a pure twist boundary. Whereas the left one still looks like the picture above - just with more interruption of the network - the right one does not convey the impression of a square network at all.
  - The answer is that these grain boundaries must accommodate more that just a pure twist: There is also a tilt component and the plane of the grain boundary is not exactly **{100}**. We will pick up this subject again in the next paragraph; more information about the right-hand side picture can be found in the link.

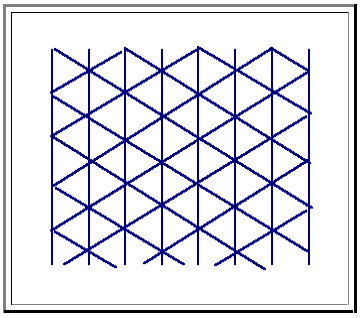
7.2.3 Case Studies: Small Angle Grain Boundaries in Silicon II

- The recipe for "making" small angle grain boundaries in Silicon given in the preceding paragraph can be used for twist boundaries on any plane, besides the **{100}** plane the **{111}** planes are particular interesting.
  - The structure will be much more complicated and serves to illustrate the importance of the grain boundary plane for given orientations. The picture below is a bright-field **TEM** micrograph (obtained under the specific bright-field conditions that rival weak-beam resolution mentioned before) and shows all dislocations present.

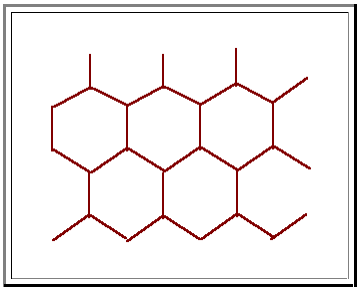


  - This is also an example of what may happen to you when you sit down at an electron microscope with your specimen and start to look at it.
  - You know what to expect (a small angle twist grain boundary) in general, but now you see fascinating things - can you understand what you see?
  - And, in extrapolation, can you understand what you see if you do not know beforehand what to expect?

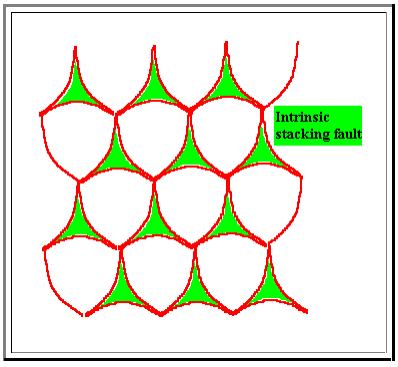
- Well, we can understand most of the structure seen above. Lets construct it step by step



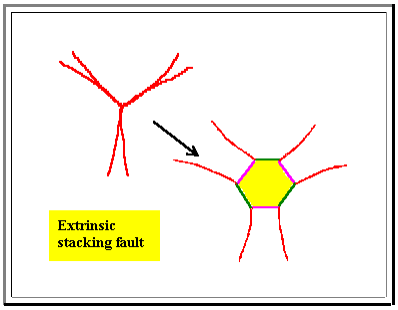
  - There must be a network of screw dislocations with **b = a/2<110>**. Since three Burgers vectors of this kind are contained in a given **{111}** plane, we expect a hexagonal network as shown on the left.



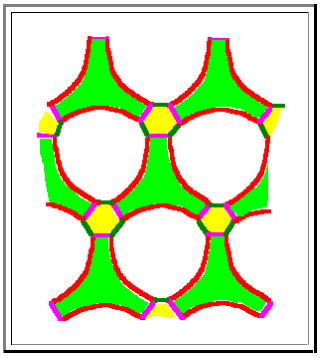
  - The knots where six dislocations meet can not be expected to be stable; we would expect a splitting leading to the honeycomb pattern illustrated on the left (with a changed scale for clarity).



  - In contrast to the {100} twist boundary, the dislocations now can split into partial dislocations in the plane of the boundary, we expect that the dislocations are split in this case. Working through the geometry we see that everything fits at one knot, it can easily be extended in the way shown. This optimizes the energy gain by large separations between the partials while at the same time keeping the stacking fault area small.
  - The "constricted" knots now look "funny" - again **6** dislocations meet at one point. Can we split the knot to something more favorable?

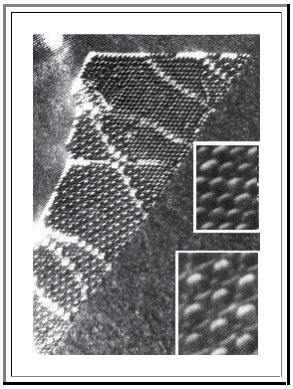


- Indeed, we can, as shown on the left. However, we *only* can do this by introducing more Shockley dislocations and extrinsic stacking faults



- Putting everything together we obtain a network of Shockley dislocations that corresponds exactly to what we see in the regular parts of the micrograph above. The exact geometry, of course, depends mainly on the stacking fault energies - here we may find differences if we would look at similar grain boundaries in other **fcc** materials.

- It remains to explain the various non-regularities of the picture.
  - Most conspicuous are the large "blobs" with just a trace of some hexagonal structure. They are simply **SiO_2** precipitates left over from the welding process; the hexagonal structures are **Moirée patterns** that always appear whenever two regular structures are put on top of each other.
- The other irregularities are formed by a superposition of:
  - A few edge dislocations to accommodate some tilt component.
  - Dislocations that moved from somewhere in the crystal into the grain boundary where they were caught and incorporated into the network. These dislocations are called extraneous or **extrinsic grain boundary dislocations** because they are not an integral part of the grain boundary structure.
  - Dislocations needed to accommodate steps, i.e. changes of the grain boundary plane measuring a few atomic distances.
- It is not always clear or easily analyzed exactly what it is that you see. Especially the connection between steps and intrinsic or extrinsic dislocation is, in general, quite complicated, because on the one side most, but not all grain boundary dislocations automatically introduce steps, while on the other side most, but not all steps introduce dislocations. We will deal with that matter in more detail when dealing with phase boundaries.
- But we are not yet done with the low-angle twist boundary on **{111}**. The micrograph above shows only part of the structure. The micrograph below shows more:



- The lower inset shows a magnified view of the network from the lower half of the boundary; the upper inset from the upper half. (Click on the picture for an enlargement and more information),
- Whereas the lower part shows the network discussed above, the upper part shows something new: A rather simple network with reduced spacing. Detailed analysis reveals that the dislocations in the upper part have Burgers vectors ***b* = a/6<112>**, but they are **not** proper Shockley partials, because there are no stacking faults between the dislocations.
- They are rather dislocations in the **DSC** lattice of a **$\Sigma = 3$** boundary - in other words, the low angle twist boundary has split into two twin boundaries with a superimposed dislocation network in one of the twin boundaries.
- We may ask: Why? And in which twin boundary are the dislocations? Why are they not in the perfect lattice between the boundaries?

- It seems to ever end. But in order not to have too much detail in the main backbone part, these complications will be discussed in an advanced section.

7.2.4 Generalization

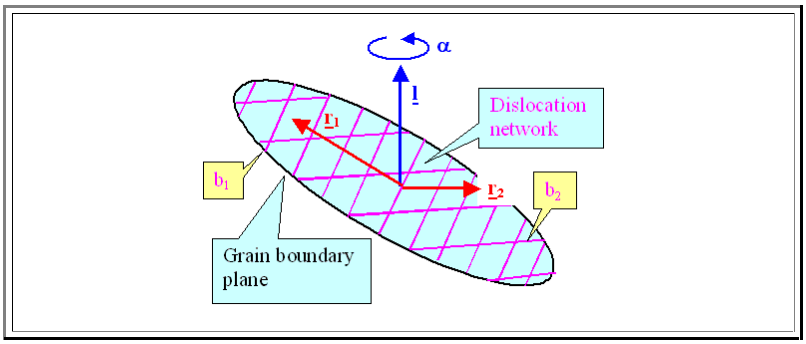
- The relation between the spacing of the dislocations and the tilt- or twist angle in the special cases given was simple enough - but what about arbitrary small angle grain boundaries with twist and tilt components? What kind of dislocation structure and what geometry should be expected?
  - As we have seen, the detailed structure of the network can be quite complicated and depends on materials parameters like stacking fault energies. We can not expect to have a simple formula giving us the answers.
  - The relation giving the *distance* between dislocations in a boundary and the orientation relationship for arbitrary low-angle orientations (meaning that the two rotation angles needed for a general description are both small, lets say ≤ **10^o - 15^o**) was first given by **Frank**. It is *Franks formula* referred to before.
- Franks formula is derived in the advanced section, here we only give the result. The low-angle grain boundary shall be described by:
  - Its dislocation network consisting of dislocations with Burgers vectors $\underline{\boldsymbol{b}}$.
  - An arbitrary vector $\underline{\boldsymbol{r}}$ contained in the plane of the boundary.
  - A (small) angle α around an arbitrary axis described by the (unit) vector $\underline{\boldsymbol{l}}$ (then one angle is enough) that describes the orientation relationship between the grains. We may then represent the rotation by a polar vector $\underline{\mathbf{R}} = \alpha \cdot \underline{\boldsymbol{l}}$
- Franks formula then is:

$$\underline{\mathbf{B}} = (\underline{\boldsymbol{r}} \times \underline{\boldsymbol{l}}) \cdot 2\sin\frac{\alpha}{2}$$

  - with $\underline{\boldsymbol{B}}$ = sum of all the specific Burgers vectors $\underline{\boldsymbol{b}}_i$ cut by $\underline{\boldsymbol{r}}$; i.e. $\underline{\mathbf{B}} = \Sigma_i \underline{\boldsymbol{b}}_i$.
  - Since the formula is formally applicable to any boundary, but does not make much sense for large angles α (can you see why?) we only consider low-angle boundaries. Then we can replace **sin α/2** approximately by **α/2** and obtain the simplified version

$$\underline{\mathbf{B}} = (\underline{\boldsymbol{r}} \times \underline{\boldsymbol{l}}) \cdot \alpha$$

- Let's illustrate this:



  - Shown are two vectors $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ contained in a boundary plane with an arbitrarily chosen dislocation network consisting of two types of dislocations having Burgers vectors $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$.
  - Franks formula ascertains that $(\underline{\boldsymbol{r}} \times \underline{\mathbf{l}}) \cdot \alpha$ equals the sum of the Burgers vectors encountered by $\underline{\boldsymbol{r}}$, i.e. $\underline{\boldsymbol{B}} = 2\boldsymbol{b_2} + 3\boldsymbol{b_1}$ for $\boldsymbol{r_1}$, and $\underline{\boldsymbol{B}} = 3\underline{\boldsymbol{b}}_1$ for $\underline{\boldsymbol{r}}_1$ in the picture above.
  - This is a major achievement, but not overly helpful when you try to find out the geometry of the network for some arbitrary boundary, because their is no simple and unique way of decomposing a sum of Burgers vectors into its individual parts.
- This "simple" formula, however, contains the special cases that we have considered before, and leaves enough room for complications. It does not, however, say anything about *preferred planes or network geometries*. For this one needs the full power of Bollmanns **O**-lattice theory.

- Franks formula is not applicable to large angle grain boundaries because the distance between the dislocations would become so small as to be meaningless.
  - In this case the grain boundary may be viewed as the (dislocation free) "low Σ" boundary closest to the actual orientation with a superimposed low-angle grain boundary formed by dislocations in the corresponding **DSC lattice**.
  - Franks formula then can be used for the low angle part and will give the correct over-all Burgers vector count.
- The precise geometry of the network, however, can become hopelessly complicated because all the additional features, e.g. extrinsic dislocations and steps, are not only still present but become more complicated, too.
  - In addition, some special perversions may evolve, e.g. the splitting of DSC lattice dislocations into partial dislocations in the **DSC** lattice, producing stacking faults in the **DSC** lattice!
  - And we are implicitly talking grain boundaries between cubic crystals! In less symmetric crystals everything is even more complicated - it is time then to study **O**-lattice theory!
- One last feature should be mentioned that now can be understood: The reactions of lattice dislocations that move into grain boundaries. So far, a grain boundary was just seen as an internal surface on which a dislocation can somehow end.
  - Now we know better: It simply decomposes into the intrinsic (**DSC**-lattice) grain boundary dislocations present. This is quite satisfying because the logical problems encountered when thinking more in more detail about how a dislocation "just ends" on a grain boundary. It is also what one sees if looking closely, an example is shown in the illustration.

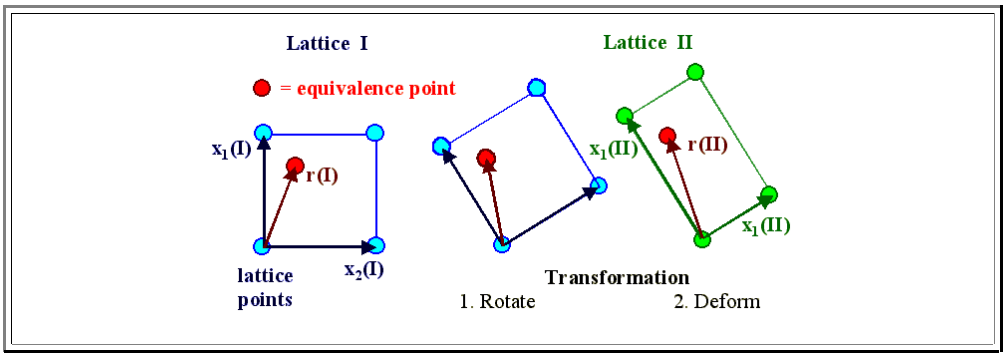
7.3 O-Lattice Theory

7.3.1 The Basic Concept

- Formulas here have plenty of indices, underlining etc. - and we will now give up the *cursive* font normally used for variables because it gets too cumbersome.

Basic Idea

- The Coincidence Site lattice (**CSL**) provided a relatively easy way to grasp the concept of special orientations between grains that give cause for special grain boundaries. With the extension to grain boundary dislocations in the **DSC** lattice, the **CSL** concept became in principle applicable to all grain boundaries, because any arbitrary orientation is "near" a **CSL** orientation. But yet, the **CSL** concept is not powerful enough to allow the deduction of grain boundary structures in all possible cases. The reasons for this are physical, practical and mathematical:
  - The **CSL** by itself is meaningless; meaningful is the special grain boundary structure that is possible if there is a coincidence orientation. The grain boundary structure is special, because it is periodic (with the periodicity of the **CSL**) and contains coincidence points (cf. the picture).
  - But we have no guarantee that periodic grain boundary structures may not exist in cases where no **CSL** exists; i.e. by only looking at **CSL** orientation, we may miss *other special orientations*. That will be certainly true whenever we consider boundaries between different lattices - be it that lattice constants of the same materials changed ever so slightly because one grain has a somewhat different impurity concentration, or that we look at phase boundaries between different crystals. If the lattice constants are incommensurable, there will be no **CSL** at all.
  - As we have seen, even a **CSL** with Σ **= 41** is significant, even so it is virtually unrecognizable as anything special in a drawing. This is an expression of the mathematical condition, that you either have perfect coincidence or none. If two points coincide almost, but not quite, no recognizable **CSL** will be seen. If two lattice points coincide except for, lets say, **0,01 nm**, we certainly would say we have a physical coincidence, but mathematically we have none.
  - The same is true if we rotate a lattice away from a coincidence position by arbitrarily small angles. *Mathematically*, the coincidence is totally destroyed and the situation has completely changed, whereas *physically* an arbitrarily small change of the orientation would be expected to cause only small changes in properties.
  - Only a very small fraction of grain orientations have a **CSL**. The "trick" we used to transform any orientation into a coincidence orientation by introducing grain boundary dislocations in the **DSC** lattice is somewhat questionable: the effect (= **CSL**) comes before the cause (= dislocations in the **DSC** lattice), because at the orientation that we want to change no **CSL** and therefore no **DSC** lattice exists.
- It becomes clear that the main problem lies in the discreteness of the **CSL**. Any useful theory for special grain boundary (and phase boundary) structures must be a *continuum* theory, i.e. give results for continuous variations of the crystal orientation (and lattice type).
  - This theory exists, it is the so-called "**O-lattice theory**" of **W. Bollmann**; comprehensively published in his opus magnus "Crystal Defects and Crystalline Interfaces" in **1970**.
  - The **O**-lattice theory *is not particularly easy to grasp*. (Sorry, but it took me many hours, too). It is well beyond the scope of this hyperscript to go into details. What will be given is the basic concept, the big ideas; together with some formulas and a few examples. You should first read just over it, trying to get the basic ideas, than study it point by point. If you don't get it the first time - don't despair, you are in good if not excellent company!
- There are two basic ideas behind the **O**-lattice theory:
  - **1.** Take a *crystal lattice I* and transform it in any way you like. That means you can not only rotate it into an arbitrary orientation relative to crystal **I**, but also deform it by stretching, squeezing and shearing it. The *crystal lattice II* generated in this way from a simple cubic lattice **I** thus could even be an arbitrarily oriented triclinic lattice.
  - **2.** Now look for coincidence points between lattice **I** and lattice **II**. But do not restrict the search for coinciding *lattice points*, but expand the concept of coincidence to all "*equivalence points*" within two overlapping unit cells. What **equivalence points** are becomes clear in the illustration.



- Points in lattice **I** and lattice **II** are called equivalent, if their space vectors are identical (always in their respective lattice coordinate system).

The Basic Formula

- Let's look at the example illustrated above. Lattice **I** is deformed by first rotating it and then stretching the axis x_1; this produces lattice **II**
  - An arbitrary point within the elementary cell of lattice **I** is described by a vector **r(I)**
  - **r(I)** transforms into a vector **r(II)** by the transformation applied. The point reached within the unit cell of lattice **II** by **r(II)** is then an *equivalence point* to the one in crystal **I**.
- Of course there is more than one equivalence point; there is always an infinite set defined by *one* point *plus* all points reachable by a lattice translation vector **T** from this particular point.
  - Any point **r'(II)** in lattice **II** belonging to the set as defined above can be described in the coordinate system of lattice **II** (defined by the units vectors $\underline{x}_1$**(II)** and $\underline{x}_2$**(II)**) by

$$\underline{r}'(II) = r(II) + \underline{T}(II)$$

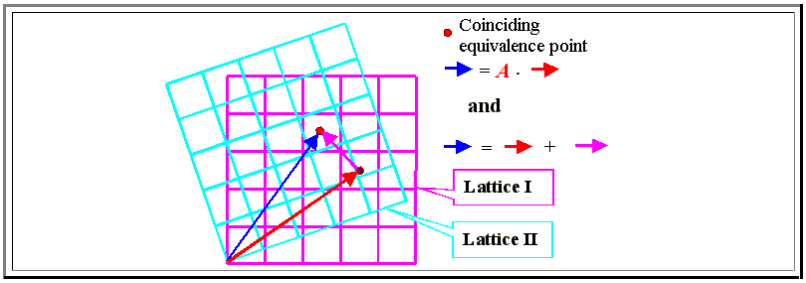
  - With **T**(**II**) = any translation vector of lattice **II**, or

$$\underline{T}(II) = n \cdot \underline{x}_1(II) + m \cdot \underline{x}_2(II)$$

  - And **n, m = 0, ± 1, ± 2, ...**
- All these points are *by definition equivalence points* to the corresponding *set of points* in lattice **I** given by

$$\underline{r}'(I) = \underline{r}(I) + \underline{T}(I)$$

- Let us designate the *set* of all equivalence points defined above in lattice **I** by C_1 and the corresponding set in lattice **II** by C_2 and, for the sake of clarity, all vectors pointing to equivalence points of the respective sets by $\underline{r}(C_1)$ and $\underline{r}(C_2)$.
- If we now look at a certain equivalence point in lattice **II**, it always originated from lattice **I** by the general transformation as shown in the picture below



- The blue lattice was obtained from the pink one by some transformation; in this case by a simple rotation.
  - The dark point with the red vector pointing to it is an arbitrary point in lattice **I** (for the sake of easy recognition about in the center of a lattice **I** cell).
  - After the transformation, it is now the red point at the apex of the blue and pink arrows in lattice **II**. It is still about in the center of a cell in lattice **II**, but for the *particular* transformation shown, it is now also about in the center of a lattice **I** cell - there is (about) a *coincidence of equivalence points*.
  - Let's assume perfect coincidence, then the red point denotes *coinciding* equivalence points, i.e. equivalence points that are "on top of each other".

We need a precise mathematical formulation that gives us the conditions under which *coincidence of equivalence points* occcurs.

- This is easy, we just have to consider that for coinciding equivalence points the blue vector in lattice **II** can be obtained in *two* ways:

  1. By the transformation equation from the corresponding red vector of lattice **I** (valid for *all* equivalence points) or, since the coincidence point belongs to *both* lattices at once, by
  2. adding some translation vector of lattice **I** to the red vector. This is symbolically shown in the picture.

- In formulas we can write for *any* vector in lattice **II** pointing to some equivalent point of the set $\mathbf{C_2}$:

$$\underline{\mathbf{r}}(\mathbf{C_2}) = \mathbf{A}\,\{\underline{\mathbf{r}}(\mathbf{C_1})\}$$

$$\underline{\mathbf{r}}(\mathbf{C_1}) = \mathbf{A^{-1}}\,\{\underline{\mathbf{r}}(\mathbf{C_2})\}$$

- With **A** = transformation matrix (we will encounter examples for **A** later; see also the basic module for matrix calculus) since this simply describes how lattice **II** originates from lattice **I**. This was the first way mentioned above.
- On the other hand, we can obtain new equivalence points in lattice **I**, i.e. other elements of the set $\mathbf{C_1}$ designated by $\underline{\mathbf{e}}(\mathbf{C_1})$ quite generally by the equation

$$\underline{\mathbf{e}}(\mathbf{C_1}) = \underline{\mathbf{r}}(\mathbf{C_1}) + \underline{\mathbf{T}}(\mathbf{I})$$

- We will now use these relations for coinciding equivalence points:

We are looking for coincidences of any one member of the set $\underline{\mathbf{r}}(\mathbf{C_1})$ with any one member of the set $\underline{\mathbf{r}}(\mathbf{C_2})$; any coincidence point thus obtained will be named $\underline{\mathbf{r}}_0$. Since this point, describable in lattice **II by** $\underline{\mathbf{r}}(\mathbf{C_2})$ must be reachable in lattice **I** by first going down $\mathbf{r}(\mathbf{C_1})$ and then adding a translation vector of lattice **I**, we obtain

$$\underline{\mathbf{r}}(\mathbf{C_2}) = \underline{\mathbf{r}}(\mathbf{C_1}) + \underline{\mathbf{T}}(\mathbf{I}) = \underline{\mathbf{r}}_0$$

- Using the transformation equation for lattice **I** from above and substituting it into the above equation yields

$$\underline{\mathbf{r}}_0 = \mathbf{A^{-1}}\,\{\underline{\mathbf{r}}_0\} + \underline{\mathbf{T}}(\mathbf{I})$$

- We wrote $\underline{\mathbf{r}}_0$ instead of $\underline{\mathbf{r}}(\mathbf{C_2})$ because we do not need the distinction between the sets $\mathbf{C_1}$ and $\mathbf{C_2}$ any more because $\underline{\mathbf{r}}_0$ belongs to *both* sets.
- Rearranging the terms following matrix calculus by using the identity or unit transformation matrix **I**, we obtain the **fundamental equation of O-lattice theory**:

$$(\mathbf{I} - \mathbf{A^{-1}})\,\underline{\mathbf{r}}_0 = \underline{\mathbf{T}}(\mathbf{I})$$

What does that equation mean?

- For a given transformation, i.e. for given orientation relationship between two grains, its solution for $\underline{\mathbf{r}}_0$ defines all the coincidence points or **O-points** of the lattices. The coincidence of lattice points is a subset of the general solution for the coincidence of equivalence points.
- The question comes up if there are *any* solution of this equation. Algebra tells us that this requires that the determinant of the matrix, $|\mathbf{I} - \mathbf{A^{-1}}|$, must be $\neq \mathbf{0}$. This will be generally true (but not always), so generally we must expect that solutions exist, i.e. that a **CSL** (= **O**-lattice) for some equivalence points (= **O**-points) exists - for any possible combination of lattice **I** and lattice **II**.

How do we solve the **O**-lattice equation, i.e. obtain the set of **O**-points for a given lattice and transformation? Simply by inverting the matrix we obtain:

$$\underline{\mathbf{r}}_0 = (\mathbf{I} - \mathbf{A^{-1}})^{-1} \cdot \underline{\mathbf{T}}(\mathbf{I})$$

That is all there is to do; it looks easy. If we have a given transformation matrix **A**, the equation above gives us the set of vectors defining the equivalent points, or as we are going to call them, the **O**-points of the two lattices.

- However, the diffusion equations look easy, too, but are not easy to solve. Also, we do not yet know what the solution, the **O**-lattice, really means with respect to grain- or phase-boundaries.

We will look at this more closely in the next paragraph; but first we will discuss a simple example.

Example for Caculating the O-Lattice

To keep the matter simple, we look at a *two-dimensional* situation where a square lattice rotates on top of another one. This will include our former example of the **Σ = 5 CSL** case. (A word of warning: In Bollmanns book are occasional mistakes when it comes to the **Σ5** orientation (which is frequently used for illustrations)).

- The transformation matrix is a pure rotation matrix, for the rotation angle α it writes

$$\mathbf{A} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

- From this we get

$$\mathbf{A}^{-1} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$$

$$\mathbf{1} - \mathbf{A}^{-1} = \begin{pmatrix} 1-\cos\alpha & -\sin\alpha \\ \sin\alpha & 1-\cos\alpha \end{pmatrix}$$

$$(\mathbf{1} - \mathbf{A}^{-1})^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2}\cot\alpha/2 \\ \frac{1}{2}\cot\alpha/2 & \frac{1}{2} \end{pmatrix}$$

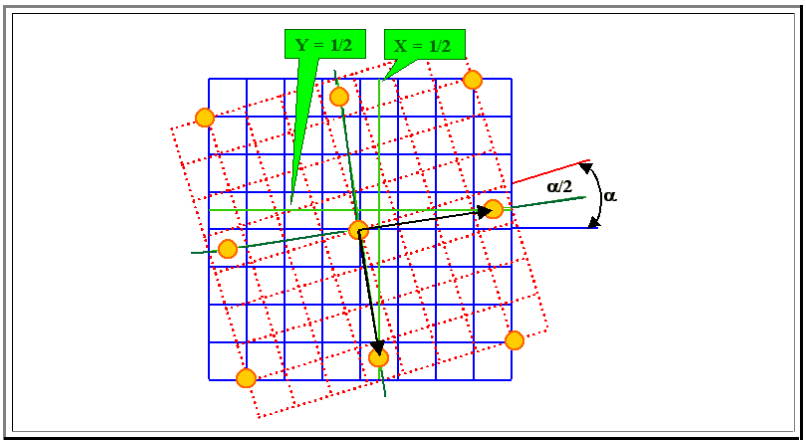
Now let's do an example. The base vectors of the square lattice **I** are $\underline{x}_1(I) = (1, 0)$, $\underline{x}_2(I) = (0, 1)$.

- If we use them as the *smallest* possible translation vector $\underline{T}(I)$ of lattice **I**, we obtain by multiplication with the last matrix the *smallest* vectors of the **O**-lattice which then must be the **unit vectors of the O-lattice**, $\underline{u}_1$ and $\underline{u}_2$:

$$\underline{u}_1 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \cdot \cot(\alpha/2) \end{pmatrix}$$

$$\underline{u}_2 = \begin{pmatrix} -\frac{1}{2} \cdot \cot(\alpha/2) \\ \frac{1}{2} \end{pmatrix}$$

- This is easily graphically represented, but the pictures get to be a bit complicated:



- Lattice **I** is the blue lattice, lattice **II** the red one; it has been rotated by the angle α.
  - The unit vectors of the **O**-lattice can be determined by the intersection of the light- and dark-green lines (remember the definition of **tan** and **cotan**!); they are depicted in black.
  - The **O**-lattice then can be constructed, its lattice points are shown as orange blobs.
- Note that a three dimensional expansion would not produce much that is new. On any plane above or below the drawing plane, the situation is exactly what we have drawn. This has one interesting consequence, however:
  - In three dimension, we have no longer **O**-*points*, but **O**-*lines* in this case.
  - The **O**-lattice in this case thus is not a point lattice, but a *lattice of lines* perpendicular to the plane of rotation. This will come up naturally later, but it is good to keep it in mind for what follows. However, since this is not the most general case, we will keep talking of **O**-*points*.
- The picture neatly helps to overcome a *possible misunderstanding*: For any **O**-point, a vector from the origin of either crystal to the **O**-point (our vectors **r(I)** and **r(II)**) point to *a coinciding equivalence point* or **O**-point, but *different* points of the **O**-lattice may be *different* equivalence points. In the example we have **O**-points that are almost at the center of **both** unit cells, or almost at a lattice point of **both** unit cells - the **O**-lattice seems to be constructed of two kinds of coinciding equivalence points; *but*:
  - If we would include more cells of the **O**-lattice, we would see that equivalence points shift slightly for the example given. A few **O**-lattice cells away, they would be more off-center or more distant to a lattice point than close to the origin of the **O**-lattice.
  - Just how many equivalence points of the set of equivalence points (which has an infinite number of members) are needed for an **O**-lattice is an important (nontrivial) question which we will take up later again.
- We can rephrase this *important question*:
  - Is the pattern of equivalence points *periodic* (= finite number of equivalence point) or *non-periodic* (infinite number)? In other words: If any one point of the **O**-lattice defines a specific equivalence point in the crystal lattices, does this specific point appear again at some other point in the **O**-lattice (apart from the trivial symmetries of the **O**-lattice)?
  - We will come back to this question later; it is the *decisive feature* of the **O**-lattice for defining the **DSC-lattice**.
- How do we get the **CSL** from the **O**-lattice? That is easy: It must be that particular subset of all possible **O**-lattices where all **O**-points are also lattice points in both lattices.
  - Looking at the unit vectors of the **O**-lattice, however, there is no way of expressing them in integer values of the base vectors of lattice **I**, because one component is always **1/2**. How about that?
- This is not a real problem, best illustrated with an example: If we chose **α = 36°52,2'**, we have

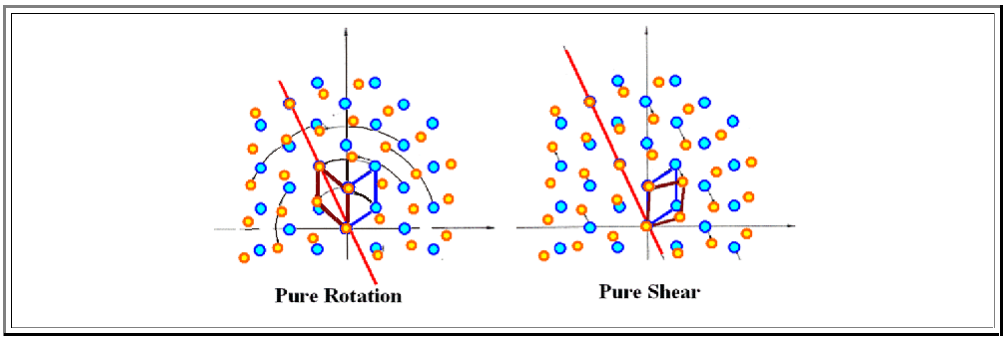
$$\underline{u}_1 = \begin{pmatrix} ½ \\ ½ \cdot \text{cotan}\,(\alpha/2) \end{pmatrix} = \begin{pmatrix} 1/2 \\ 3/2 \end{pmatrix}$$

$$\underline{u}_1 = \begin{pmatrix} -½ \cdot \text{cotan}\,(\alpha/2) \\ ½ \end{pmatrix} = \begin{pmatrix} -3/2 \\ 1/2 \end{pmatrix}$$

  - Thus every *second point* of the **O**-lattice is a lattice point in both lattices (depicting **O**- points of the equivalence class **[0,0]**), these points thus define the **Σ = 5 CSL**. The other **O**-points are of the equivalence class **[1/2,1/2]**. **CSL** lattices (two-dimensional case) thus correspond to specific **O**-lattices, but with lattice constant possibly larger by some integer value. This is quite important so we will illustrate this in a special module.
- Note, too, that in this case the pattern of equivalence point is obviously periodic, so we have a first specific answer to the question asked above.
- Before we delve deeper into the intricacies of **O**-lattice theory, we shall first discuss some of its general implications in the next paragraphs.

7.3.2 Working with the O-Lattice

- The math and physics of the **O**-lattice is not particularly easy because there are some tricky details to keep in mind. In this paragraph some of the problems, tricks and helpful definitions are just summarized; in due time they may be specified in more details.
- It is useful in many cases to decompose the transformation matrix **A** into matrices that describe the *volume deformation* (elongating or shortening only the axes of the crystal), the *shear deformation* (only changing the angles between the axes), and the *rotation* of the coordinate system of crystal **I** separately.
  - This may allow a better grasp of the real situation and helps, if necessary, to use approximations only for suitable parts of the system.
  - The main reason, however, lies in the fact that the *pure rotation is not unambiguously defined*. Depending on the basic symmetries of the system, the same final state of orientation can be obtained by many different rotations - but only one (or one set) may make sense physically. This leads to the next point:
- The choice between various possible transformations **A**. There are many possible ambiguities, not only with respect to the rotation part, but also, e.g., in the relations between pure shear and pure rotation; an example is shown below.
  - Starting from a given lattice **I**, identical lattices **II** can be produced either by pure shear or by pure rotation:



- *Mathematically*, there is no difference, but *physically* the two transformations are not the same because the atoms involved have to move in quite different ways. Which one is the physically sound one? As it turns out, the criterion is to *preserve nearest neighbor relationships*.
  - Mathematically, this means that from all possible transformation matrices **T**, the particular one that has to be chosen is the one *with the smallest numerical value of its determinant* $|T|$.
  - This ensures that the unit cell of the **O**-lattice generated will have the *largest possible value* (it is directly given by $1/|T|$), which will give the smallest possible dislocation content. This requires, of course, that you *know* all the possibilities for **A** in the first place - not a satisfying condition for a mathematician.
  - It may be noted in passing, that this *ambiguity* limits the usefulness of the **O**-lattice theory. There are cases, where the choice of the transformation matrix following the rules of **O**-lattice theory, does *not* lead to the "correct" solution as ascertained by looking at what the crystal does (by **TEM**).
- Another generalization comes from looking at the essentials of solutions to matrix equations. Consider the solutions of the basic equation

$$(I - A^{-1})\,\underline{r_0} = \underline{T}$$

  - From basic matrix algebra we know that the *type of solution* depends on the **rank** of the matrix ***A***.
  - We have the following cases:

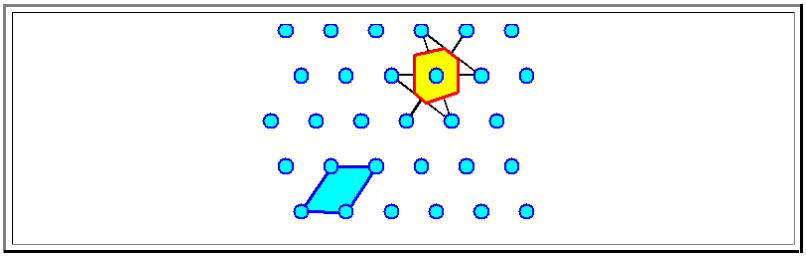
    - *Rank* **(A) = 3**
      The solutions define *points* in **O**-space, i.e. an ***O**-lattice*.
    - *Rank* **(A) = 2**
      the solutions are ***O**-lines*.
    - *Rank* **(A) = 1**
      The solutions are ***O**-planes*.
    - *Rank* **(A) = 0**,
      we have the trivial case of *identity* ***1***.

- *This is an issue of prime importance!*.
  - Since we can produce *all* grain boundaries (but *not* all phase boundaries) by just rotating crystal **II** around *one* properly chosen axis, the rank of the transformation matrix does not have to be larger than **2**.

- What does this mean? Well - for grain boundaries, there is no such thing as a **O**-point *lattice* - it is rather a *lattice of lines*. We have essentially a *two-dimensional* problem.
- Nevertheless, for the sake of generality, we will continue to discuss the "**O**-point lattice", knowing that it often is just a line lattice.

Solving the basic equation produces the **O**-lattice and therefore also the unit cell of the **O**-lattice. However, the "natural" unit cell obtained by simply connecting **O**-lattice points in some "obvious" manner may not be the physically most sensible one!

- As taught in basic crystallography, there are many ways of defining unit cells - we have another ambiguity!
- We will tend to take the **Wigner-Seitz cell**. Why? Who knows at this point - just go along. What this means is illustrated below:



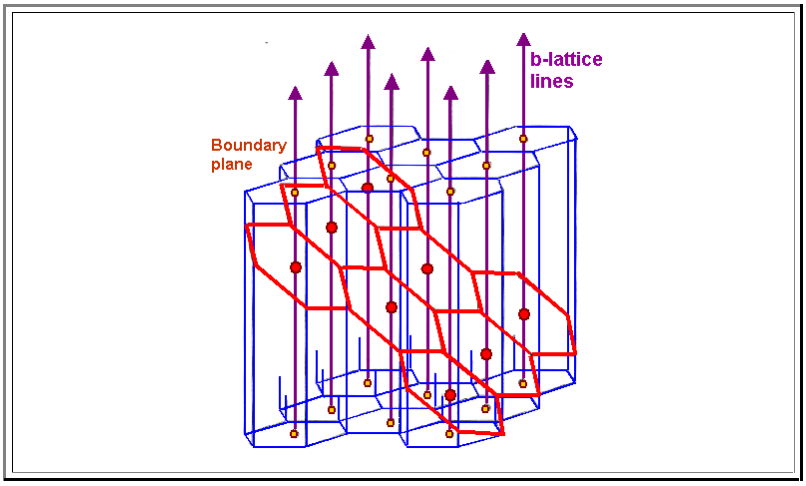
- Again, the right choice must come from the physical meaning of the **O**-lattice. This we will discuss in the next paragraph. Here we note that **O**-lattice defined in this way resembles nothing so much as a *honeycomb* - just remember again, that the **O**-points are lines. An illustration that comes fairly close in a slightly different context ("Bollmanns view of Franks formula") can be accessed via the link.

7.3.3 The Significance of the O-Lattice

- Lets pretend we are considering an actual grain boundary. We have found a suitable transformation matrix that produces crystal **II** out of crystal **I** with the right orientation, we have solved the basic equation, and we have constructed a suitable **O**-lattice. What does that give us?
- We now must address the essential question: What is the significance of the **O**-lattice for grain- and phase boundaries? What is the physical meaning? There is an *easy* answer and a *difficult* implementation:
  - First of all, *the **O**-lattice in itself has no physical meaning whatsoever* - in this it is exactly like the **CSL**.
  - However, since it *always* exists (unlike the **CSL**) *and* is defined in both crystals, if you were to design a boundary between *two* crystals of given orientation (and thus with *one* well-defined **O**-lattice) *that intersects as many **O**-lattice points as possible*, you will obtain the best physical fit along the boundary, i.e. *probably* the lowest grain boundary energies.
  - "Best physical fit" is not a very quantitative way of putting it. It means that the atoms to the left and right of the boundary will not have to be moved very much to the positions they will eventually occupy in the real boundary. This also can be expressed as "**minimal strain**" situation; the expression Bollmann uses.
  - If atoms happen to sit on an **O**-lattice point, they do not have to move at all because then then occupy equivalent positions in both crystals; if they are close to an **O**-lattice point, they only move very little, because at the **O**-points the fit is perfect.
  - The misfit increases moving away from an **O**-lattice point and reaches a maximum between **O**-lattice points.
- The crystals then can be expected to increase the area of best fit between **O**-lattice points and to *concentrate the misfit in the regions between **O**-lattice points* - this will be a dislocation with *Burgers vector = lattice vector*. We cannot, at this stage produce grain boundary dislocations, i.e. we are still limited to *small angle grain boundaries*.
  - There is a direct important consequence from this for the basic equation: We can replace ***T*(I)** by **b**(I), the set of possible Burgers vectors because they are *always* translation vectors of the lattice and obtain

$$(I - A^{-1})\,\underline{r}_0 = \underline{b}(I)$$

  - Remember that all translations vectors of the lattice are possible Burgers vectors; this came straight from the Volterra definition of dislocations. The fact that *observed* Burgers vectors are always the smallest possible translation vectors does not interfere with this statement - all it means is that a "Bollmann" dislocation with a large Burgers vector would immediately decompose into several dislocations with smaller vectors.
- Our basic equation yields the base vectors of the **O**-lattice if we feed it with the base vectors, i.e. the smallest possible translation vectors, of the crystal lattice. Since the Burgers vectors in a given lattice are pretty much the smallest possible translation vectors, too, we may see the **O**-lattice as some kind of *transformation* of the **b**-lattice, the lattice defined by taking the permissible Burgers vectors of a crystal as the base vectors of a lattice.
  - The crucial point now is to realize that the lines of intersection of the the actual plane of the boundary with the cell walls of the **O**-lattice (which, remember, looks like a honeycomb)), *are* the dislocations in the grain boundary. Whenever we cross over from one cell in the honeycomb structure to the next, we moved *one* Burgers vector apart in the real lattices. It is helpful at this point, to study the case of a small angle grain boundary treated in the advanced section under " Bollmanns view of Franks formula"; the essential picture is reproduced below.



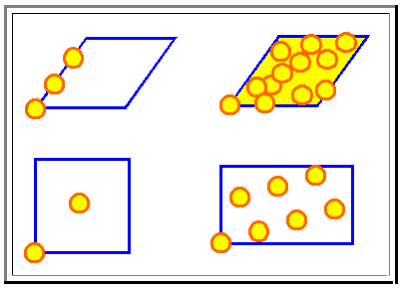
- The magenta lines are the **O**-lattice lines; the honeycomb structure is shown in blue, and the intersection with an arbitrary boundary plane produces the red dislocation network.
  - This is why it becomes important what kind of unit cell we pick for the **O**-lattice as mentioned before. As always, there are many possible choices.

- Bollmann gives precise directions for the choice of the "right" unit cell of the **O**-lattice - simply take the largest one possible (producing as few dislocations as possible). We will not reproduce the mathematical arguments; here we just note that it is possible to define an optimal **O**-lattice.

- We now have a big difference in the mental construction of a grain boundary between the **O**-lattice theory and the **CSL** theory. From the former *we now have a rule* for finding the optimal plane of a grain boundary for *any* given orientation - whereas the **CSL** model provides this information only for CSL orientations.
  - This rule will prove to be very general: We will be able to carry it over to the case of large angle grain boundaries (remember, that all complications notwithstanding, we implicitly deal only with small angle grain boundaries so far).
  - We also can obtain quantitative information about the dislocation structure in the chosen plane as long as we we restrict ourselves to small angle grain boundaries.
  - In this case the **O**-lattice theory is just a generalization of Franks formula - all you have to do is to replace "**sin**α" in the transformation matrix by "α" (and use the corresponding linearizations of all other trigonometrical functions for small angles) - Franks formula will result.
- In other words, as long as the spacing of the **O**-lattice is large compared to the crystal lattices, all of this makes sense, and this condition is always met for small deformations, i.e. for small angle boundaries.
  - For **O**-lattices with lattice constants in the same order of magnitude as the crystals, however, the spacing between the dislocation would be too small as to be physically meaningful - exactly as before. So what is new?
  - Well, the **O**-lattice theory as a generalized version of Franks formula, is not just applicable to small angle grain boundaries, but to "small deformation" boundaries of *any* kind, including phase boundaries. This is already a remarkable achievement.
  - But, as we will see, the complete **O**-lattice theory also incorporates arbitrary ("large angle") boundaries of all types, too.
- In order to progress, we now must ask the question: Are there any "special" **O**-lattices, or, in other words, special orientations the crystals would prefer?
  - We already know parts of the answer: Yes, there are preferred orientations for grain boundaries; the **CSL** orientations, which, after all, must also be expressible in the **O**-lattice concept.
  - From this we can go on and this will be dealt with in the next chapter.

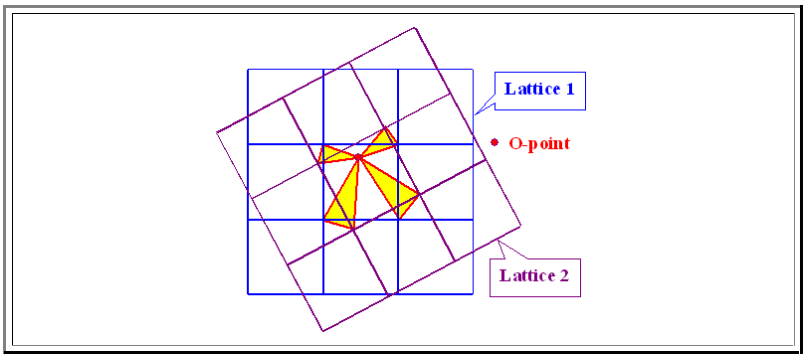
7.3.4 Periodic O-Lattices and Pattern Elements

- A CSL lattice by definition has coincidence points in both lattices; the **CSL** points thus are always **O**-lattice points, too. The converse is not necessarily true (as we already have seen in the example):
  - The **O**-lattice of two crystals in a **CSL** orientation thus must include the **CSL** lattice points as **O**-lattice points. This **O**-lattice, however, may also have additional **O**-lattice points - all we can deduce at this point is that the **CSL** lattice points must be a *subset* of the **O**-lattice points which belong to the **O**-lattice that includes the specific **CSL** orientation.
  - We know that the **CSL** points are **O**-points which are always of the same equivalence type - they are lattice points, to be precise. In other words, the **O**-lattice belonging to a certain **CSL** lattice, if drawn into the coordinate system of one of the crystals *is periodic in this reference system*.
  - This is *not* a general property of an **O**-lattice - in general, every equivalence point defined by an **O**-point could be different from all the others and there would be no periodicity.
- This is best visualized by drawing all equivalence points encountered for a given **O**-lattice (which, of course, always has infinitely many points) into the unit cell of one of the crystals - we obtain the so-called **reduced O-lattice**.
  - For a *periodic* reduced **O**-lattice, there would be a *finite* number of equivalence points; a *non-periodic* lattice would lead to an *infinite* number of equivalence points in the reduced **O**-lattice.
  - Lets look at some examples:



  - Shown are elementary cells of one lattice (blue) with the equivalence points occurring in the **O**-lattice drawn in. In three cases the **O**-lattice would be periodic; in the case in the upper right, it would be non-periodic

- Periodic **O**-lattices are clearly special; and it is self-evident that every **CSL** orientation must correspond to a periodic **O**-lattice. *But there is more*.
  - At any **O**-lattice point in a periodic **O**-lattice, we have a certain arrangement of the crystal atoms around that point, a specific *pattern*. Since in a periodic **O**-lattice there are only a finite number of different equivalence points, there is only a finite number of distinct patterns, too.
  - An individual pattern is called a **pattern element**. There are as many pattern elements as there are equivalent points in the reduced **O**-lattice.
  - This is a crucial concept in **O**-lattice theory, unfortunately it is not explained very well in Bollmanns book. Let's see what is meant by pattern:

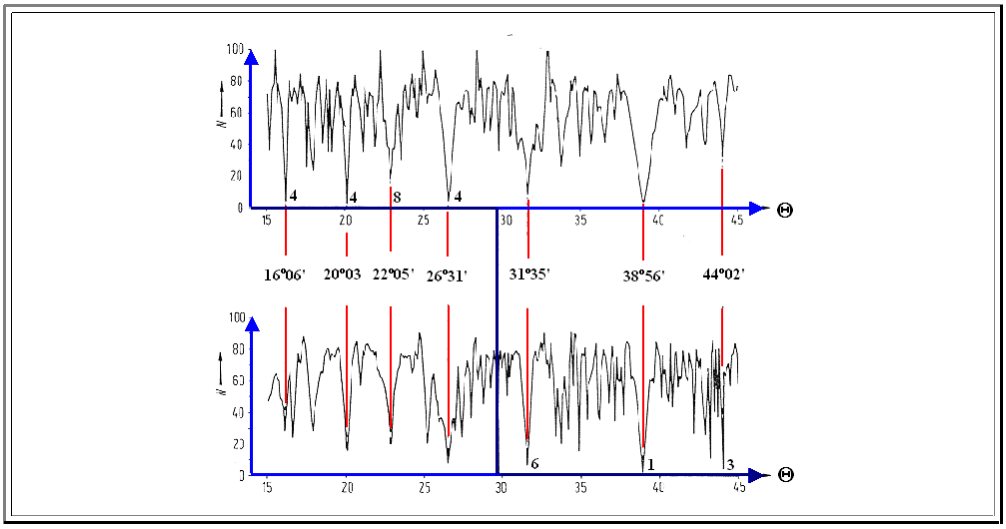


  - Shown are two lattices (blue and magenta which are superimposed) and one **O**-point (red). A *representation of the geometry* of the atoms that you may put into the lattices is given by the yellow triangles. They are simply constructed by connecting the lattice points of the two lattices "around" the **O**-point with the **O**-point.
  - The picture also demonstrates (but does not prove) an universal theorem: *Any* **O**-point can be chosen as the *origin* for the transformation that produces lattice **2** from lattice **1** (here it is a simple rotation).
- In a non-periodic **O**-lattice, the representation of patterns in the way shown is different at any **O**-point - this is also rather difficult to draw.
  - This is where **O**-lattice theory gets hard to illustrate. Nobody surpasses Bollmann who provides complicated drawings of patterns (done by hand!) in his book, one example is shown in the link.
- The question now is: Which orientations provide *periodic* **O**-lattices? It appears that there is no simple formula coming up with transformation matrices or angles for rotations that produce *periodic* **O**-lattices. We have to go the other way and ask two questions for *any* possible orientation:

- Is the corresponding **O**-lattice periodic?
- If yes, how many pattern elements (**= N**) are contained in the reduced **O**-lattice?

What we want is **N** as a function of some misorientation angle for some simple geometries. This needs some numerical calculations; lets look at the results for rotations on the **{110}** planes of cubic crystals

- The following picture shows **N** as a function of the misorientation angle:



- This again is one of Bollmanns trickier pictures (with some color added), because it is *only* understandable if you read and understood much of what has been said before in his book (it is neither explained what the difference is between the two curves - they have after all an identical coordinate system, nor what the bold lines (here dark blue) implicate).
- Well, the **N**-values are given for *two independent kinds of transformations* which both include the same rotation T (upper and lower curve), but one inlcudes a so-called "unimodular transformation" in addition. The one with the smallest determinant which, as we have seen, is the one you should use, changes from the upper curve to the lower curve at **T = 30°** and this explains the bold (or dark blue) lines. Since the two curves are different, you now see that it matters, indeed, which transformation matrix you pick.
- Don't worry; it is not necessary to understand that in detail. Just acknowledge, that **N** can be computed and that unambiguities with respect to different transformation matrices can be dealt with somehow.
- Also note that the "real" curve would be a fractal with **N =** ∞ for most values; it is smoothed here by only counting the equivalence points in **100** "pixels" of the **O**-lattice (so **N** ≥ **100** applies) and stopping the numerical procedure after some time if it does not turn periodic anyway.

It is clear that there are several "special" orientations for this geometry with small values of **N**. This looks good, however, we are not yet done. We are really looking for **O**-lattices that are periodic on a short scale, i.e. the patterns should repeat after a short distance. This requires three ingredients:

- *Periodicity* as a starter, i.e. **N** is finite (or in reality e.g. **N < 100** for numerical calculation).
- *Small values* of **N**, because the pattern repeats after **N** steps - the larger **N**, the longer it takes for a repetition. To give an example: For **N = 10** you have to go out **10** lattice constants of the **O**-lattice before the same pattern is encountered again.
- This immediately calls for *small lattice constants of the **O**-lattice* , too. Or, to be more general, for *small volumes V_O of the **O**-lattice cells*.

The real measure for the periodicity of the **O**-lattice patterns is therefore not **N**, but the *density* **N'** of periodic equivalence points given by

$$N' = \frac{N}{V_O} = \frac{N}{|T|}$$

- With **|T|** meaning the determinant of the transformation matrix, since $V_O = 1/|T|$ follows from basic matrix algebra together with the definition of the **O**-lattice.

- Now comes a major point: **N'** is nothing but the number of crystal units (volume of unit cells or lattice constants) per period of the pattern because the unit of the **O**-lattice is always (for periodic **O**-lattices) an integer number of the crystal units!
  - In other words: **N'** corresponds directly to the measure of coincidence in the **CSL** model, the number Σ ! In fact, the numerical values are identical in most (but not all) cases: **N' = Σ**.
  - The **O**-lattice theory, however, is not only much more general, but gives the recipes for calculating **N'** (or Σ). Try, for example, to find the **CSL** lattices for orientations between, say a cubic and a monoclinic lattice: All you need are the deformation matrices; the rest can be done for all possible cases by a computer program.
  - Just one case in point: What happens for perfectly well defined transformation matrices **T**, but with **|T| = 0**? **N'** in this case will be ∞.
  - Lets look at an example: Rotation of cubic lattices by an angle around a **<110>** direction:

Angle Θ	|T|	N	N'	Σ
10° 6,0'	0,031	4	129	129
13° 26,06	0,055	4	73	73
20° 3,0'	0,121	4	33	33
22° 50,4'	0,157	8	51	51
26° 31,6'	0,210	4	19	19
38° 56,6'	0,111	1	9	9
50° 28,6'	0,000		oo	11
58° 59,6'	-0,030	1	33	33
70° 31,6'	0,000		oo	3

  - Two perfectly well defined rotations lead to **|T| = 0**; their Σ values are **11** and **3**, respectively, while **N'** is infinity!
  - This tells us that these particular orientations are *much more special* then implied by their Σ values: These orientations can be obtained by simpler transformations matrices of lower rank and they correspond to grain boundaries with a particular high degree of "fitting" and thus low energy.
  - There is also a first real result: **Σ11** boundaries should be rather common, and that's what they really are.
- We will not go into more details at this point; but it should become clear that there is a lot of power behind the **O**-lattice theory.
  - However, even at this stage, calculations become tedious and need numerical methods. It would be most useful to implement the basic equations in a computer program from now on - but I do not know if this has been done.
  - And, always keep track of this: So far we have only dealt with "small deformation" boundaries and with high angle boundaries having a *periodic* **O**-lattice. We are still some distance away from a general boundary.
- We now need to do the next step - always, for easier understanding, in analogy to the **CSL** model of grain boundary structures:
  - What happens if the orientation of the two crystals (including arbitrary lattices and thus phase boundaries, too) is close to, but not exactly at a "special" **O**-lattice orientation? "Special" meaning a periodic **O**-lattice.
  - In other words, we are asking for possible structural defects which can be superimposed and will change the (generally non-periodic) **O**-lattice of an arbitrary boundary (which is always defined) just the right amount to generate a periodic **O**-lattice with a supposedly low energy?
  - This is the essentially the same question we asked for crystals close to, but not exactly at a " low Σ" orientation - but on a much higher level of abstraction and with the possibility to deal with it quantitatively.

7.3.5 Pattern Shift and DSC Lattice

The General Idea of Pattern Element Conservation

- In the **CSL** model of grain boundaries the **DSC** lattice was introduced to account for small deviations from a perfect lattice coincidence orientation. It was the lattice of *all translations of one of the crystals that conserved the given **CSL***. Translations other then those of the **DSC** lattice would destroy the coincidence of lattice points.
  - The lattice vectors of the **DSC** lattice therefore could also be interpreted as the set of possible Burgers vectors for dislocations allowed in a grain boundary without destroying the coincidence.
  - While the simple recipe for constructing the **DSC** lattice in the simple cases usually shown (two-dimensional, cubic lattices) is rather straight forward, it was neither mathematically justified, nor is it immediately clear how it should be constructed in complicated cases.
  - The **DSC** lattice, in fact, comes from the **O**-lattice theory and was simply adopted to the "easy" **CSL** model.
- Obviously we now must ask ourselves: What happens to an ***O**-lattice*, particularly a periodic one, if we translate one of the crystals?
  - This is actually one of the more complicated questions to ask, especially for the rank of the transformation matrix **A < 3** (as we expect for grain boundaries).
  - We will not go into details here because in this rendering of **O**-lattice theory we omitted some more mathematical points considering what happens to the **O**-lattice in a given situation if you shift (= translate) crystal **I** or crystal **II**. Or, in a reversed situation, how you must shift crystal **I** or **II** if you translate the given **O**-lattice.
- The first answer to the question above is:
  - In general (i.e. rank **A = 3**), the **O**-lattice is preserved, but shifted by some amount that depends on the (arbitrary) magnitude of the translation of the crystal chosen. *This is in contrast to the **CSL***, where arbitrary shifts *not* contained in the **DSC** lattice will completely destroy the **CSL**.
  - This does not help, we obviously must find a more specific criterion than just conservation of the **O**-lattice in general in order to find specific translations that correspond to Burgers vectors of grain boundary dislocations. We therefore ask more specifically:
- What happens to the *pattern elements* associated with every equivalence point in a reduced *periodic **O**-lattice* upon shifting one of the crystals?
  - Since we have seen (without proving it) that any **O**-point can be taken as the origin for the rotation transforming crystal **I** into crystal **II**; we should be able to shift lattice **I** by any vector pointing to an equivalence point in the reduced **O**-lattice without changing pattern elements. In other words we simply change the origin of the rotation (we only look at rotations in this examples).
  - The **O**-lattice then will also be shifted by some other vector which can be calculated by employing our basic equation

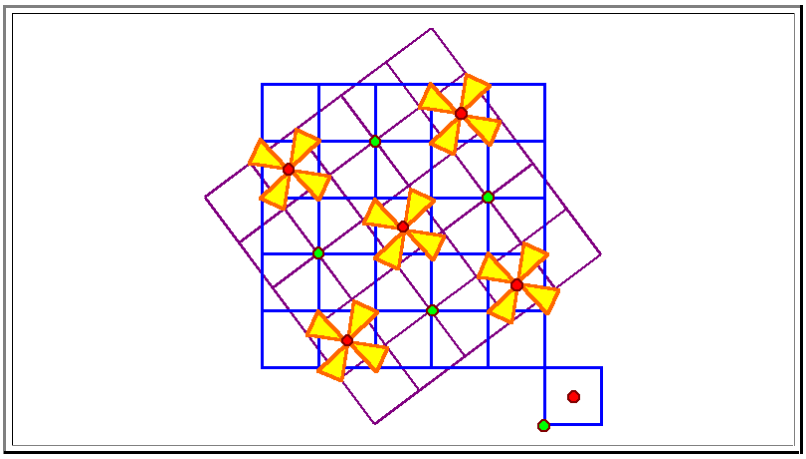
$$(\mathbf{I} - \mathbf{A}^{-1})\,\underline{\mathbf{r}}_{i0} = \underline{\mathbf{T}}(\mathbf{I})$$

  - The $\underline{\mathbf{r}}_i$ are the base vectors of the **O**-lattice if we take $\underline{\mathbf{T}}_i$ to be the set of base vectors of the crystal **I** lattice.
- Now shift the crystal **I** lattice by some vector $\underline{\mathbf{e}}$ connecting equivalence points, replace $\underline{\mathbf{r}}_i$ by $\underline{\mathbf{r}}_i = \underline{\mathbf{r}}^0_i + \Delta\underline{\mathbf{r}}_i$, with $\Delta\underline{\mathbf{r}}_i$ = shift of the **O**-lattice for a shift $\underline{\mathbf{e}}$ of the crystal lattice, and solve the equation for the $\Delta\underline{\mathbf{r}}_i$.
  - Well, lets *not* do it, but accept that there is a shift that can be calculated.
  - On *second* thoughts, this must also be true for lattice **II**. We thus may also employ vectors that translate lattice **II** by one of the vectors pointing to equivalence points in the reduced **O**-lattice.
  - And on *third* thoughts (not entirely obvious), we also must be able to translate the **O**-lattice itself by any vector that connects equivalence points. This requires that the **O**-lattice shifts by some vector - it is the reverse problem from the one outlined above.
- The trick is that all those shifts may be different, and while they all produce the same general **O**-lattice, there might be different pattern elements. But - there is a *finite* number of pattern elements and a *finite* number of possible shifts.
  - Obviously, the set of all different configurations (distinguished by pattern elements) obtainable defines the complete geometry of the particular boundary with the periodic **O**-lattice considered because no configuration is special.
  - The set of all possible displacement vectors can be expressed as the translation vectors in a new kind of lattice, the "**Complete Pattern Shift Lattice**", abbreviated by Bollmann as " **DSC lattice**", that we encountered earlier (in a much simpler form).
- Unfortunately, it is not immediately obvious how to calculate the **DSC** lattice from **O**-lattice theory. In fact, the respective chapter in Bollmanns book is particularly ***hermetic*** or obtuse.

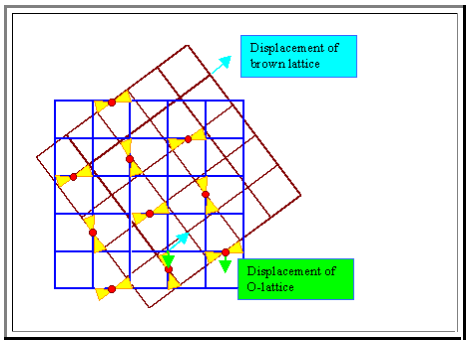
- Somewhat later (**1979**), Bollmann together with **Pond** gave the old abbreviation a new meaning: "**DSC**" now stands for "***D**isplacements* which are ***S**ymmetry **C**onserving*". But few people know what exactly **DSC** stands for - the main thing is to understand the significance of the **DSC** lattice.

Some Illustrations

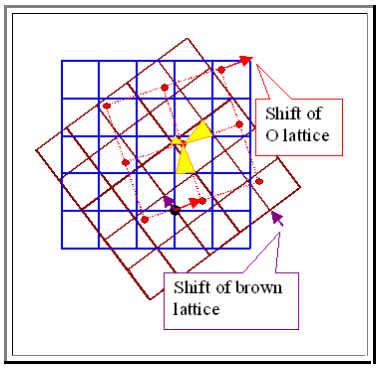
- Lets see what the various displacements discussed above really produce if applied to a simple situation. We take the (redrawn) example from Bollmanns book.
  - First, lets construct the possible set of pattern conserving translations by putting several reduced **O**-lattice cells together (for the case of rotation around **<100>** of **39^{o} 52,2'**, corresponding to the **Σ = 5 CSL**).



  - The left part shows the rotation, yielding the **O**-lattice. Coinciding *lattice* points that are also **O**-points are shown in green, the other **O**-points in red. On the right-hand side the repeated reduced **O**-lattice is shown in the blue crystal.
- Now lets displace the brown lattice by a vector pointing form the green to the red equivalence point in the above picture. Here is what you get.



  - The **O**-lattice shifted down, and some new kinds of pattern elements appear. There are no more or less special than the ones in the picture above; both belong to the complete structure of the boundary illustrated.
  - Note that we also obtain new equivalence points for the boundary (in the middle of the lines defining the square lattices).
  - Now we shift the brown lattice by one of the vectors pointing to the new equivalence point. We obtain yet another pattern element.



  - But that's it. The pattern elements shown here are all there are (Try to prove that yourself if you don't believe it).
  - We could now start to produce the **DSC** lattice, but this will just give the same kind of lattice we had in the simple **CSL** case.
- Instead we only note that there is a sufficiently clear procedure of how to create a **DSC** lattice for a given periodic **O**-lattice, that is *always* applicable - even to phase boundaries (in principle; of course only in principle).
  - In the next (and last) chapter, we will show how **O**-lattice theory now can be applied to large angle grain boundaries and discuss briefly its merits and limits.

7.3.6 Large Angle Grain Boundaries and Final Points

Application of O-Lattice Theory to Large Angle Grain Boundaries

- The basic assumption is that an arbitrary grain boundary would prefer to be in a orientation corresponding to a *periodic* **O**-lattice with a high density of equivalence points. This is a fancy (i.e. more general) way for saying that boundaries prefer to be in a low Σ orientation as we did in the simple **CSL** model.
  - Such a particular orientation can be achieved at the expense of generating some *grain boundary dislocations*.
  - For a small angle grain boundary we have seen that a cut of the boundary plane through the **O**-lattice gives directly the geometry of the dislocation network (give or take some adjustments to account for the particular dislocation properties).
  - The **O**-lattice was obtained (by calculations) relative to the two real crystals. Looking back, what we did was to use one of the crystals as a *reference for the preferred state*, the other one then described the *deviation* of the boundary from the preferred state.
- For large angle grain boundaries we now do exactly the same thing - *except* that the reference state is now the *periodic **O**-lattice that the boundary aspires to obtain*.
  - The *deviation* of the boundary from this preferred state is then described by the **O**-lattice that comes with the orientation that describes the *actual* boundary .
- The logical consequence then is that the geometry of the dislocation network necessary to obtain the preferred state is the *__O__-lattice of the two __O__-lattices* described above - a so-called **O2-lattice** or **second order O-lattice**.
  - That may sound a bit heavy, but it is really straight forward if you think about it.
  - It is also clear - in principle - how we would calculate the **O2**-lattice, but we are not going to look at this.
- If we now imagine a boundary plane cutting through our **O2**-lattice as before, we now must ask how large the translation will be that the **O**-lattice "crystals" experience when a **O2**-lattice wall is crossed. This translation will be the Burgers vector of the second-order dislocation forming the grain boundary dislocation network.
  - Well, as you would have guessed, it must be a translation that conserves the underlying pattern of the **O**-lattices, so the translation vector (= Burgers vector) is a vector from the **DSC**-lattice of one of the primary **O**-lattices.
  - While our periodic reference **O**-lattice has a defined **DSC** lattice, the other one may (and in full generality probably will) have a non-periodic **O**-lattice and thus does not have a **DSC** lattice. *This looks like a problem*.
  - However, since **O**-lattices are continuous (and smooth) functions of the misorientation angle (which the **CSL** is not), we know that the two **O**-lattices are rather similar, and we always can take the **DSC** lattice of the periodic **O**-lattice in a good approximation. So there is no real problem.
- OK. We are done. That's (almost) all there is to it.
  - The general recipe for *constructing* a grain boundary with a secondary is network is "clear". It goes exactly along the lines we derived for small angle grain boundaires - only we work in "second order **O**-lattice theory".
  - The reverse is also possible: We have a general recipe for *analyzing* the structure found in a real boundary.
  - It will just take a few months of studying the intricacies of the underlying math and some getting used to the more trickier thoughts, and you can construct and analyze all kinds of boundaries on your own.
- But most likely, you won't. This is due to some (sad) facts of life that will be the last thing to discuss in this context.

Merits and Limitations of O-Lattice Theory

- The merits of Bollmanns theory are clear:
  - It nucleated a lot of work on grain boundary structures and introduced the crucial concept of the **DSC** lattice and its dislocations.
  - It was (and is) the mathematical frame work for tackling the kind of higher-level geometry that is contained in interfaces between crystals. (It will be interesting to see if someone sometime tackles the grain boundary structure between single quasi-crystals, which could be done by extending **O**-lattice theory into a **6**-dimensional space and then project the results back into three dimensional space).
  - It allows to conceive and analyze more complex problems, where a **CSL** model is not sufficient.
- However, there are serious problems and limitations, too.
  - The recipe for the proper choice of the *one* transformation matrix you should use out of many possible ones *is not always correct*. It generally fails for (some) small angle tilt boundaries, where the **O**-lattice theory would predict a twin-like structure with no dislocations - contrary to the observations. It also fails for some other boundaries, casting some lingering doubt on the whole thing.

- It is still *too simple* to account for real boundary structures even if the limitation referred to above does not apply. Two examples might be mentioned.
- **1.** The *rigid body translations* observed in many (twin-like) boundaries, especially in **bcc** lattices.
- **2**. Tremendously complicated structures observed in crystal with more than one atom in the base of the crystal - e.g. in **Si**. What happens (and was first observed and then analyzed by Bollmann) is that dislocations in the **DSC** lattice may split into partial dislocations bounding a stacking fault in the **DSC** lattice. While this effect may be incorporated into the **O**-lattice theory, it does not make it easier.

Still, whereas newer theories concerned with the structures of grain boundaries do exist, none is quite as complete and mathematical as the **O**-lattice theory. A "final" theory has not yet been proposed

What remains for practical work is

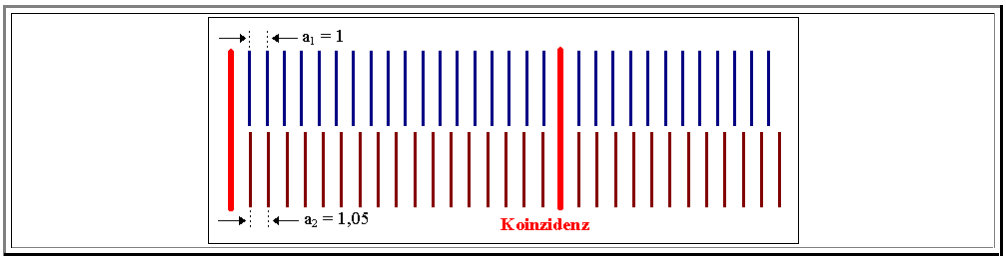
- *The **DSC** lattice*. This is certainly the most important outgrowth of the **O**-lattice theory. Grain boundaries simply cannot be discussed without reference to the **DSC** lattice. For practical importance it has all but eclipsed the **O**-lattice. As we have seen, it is (mostly) easily constructed without going into heavy matrix algebra.
- The systematic approach, always good for looking deeper into less clear situations.
- The good feeling, that something can be done about taking a deep look into grain boundary structures from a theoretical point of view, even if there are some limitations and unclear points.

8. Phase Boundaries

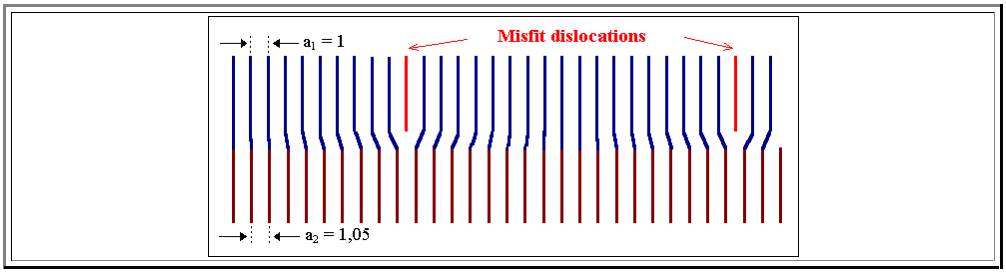
8.1 Misfit Dislocations

8.1.1 Modifications of the CSL Concept and Misfit Dislocations

- The O-lattice theory fully accounts for the *structure* of phase boundaries, too - as long as we look at two crystalline phases, of course. It is, however still not easy to use (it is waiting for someone to turn it into a user-friendly piece of software), and it can not answer a few specific question about the *development* of the structure whenever a phase boundary is formed.
- So for phase boundaries, too, it is often more easy to think in terms of the simpler coincidence lattice - *but with a grain of salt*. A few special points are:
  - In general, there will be *no suitable coincidence lattice at all*, because the lattices are different and their lattice constants are incommensurable (their quotient is an irrational number). In practice, however, we do not know the lattice constants to an arbitrary degree of precision, and you will always find some fitting relation.
  - Even if there is a **CSL**, it is *not necessarily the proper reference lattice*. This can be seen from a simple example: Two cubic crystals with lattice constants $a_1 = 1$ and $a_2 = 1{,}05$ (i.e. a misfit of **5%**) from a phase boundary:



- We have a perfect two-dimensional **CSL** structure ($\Sigma_{2d} = 20$ would hold for **2** dimensions),
  - Note that we can have situations (like *even* Σ numbers) which are simply not possible for grain boundaries, where the lattice constants are the same by definition.
- It is much more sensible to describe this phase boundary as a $\Sigma = 1$ boundary with superimposed **phase boundary dislocations** (which we practically always will call **misfit dislocations**) as shown below, because this is usually an energetically better situation than a $\Sigma_{2d} = 20$ (or whatever) boundary with no dislocations.
  - The misfit dislocations in this case are more or less lattice dislocations of the crystals - but that does not mean that **DSC** lattice dislocations never occur in phase boundaries!



- Misfit dislocations compensate for differences in the lattice constants by concentrating the misfit in one-dimensional regions - the dislocation lines.
  - Between the dislocation lines the interface is **coherent**; a phase boundary with misfit dislocations is called **semi-coherent**.
  - Misfit dislocations - in contrast to general grain boundary dislocations - *must have an edge component* that accounts for the lattice constant mismatch
- Whereas the **O**-lattice theory as applied to phase boundaries allows phase boundary dislocations in general (of which misfit dislocations are only a subset), "simple" misfit dislocations are the dominant defects in *technologically important* man-made phase boundaries.
  - Misfit dislocations are not restricted to boundaries between two chemically different types of materials. Silicon heavily doped with, e.g., Boron, has a slightly changed lattice constant and thus formally can be sen as a different phase. The rather ill defined interface between a heavily doped region and an undoped region thus may and does have misfit dislocations, an example is given in the illustration.
  - The mere existence of misfit dislocations coupled with their usually detrimental influence on electronic properties is the reason why many "obvious" devices do not exist at all (e.g. optoelectronic **GaAs** structures integrated on a **Si** chip), and others have problems. The aging of Laser diodes, e.g., may be coupled to the behavior of misfit dislocations in the many phase boundaries of the device.

- Optoelectronics in general practically always involves having phase boundaries, e.g. devices like Lasers, **LEDs**, as well as all multi quantum well structures. A very careful consideration of misfit and misfit dislocations is always needed and some special process steps are often necessary to avoid these defects.

- However, not every (**Σ = 1**) phase boundary with some misfit between the partners contains misfit dislocations - provided one of the phases consists of a *thin layer* on top of the other phase. Only if the thickness of the thin-layer phase exceeds a *critical value*, misfit dislocations will be observed. It is easy to understand why this is so:
  - For thin layers, it may be energetically more favorable to deform the layer *elastically*, so that a perfect match to the substrate layer is achieved. The total elastic energy contained in the "**strained layer**" scales with the thickness of the layer and the expenditure in elastic energy below a **critical thickness** for an epitaxial layer may be smaller than the energy needed to introduce misfit dislocations.
- This is a situation not dealt with in the **O**-lattice theory or its simple **CSL** version. A *new theory* is needed.

8.1.2 Energy of Misfit Dislocations and Critical Thickness

- The critical thickness for the introduction of misfit dislocations can be obtained by equating the energy contained in a misfit dislocation network with the elastic energy contained in a strained layer of thickness ***h***.
  - Since the elastic energy increases directly with ***h***, whereas the energy contained in the dislocation network increases only very weakly with ***h***, the thickness for which both energies are equal is the critical thickness h_c. Thicker layers are energetically better off with a dislocation network, thinner layers prefer elastic distortion.
  - This computation was first done by **Frank** and **van der Merve** in **1963**; the resulting *Frank and van der Merve formula* became quite famous.
- Somewhat later in **1974** **Matthews** and **Blakeslee** reconsidered the situation and looked at the forces needed to move a few pre-existing dislocations into the interface in order to form the misfit dislocation network. They obtain the same formula for the critical thickness as van der Merve (i.e. the equilibrium situation), but their treatment also allows to consider the *kinetics* of the process to some extent (i.e. how the network is formed) and is therefore widely used.
  - We are looking at the situation *retrospectively* by studying an article of the possibly most famous **TEM** and defect expert, **Peter Hirsch** from Oxford University, or, to be precise, **Sir Peter** as he must be called after his nobilitation by **Elizabeth I**, Queen of England.
  - This is to show that - honorwise - a defect expert can go just as far as a rock star (several of which have been knighted by the queen - most famous the **Beatles**). Moneywise, however, it is a completely different matter.
  - We use parts of **his article** printed in the *Proceedings of the 2nd International Conference on Polycrystalline Semiconductors (Schwäbisch Hall, Germany, 1990, p. 470)*. *Do look it up* - it is part of the lecture!

- As you saw, great minds sometimes make great steps and are not immune to small errors! If you didn't see that, consider:
  - How exactly do you get **eq. 1**?
  - Why is the strain for minimum energy calculated in **eq. 3** equal to the unrelaxed elastic strain at the point of the introduction of dislocations?
  - What is ***h***, the thickness of the layer, doing in an equation for the *critical thickness* h_c (**eq. 5**)? After all, the critical thickness can not possibly depend on the thickness itself.
- Well, if you want to know, turn to the annotated version of Sir Peters paper.

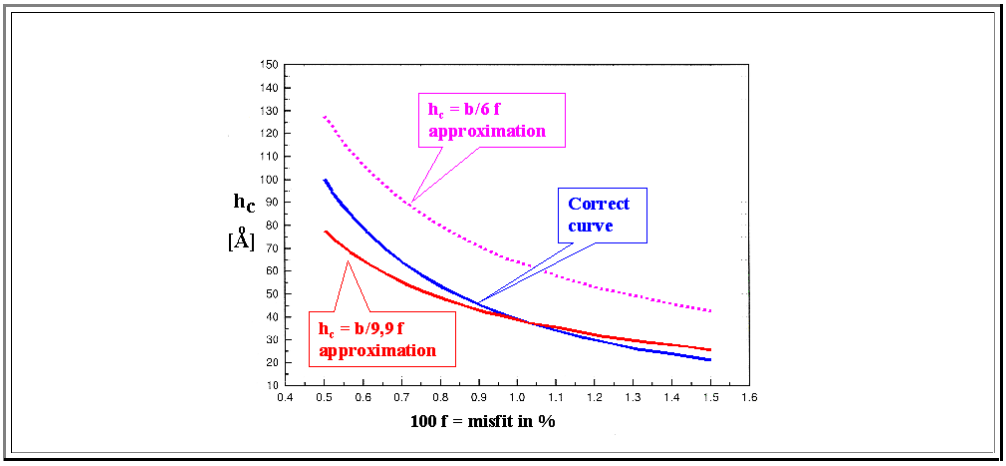
- Still, Sir Peter got it right in principle, and his derivation of the critical thickness is short and most elegant. The final formula for the critical thickness h_c is

$$h_c = \frac{b}{8\pi \cdot f \cdot (1 + \nu)} \cdot \ln \frac{e \cdot h_c}{r_0}$$

  - With ***b*** = Burgers vector of the misfit dislocations (actually only their edge component in the plane of the interface), ***f*** = misfit parameter, i.e $\Delta a/a$, **e = e = 2,7183... =** base of natural logarithms, and r_0 **=** core radius of the dislocations.
  - This transcendental equation may be roughly approximated by

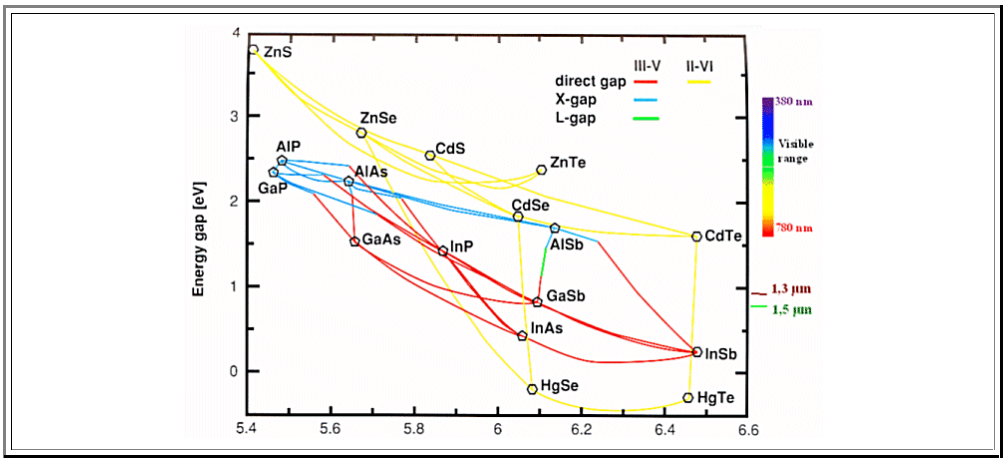
$$h_c \approx \frac{b}{9.9 \cdot f}$$

- Lets see what the calculations tell us for real phase boundaries (for a ***b*** value of **0.376 nm** (which applies to **Si**)). We note that misfit dislocations are only to be expected if the layer thickness ***h*** exceeds the critical value h_c.



  - For a misfit of **1%** the critical thickness is about **4 nm** - *not much at all*!

- This situation provides for many technological problems, especially in semiconductor technology. It imposes severe limits on "heterojunctions", i.e. electronic junctions between two materials because a misfit dislocation network will invariably "kill" your device - if not immediately, somewhat later (which is often worse!).
  - Looking at common technical semiconductors, we realize that we have major problems in making heterojunctions:



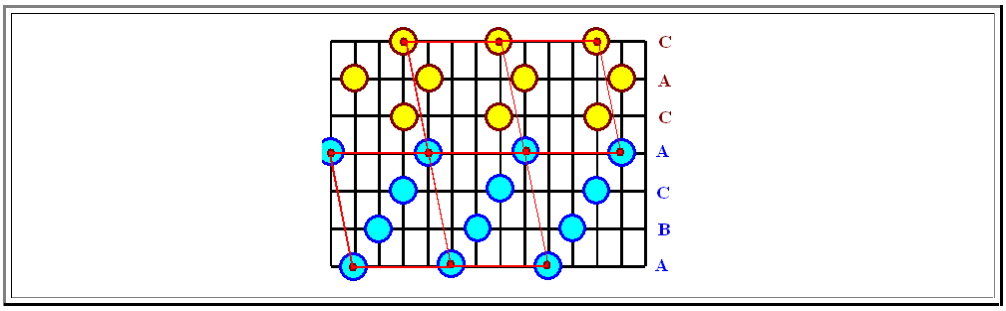
  - Misfits between two materials tends to be large (not even considering **Si** with a lattice constant of **a = 0,532 nm**), and dislocation free interfaces do not come easy, if at all.
  - A large group of researchers has been (and still is) looking for ways to beat the critical thickness limitations. There are many tricks (the link contains a few), but hard work is needed just as much as some luck and good ideas. A particularly clever recent idea known under the heading of "**compliant substrates**" is described in an advanced module.

- Experiments confirm the theory. Very thin epitaxial layers of a second phase do not show dislocations in the interface, but with increasing thickness misfit dislocations will appear.
  - Considering that misfit dislocations are usually unwanted but that they must appear with increasing layer thickness - however not out of thin air - we ask an important question:

- Exactly how are misfit dislocations produced and incorporated into the interface if the critical thickness is reached. More to the point: How can I prevent this *nucleation* and *migration* process?
  - Suffice it to say that while this question has not been fully answered, there are many ways and tricks to keep misfit dislocations from appearing at the earliest possible moment.
  - The issue is sufficiently important in optoelectronics to merit more discussion. More information to this point can be found in various modules of the Hyperscript "**Semiconductors**".

8.1.3 Other Defects in Phase Boundaries

- Even coherent phase boundaries can still contain other defects besides misfit dislocations (not to mention incoherent phase boundaries). In particular, we must expect:
  - *Dislocation networks* besides the misfit dislocations that compensate for small *tilt and twist components* in analogy to the small angle grain boundaries.
  - *Steps* associated with dislocations (so-called *incoherent steps*) in analogy to the steps encountered in grain boundaries.
  - *Coherent steps* (without any dislocation character) as something new.
- After all, the surface of a substrate on which we deposit a layer of a second phase will, in general, not be atomically flat. **Steps** thus must be expected to be an integral part of the phase boundary. We will examine some examples for this in the next subchapter.
- Next, it is important to realize that semicoherent phase boundaries can have other **CSL** relation besides **Σ = 1**, in particular **Σ = 3**, but other values, too.
  - As the most important example, consider a *hexagonal lattice* matched to the **{111}** plane of a *fcc lattice*. It is found to be in a **Σ = 3** relationship, easily seen if you compare the stacking sequences in the picture below:



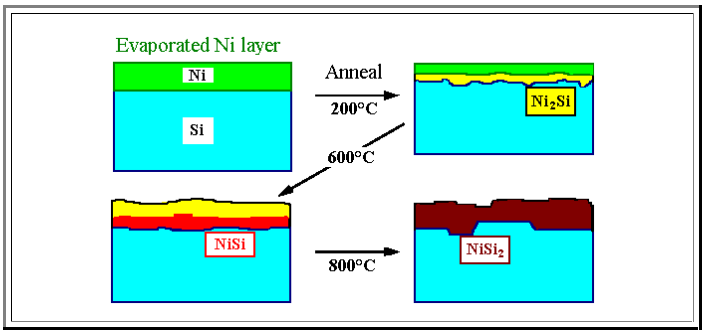
  - The **ACACACA...** stacking sequence of the hexagonal lattice fits right on the **ABCABCA..** stacking sequence of the **fcc** lattice on a **{111}** plane. The **Σ = 3** relationship is clearly visible; it is indicated by red dots and lines.
  - There are of course more complex geometries - if the **CSL** concept is not applicable; the **O**-lattice concepts has to be used.
- Sorting out the various types of possible defects is no longer an easy task. The interpretation of **TEM** micrographs may become quite involved.
  - Some examples will be discussed in the case studies in the next subchapter.

8.2 Case Studies

8.2.1 Ni Silicides

We will look in some detail on the system **Si - silicide - metal**, where many phase boundaries can be observed. The basic experiment consists of depositing a metal (here **Ni**) on **Si** (either in a **{100}** or **{111}** orientation), and induce some reaction by heating.

- Three different **Ni**-silicides will form consecutively:



Altogether five different phase boundaries may be encountered, some of which are shown in the picture above:

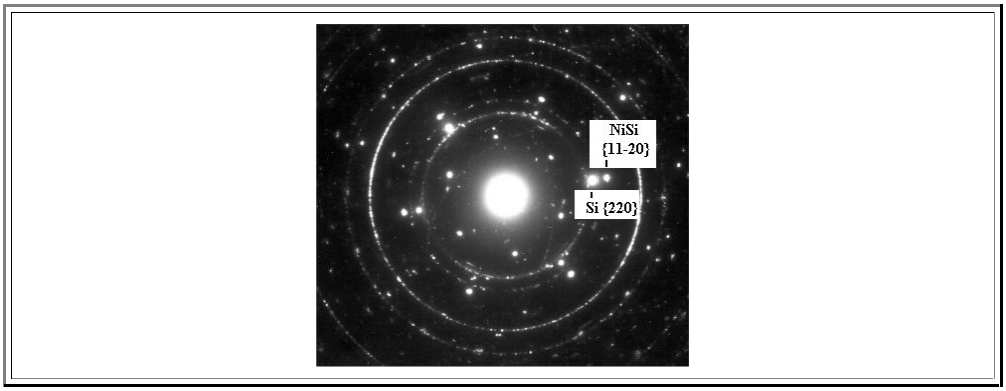
- **Si - Ni**,
  Si - Ni_2Si and **Ni_2Si - Ni**,
  Si - NiSi and **NiSi - Ni_2Si**,
  Si - $NiSi_2$, and **$NiSi_2$ - NiSi**.

Major findings are:

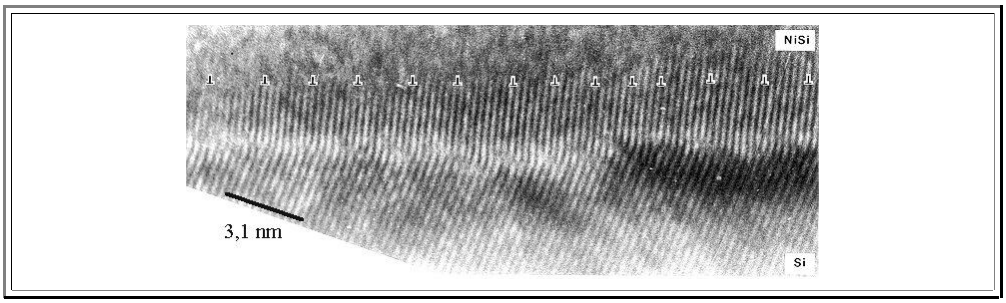
- The interface between **Si** and **Ni** does not really exist because immediately after the (room temperature) evaporation, a thin **Ni_2Si**-silicide layer forms between the **Si** and the **Ni**.
- The **Ni_2Si** layer is polycrystalline; the interface between **Si** and **Ni_2Si** seems to be incoherent - i.e. if there is any structure it is not observed with "normal" **TEM**.

The interface between **{111} Si** and **NiSi** is epitaxial, however, and thus semicoherent *against all expectations*:

- **NiSi** is reported to crystallize in an *orthorhombic* lattice; on **{111}** Si substrates, however, a *hexagonal* lattice is observed (which can be cobtained from an orthorhombic lattice by slight adjustments of the lattice parameters).
- The misfit is *extremely large* (ca. **15%**) and would require a distance of **0,6 nm** for **b = *a*/2<110>** misfit dislocations. Such a small spacing is usually considered to be too small to be meaningful - epitaxial relationships thus should not exist. The diffraction pattern, however, indicates a clear epitaxial relationship (with a bit of polycrystallinity as indicated by the rings):



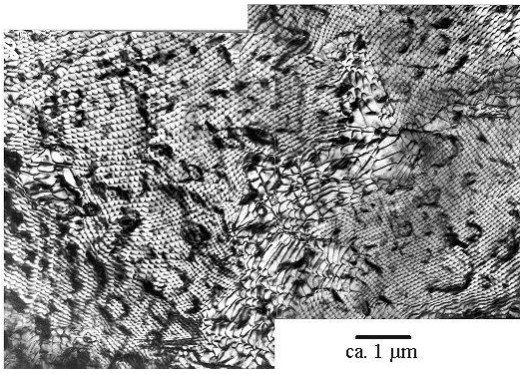
- While no structure can be seen in conventional **TEM**, high-resolution **TEM** shows pronounced misfit dislocations relieving some of the stress at a spacing of about **1,6 nm**. This is one of the densest misfit dislocation networks ever observed. The ending lattice planes are indicated by the edge dislocation symbol somewhat above the actual interface plane.



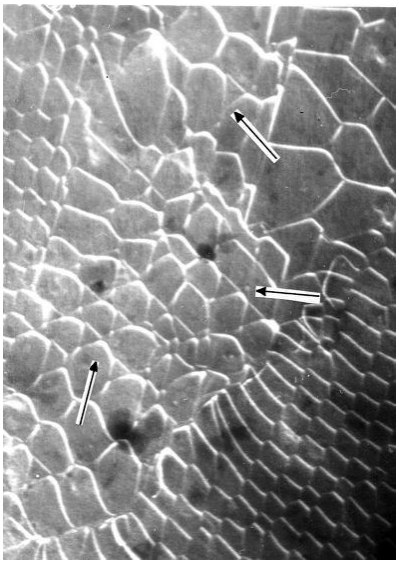
- The most interesting phase is **$NiSi_2$**; it is the final product after sufficient annealing at **800 °C**.
  - **$NiSi_2$** crystallizes in the cubic **CaF_2** - structure with a lattice constant that is only **0,3%** smaller than that of **Si**.
  - We thus can expect an epitaxial relationship with a misfit dislocation network at a spacing

$$p = b \cdot \frac{b}{(a_e - a_m)/a_m} = \frac{b}{0{,}003}$$

  - With a_{Si} **= 0,54 nm** and ***b* = *a*/2<110> = 0,382 nm** we would expect a network with a spacing of about **130 nm**.
  - What we see for an interface on a **{111}** plane looks like this:



  - This looks rather interesting. We seem to have a simple hexagonal network of dislocations, but we see some additional features: "Blackish" areas and an island with rather coarser structures embedded in a sea of something with a possible hexagonal symmetry.
- The reasons for these complications are two peculiarities of this interface, which can also be found in similar systems; in particular in the **Si - $CoSi_2$** interface.
  - *First*, it "likes" to be on **{111}**-planes. This leads to heavy facetting if the **Ni** layer is deposited on a **Si {100}** plane, but also to some facetting on **{111}**. This can be seen best in cross-section; an example is given in the illustration. We must expect that the accommodation of steps will introduce irregularities into the network.
  - *Second*, the interface is mostly *not* in a **Σ = 1** relation, i.e. with a direct continuation of the lattices, but in a **Σ = 3** relation. This means that the **$NiSi_2$** is *twinned* with respect to the substrate. An overview picture is shown in the link. This somewhat surprising result can be obtained from a careful contrast analysis of the network with micrographs taken at higher magnifications. The network then looks like this:



- Shown is one of the "islands" in a sea of regular hexagonal dislocations. Its structure looks somewhat familiar: The arrows point to extended stacking fault knots as in the case of the small angle twist grain boundary on **{111}** in **Si**.
- But in contrast to the network in the small angle twist boundary, all dislocations now are *edge dislocations*; as expected for misfit dislocations. The distance is also what would be expected for a almost fully relaxed layer of **$NiSi_2$**.

The question is, of course, why this mix of **$\Sigma = 1$** and **$\Sigma = 3$** relations? As in the case of the low angle twist boundary encountered before, nobody knows for sure. Obviously, the energy balance is rather similar for the two cases.

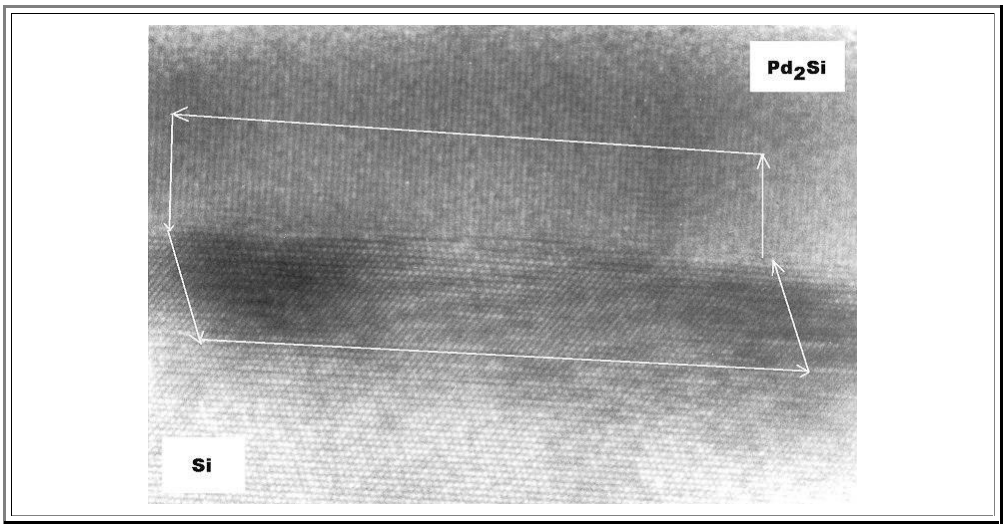
- Very similar interfaces have been observed in the case of **Si - $CoSi_2$** interfaces, which, except for a slightly larger misfit, have essentially the same geometry.

Despite the structural similarity to the small angle grain boundaries, the phase boundaries add new features and open questions. To get more insights, we will now discuss the case of the interface between (cubic) **Si** and (hex.) **Pd_2Si**.

8.2.2 Case Study for Pd Silicides

If **Pd** instead of **Ni** is evaporated on a clean **Si** surface, hexagonal **Pd_2Si** (***a* = 0,653 nm**, ***c* = 0,344 nm**) develops around **200°C - 300°C**. Increasing the annealing temperatures produces no new phases.

- The misfit of the hexagonal **{001}** plane to a **Si {111}** plane is **1,8%**; we can expect an epitaxial relationship with a misfit dislocation network at a spacing of about **10 nm**. This is around the resolution limit of **TEM** in a regular contrast mode, so we have to resort to **HRTEM** and cross-sectional specimen.
- A **HRTEM** image of the interface is shown below (*This picture from **1980** is of historical interest, too: It was, to the best of my knowledge, the first **HRTEM** picture ever obtained from a phase boundary*).



We clearly have an epitaxial layer of **Pd_2Si**. No ending lattice fringes denoting misfit dislocations are unambigously visible. Therefore a Burgers circuit has been drawn, somewhat analogous to the procedure used to obtain Franks formula. It goes up in the **Pd_2Si**, then to the left crossing **90** lattice fringes, back to the boundary, **90** fringes to the right and up to the boundary again. It does *not* close, although it is clear that it would have closed on a perfectly flat and misfit-dislocation free interface.

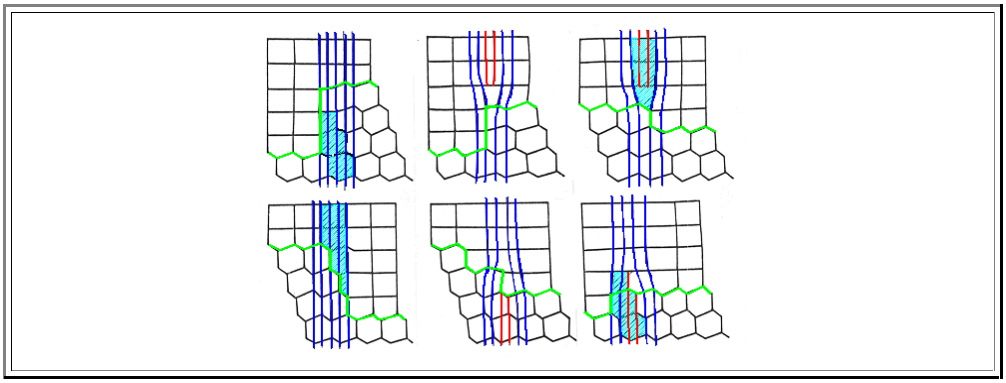
- Does this mean that the phase boundary contains misfit dislocations?
- This is not clear. We *may* have dislocations in the interface, but we *certainly* have steps as can be seen directly. We now have to pay some attention to the *relationship of dislocations and steps*, their images in a **HRTEM** picture, and their consequences for a Burgers circuit.

This will lead us into new and quite complicated territory. We will consider the relationship between steps and dislocations only for the example of hexagonal lattices on cubic lattices, or, more generally, for **Σ = 3** relations. This will be sufficient to gain an idea of the added complexity.

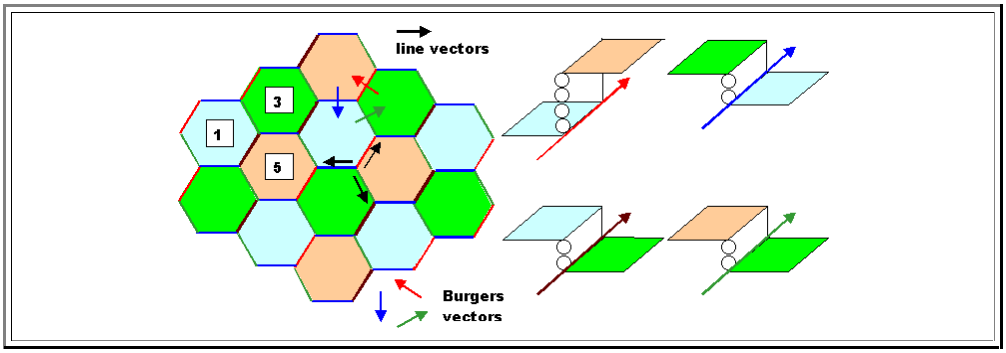
8.3 Steps in Interfaces

8.3.1 The Relation Between Steps and Dislocations in Σ = 3 Boundaries

- Playing with models of a perfectly fitting phase boundary between a hexagonal and a cubic lattice, one realizes quickly that steps can be incorporated without problems *and without dislocations* as long as the step height comes in multiples of **3** (in units of the translation vectors of the **CSL**). This, together with some other cases, is shown below:



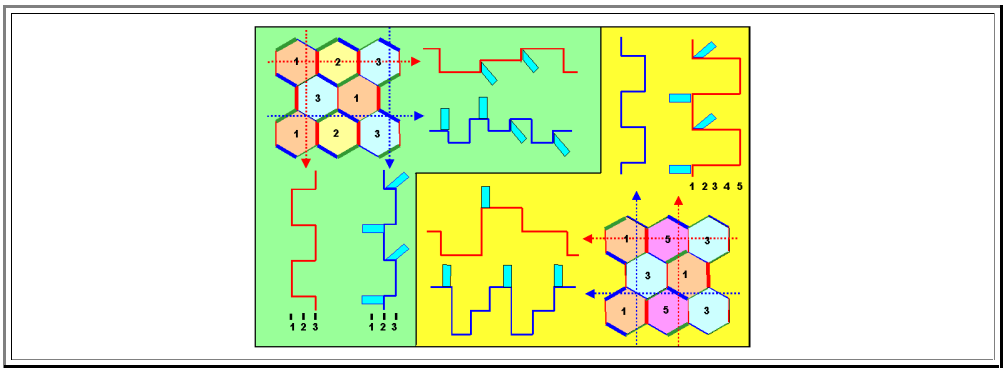
  - The left two pictures show *pure steps* (or **coherent steps**). For ease of interpretation, some lattice *planes* of the **DSC** lattice are shown in blue; ending lattice planes of the **DSC** lattice are red. Ending lattice *fringes*, as seen on a **HRTEM** micrograph, are indicated in light blue - note that they are not the same thing as lattice planes. The phase boundary itself is shown in light green.
  - Included in the drawing are also two *pure **DSC** lattice dislocations* (middle pictures). They are true dislocations because they can be constructed with the Volterra method as demonstrated before, and they have the **1/*r*** stress field that is a hallmark of dislocations. Sometimes they are called **coherency dislocations**.
  - Finally, a mix of pure dislocations and pure steps is shown on the right. It is evident that steps going just one plane up or down must be a mix of pure steps and pure dislocations. The same is true for steps going **4** planes up or down and so on. These dislocations are sometimes called **anti-coherency dislocations**.
  - Note that there is no way of having a combination of pure steps and dislocations with **step height zero**.
- An unexpected property emerges: **Pure steps** (sometimes also called **coherent ledges**) show ending lattice fringes in a **HRTEM** micrograph, whereas true dislocations in this case are *not* associated with ending lattice fringes.
  - Where does that leave us? Is Franks formula, which after all counted Burger vectors in a circuit not unlike the one shown before, not applicable to non-planar boundaries? Why can we see ending lattice fringes in a **TEM** picture and there is no dislocation?
  - Well, ending lattice *fringes* (again note we call it lattice *fringe* on purpose) are not lattice *planes*, and at a grain- or phase boundary *all lattice planes of one kind end* and some of a new kind start. The fact that some visible fringes appear to be continuous in a "fuzzy" projection of the lattices, has no particular meaning in itself. Of course, one lattice plane ending in one crystal may give rise to a lattice fringe ending on a micrograph, too, and thus signify an edge dislocation, *but this must not be generalized*.
- We may, however, make an important generalization of a different kind: A semi-coherent phase or grain boundary, in general, needs at least *two qualitatively different kinds of defects* in its interface:
  - *Pure **DSC**-lattice dislocations* (generally associated with an intrinsic step), and
  - *Pure steps* (without dislocation character, i.e. without a long range stress field).
- In general *both* defects are required as we will see if we now compose a (hexagonal) dislocation network in a **Σ = 3** boundary.
  - Whereas the dislocation network has a perfect threefold rotational symmetry, the boundary is less symmetric. This can be seen when we consider the steps introduced by the dislocations (use the picture above), too:



- As soon as we defined the line- and Burgers vectors, we realize that the steps associated with pure dislocation are always the same: Looking in the line direction, crossing a pure dislocation would always lead two steps down in this example. If we start a closed circuit at the hexagon labelled "(level) **1**" and go across the green dislocation, we end up two atomic planes down on level **3**. The same happens if we now cross the blue dislocation; we are down to level **5**.
- *However*, in closing the circuit, we must necessarily come up to level **1** again, This is only possible if the dark red dislocations break the symmetry and contain *two* steps (**–2 + 6 = 4**). Thus we go two levels down and **6** levels up, which is just right.

This feature, which is clearly a general feature of all boundaries, opens up a whole new can of worms.

- There is more than one way to combine pure steps and pure dislocations to create a network that satisfies the requirements for accommodating misfit (this needs the dislocations) and to compensate for the steps introduced by the dislocations (this needs pure steps).
- The image of a given boundary in cross-sectional **HRTEM** can look very different, depending on what kind of possible configuration is cut which way. Lattice fringes may end in several different ways.
- The closing failure of a Burgers circuit that counts lattice fringes around a large part of an interface thus cannot be considered as a net count of dislocation Burgers vectors. Combined with a net count of the steps in the interface, however, it may be useful.
- The following graphic illustrates these points



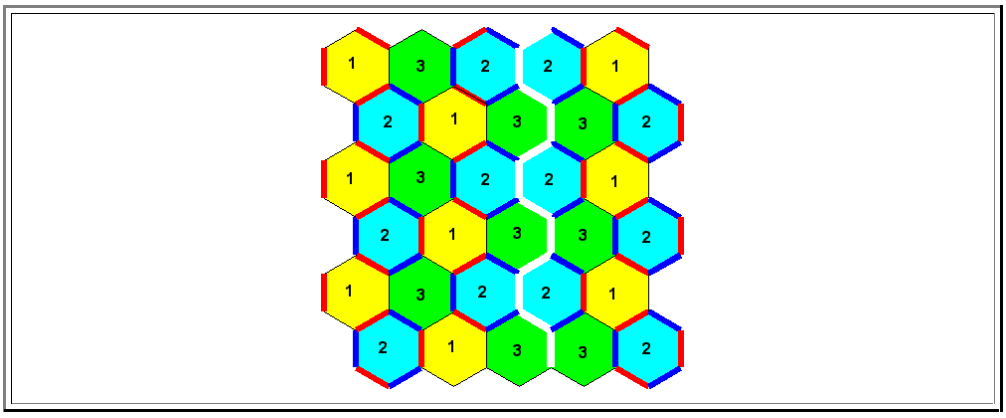
Shown are two possible combinations of dislocations and steps in **Σ = 3** boundaries (of any kind). Dislocations in combination with a coherent step are indicated in bold lines; the numbers in the hexagons indicate the level of the boundary

- Two possible geometries are shown in the upper left-hand corner and the lower right-hand corner
- Four cross-section through the dislocation/step network are drawn in together with their schematic image in **HRTEM**. Ending lattice fringes are indicated in light blue (assuming without justification that the image of dislocation/step combinations that are inclined with respect to the electron beam add no further complications).

It becomes clear that the interpretation of an **HRTEM** image can be a demanding task which will not necessarily give an unambiguous answer. You may try your skills at the picture in the illustration.

But we are still not done with the discussion of the intrinsic geometry of a simple **Σ = 3** boundary. Even if we assume that we have a dislocation/step network of a defined kind (e.g. the one from the upper left-hand corner of the above drawing) and that the boundary is flat apart from the ups and downs of the dislocations, we must expect an *added complication*:

- Since the dislocations/step network most likely formed in small patches and then spreads out, individual patches may be out of "synch", i.e. whenever they meet they will not fit together. This is illustrated below



- Dislocations in combination with a coherent step are shown in bold; the color now denotes if the step goes up or down as seen from the inside of the hexagons completely enclosed by "bold" dislocations. The dislocations/step network on the left and right side are identical, but displaced relative to each other by one hexagon.

- Along the white line, they obviously don't match. We would need dislocations with a step height of zero, which as we have seen before, do not exist in this geometry.

- The only way out is to postulate a new kind of defect, some kind of stacking fault in the dislocation/step lattice. To the best of my knowledge, such a defect has not yet be named or discussed in detail - although it is clearly a necessary feature of general phase- or grain-boundaries.
  - This serves to illustrate that the last word about structural aspects of defects in crystal has not yet come in. One may ask, of course, if esoterica like the dislocation/step network considerations are of any importance. The answer is: *Who knows?*
  - Considering however, that many materials (including natural minerals) are full of phase and grain boundaries, that many properties of theses boundaries are directly linked to their structure and that not much is known about the atomic structure of non-trivial boundaries, it is not totally unrealistic to expect that research will go on.
- In the next (and last) subchapter we will briefly look at some more questions in relation to phase boundaries.

8.3.2 Open Questions

- The final conclusion in the phase boundary chapter is simple: *There is still much to do!* Some open questions will just be mentioned:
  - If the phase boundary moves into the interior of some material in a reactive process like silicide formation, the phase boundary dislocations must climb. How do they do this?
  - Supposedly, the climb of phase boundary dislocations needs a specified current of point defects (whatever is needed to accommodate the climb rate given by the speed of the advancing interface). Will the point defects assisting climb affect the kinetics of the phase boundary movement? How?
  - Does the hexagonal dislocation network have a preferred direction. How can this be proven?
  - What kind of defect forms the boundary between different network domains?
- Many more questions can be formulated to understand just the structural properties of interfaces in general. But here I stop. I have done my share of the work. Now it is up to you.