1. Introduction

1.1 Scope of the Course

Thou shall live in interesting times!
Chinese curse

1. Introduction

1.1 Scope of the Course

1.1.1 Goals and Contents

- This course tries to do something next to impossible: Give a review of all important or interesting semiconductor materials, products and technologies, *excluding only Silicon* as far as it is used for microelectronics! Moreover, the course aspires to go into some detail with respect to the fundamental material properties of semiconductors, i.e. it requires a heavy dose of semiconductor physics.
- The course is intended for Materials Science and Engineering students pursuing a Master's degree. A number of problems come to mind:
  - One term is surely not sufficient to cover this field adequately. We will therefore focus on whatever seems most interesting for the last third of the course (and then, maybe, finish the still open chapters).
  - All semiconductor physics text books above a very basic level invariably assume that the reader is thoroughly familiar with quantum theory and at least some solid state physics and thermodynamics. This is not true for the average Materials Science and Engineering student with a Bachelor degree. We will try to cope as best as we can.
  - There is no book (as far as I can tell after searching around for a while) that comes even close to what is intended here, so you must be content with this script.
- Fortunately, there are some helpful circumstances, too:
  - By the time they take this course, *all* materials science students should be well acquainted with **Si** technology as far as it relates to microelectronics, with solid state physics in general, and with classical thermodynamics.
  - Some, if not most, also took the electives "Quantum theory" and "Statistical Thermodynamics" - that will help.
  - Moreover, all Materials Science Bachelors in Kiel had a heavy dose of basic semiconductor physics and devices in their 4th semester where they had to take "Introduction to Materials Science II".
- What will be the content? Who knows before actually doing it? What will be attempted can be seen in the link, which gives a reasonably detailed outline of the intentions.
  - What will be really presented will depend on how fast we will be able to move and on the questions and problems coming while teaching the course.
- All things considered, the general background should be sufficient to understand the basic concepts and to translate that understanding to an appreciation of general semiconductor physics, products and technology. In summary:

Let's go!

1.1.2 Relation to Other Courses

- The graduate course "Semiconductors" interacts with several other courses in the materials science curriculum. A certain amount of overlap is unavoidable.
- Of course, "semiconductors" relates to the truly basic courses in materials science, like *thermodynamics*, *analytics*, or *solid state physics* (all of which are required for all students) - it simply relies on the stuff covered there.
  - Below is a list of courses that are electives (at least for some), but have some relation to "Semiconductors".

- **Electronic Materials** (Werkstoffe der Elektrotechnik und Sensorik I; Prof. Föll) or its successor **Advanced Materials B**..
  - Required for Masters students.
  - Focuses on a short introduction to Silicon technology and materials but covers mainly dielectric and magnetic materials.
- **Si Technology I + II** (Prof. Wagner)
  - Elective for Master students.
  - Covers **Si** and **Si** technology (with emphasize on **MEMS**), which is not included in this course.
- **Quantum Theory for Materials Scientists** (Dr. Carstensen)
  - Elective for all students
  - A *must* if you really want to understand semiconductors
- **Thin Solid films I** (Various lecturers)
  - Elective for Masters students
  - Perfectly complements the technological part of "Semiconductors".
- **Defects in Crystals** (Prof. Föll)
  - Elective for all students
  - Most semiconductors except **Si** are plagued by crystal defects, so some basic knowledge of crystal lattice defects is helpful but not really necessary.

1.1.3 Required Background Knowledge

Mathematics

- Advanced Math as taught in all undergraduate courses should be enough. You should not be intimidated by long equations and have a feeling for the important parts in mathematical deductions.

General Physics and Chemistry

- A general undergraduate level of basic physics should be sufficient. You should be comfortable with units and conversion between units. Not much chemistry will be needed, but you should know, e.g., that **KOH** is a base and **HF** an acid and that **SiO_2** is quartz which is pretty stable in most chemicals, and so on.

General Materials Science

- You must be comfortable with all crystallographic notions and the reciprocal lattice. Thermodynamics, especially in its statistical form as expressed by distribution functions, is a must. Some preliminary understanding of semiconductors - conductivity, holes and electrons, junction in general - is necessary.

Quantum Theory

- A good understanding of quantum theory up to the level of the free electron gas model is essential.

Thermodynamics

- A working knowledge of classical thermodynamics is helpful. Simple statistical thermodynamics as expressed by distribution functions is a must. Knowing in detail what "chemical potential" means would be nice, but it's not really required.

1.1.4 Organization

- The course will consist of two parts:
  - A regular *lecture*, lasting **90** minutes, once a week
  - Exercises, which may be held as follows:
    - As *regular exercises*, i.e. solving problems.
    - As a *seminar* , where a team of two students prepares and delivers a **45** min. lecture on a specific subject.
    - As a *mixture*, i.e. first half with regular exercises, second half with seminar.
- All information concerning the running term (including details like schedules, grading, exams, etc.) can be found in a separate document via the link.

2. Semiconductor Physics

2. Semiconductor Physics

2.1 Basic Band Theory

2.1.1 Essentials of the Free Electron Gas

- This course requires that you must be familiar with some solid state physics including a working knowledge of thermodynamics and quantum theory.
  - The **free electron gas** model is a paradigm for the behavior of electrons in a crystal, you should be *thoroughly familiar* with it.
  - In case of doubt, refer to the Hyperscript "MaWi II" – which, however, is in *German*.
  - In the following, the *essentials of the model* are repeated – briefly, without much text. If you have serious problems with the topic already here, you do indeed have a problem with this course!

The Energy Levels of an Electron in a Constant Potential

- The **free electron gas** model works with a **constant potential**. This is, of course, a doubtful approximation; essentially only justified because it works – up to a point.
  - **Approximations:** Constant potential $U = U_0 = 0$ within a crystal with length L in all directions; only *one* electron is considered.
  - The true potential outside the crystal will turn out to be irrelevant due to *periodic* boundary conditions (see below).
- The major formulas and interpretations needed are:
- Time independent one-dimensional **Schrödinger equation** :

$$\boxed{-\frac{\hbar^2}{2m_e}\left(\frac{\partial^2\psi(x,y,z)}{\partial x^2}+\frac{\partial^2\psi(x,y,z)}{\partial y^2}+\frac{\partial^2\psi(x,y,z)}{\partial z^2}\right)\underbrace{+\,U(\underline{r})\cdot\psi(x,y,z)}_{=\,0} \;=\; E\cdot\psi(x,y,z)}$$

  - **ħ = h "bar" = $h/2\pi$** = Plancks constant/**2**π
    m_e **=** electron mass
    ψ **=** wave function
    ***E* = total energy =** kinetic energy **+** potential energy. Here it is identical to the *kinetic energy* because the *potential energy* is zero.
- **Potential $U(x)$** as defined above; i.e. $U(x) = U_0 = \text{const} = 0$ for $0 \leq x \leq L$. (Remember that L is the macroscopic size of the crystal.)
  - For the (potential) energy, there is always a free choice of zero point; here it is convenient to put the bottom of the potential well at zero potential energy. We will, however, change that later on.
- **Boundary conditions**: Several choices are possible! Here, we use **periodic conditions** (also called **Born–von Karman** conditions), leading to a so-called supercell:

$$\boxed{\psi(x + L) \;=\; \psi(x)}$$

- **Solution** (for **3**-dimensional case)

$$\psi \;=\; \left(\frac{1}{L}\right)^{3/2} \cdot \exp\,(\mathrm{i} \cdot \underline{k} \cdot \underline{r})$$

$$\begin{aligned}
\underline{r} &= \textit{position vector}\\
&= (x,\, y,\, z)\\[1ex]
\underline{k} &= \textit{wavevector}\\
&= (k_x,\, k_y,\, k_z)\\[1ex]
k_x &= \pm n_x \cdot 2\,\pi/L\\
k_y &= \pm n_y \cdot 2\pi\,/L\\
k_z &= \pm n_z \cdot 2\pi/L\\[1ex]
n_x\,,\, n_y,\, n_z &= \text{quantum numbers}\\
&= 0, 1, 2, 3, \ldots\\[1ex]
\mathrm{i} &= (-1)^{1/2}
\end{aligned}$$

- A somewhat more general form for crystals with unequal sides can be found in the link.

There are infinitely many solutions, and every individual solution is selected or described by a set of the three quantum numbers n_x **,** n_y, n_z . The solution ψ describes a plane wave with amplitude **$(1/L)^{3/2}$** moving in the direction of the **wave vector $\underline{k}$** .

Next, we extract related quantities of interest in connection with moving particles or waves:

- The **wavelength** λ of the "electron wave" is given by

$$\lambda \;=\; \frac{2\pi}{|\underline{k}|} \;=\; \frac{2\pi}{k}$$

- The momentum $\underline{p}$ of the electron is given by

$$\underline{p} \;=\; \hbar \cdot \underline{k}$$

- From this and with **m** = electron mass we obtain the velocity $\underline{\mathrm{v}}$ of the electron to be

$$\underline{\mathrm{v}} \;=\; \frac{\underline{p}}{\mathrm{m_e}} \;=\; \frac{\hbar \cdot \underline{k}}{\mathrm{m_e}}$$

The numbers n_x, n_y, n_z are **quantum numbers** ; their values (together with the value of the **spin**) are characteristic for one particular solution of the Schrödinger equation of the system. A unique set of quantum numbers (alway plus one of the two possibilities for the spin) describes a **state** of the electron

- Since these quantum numbers only appear in the wave vector $\underline{k}$, one often denotes a particular wave function by *indexing it with **k** instead of $n_{x,\, y,\, z}$* because a given $\underline{k}$ vector denotes a particular solution or **state** just as well as the set of the three quantum numbers.

$$\psi_{n\mathrm{x},\, n\mathrm{y},\, n\mathrm{z}}(x{,}y{,}z) \;=\; \psi_{\mathrm{k}}(x{,}y{,}z) \;=\; \psi_{\mathrm{k}}(\underline{r})$$

- In other words, in a formal, more abstract sense we can regard the wave vector as a kind of vector quantum number designating a special solution of the Schrödinger equation for the given problem.

Since in the present case the total energy E is identical to the kinetic energy $E_{\mathrm{kin}} = ½\mathrm{m_e v}^2 = ½p^2/\mathrm{m_e}$, we have

$$E \;=\; \frac{\hbar^2 \cdot k^2}{2\mathrm{m_e}}$$

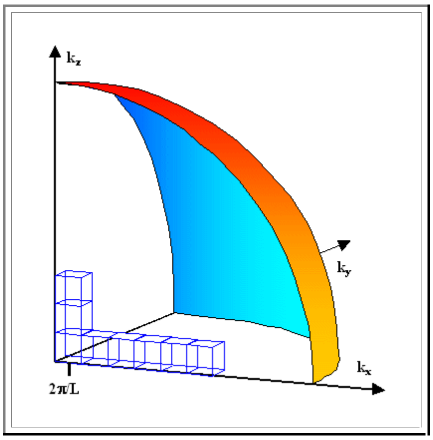
- We now have expressed the *total energy* as a function of the *wave vector*. Any relation of this kind is called a **dispersion function**. Spelt out we have

$$E = \frac{\hbar^2}{2m_e} \cdot \left(\frac{2\pi}{L}\right)^2 \cdot \left(n_x^2 + n_y^2 + n_z^2\right)$$

- This is the first important result: There are only *discrete* energy levels for the electron in a box with constant potential that represents the crystal.
  - This result (as you simply must believe at this point) will *still be true* if we use the *correct potentials*, and if we consider *many* electrons. The formula, however, i.e. the relation between energy and wave vectors may become *much more complicated*.
- The boundary conditions chosen and the length ***L*** of the box are somewhat arbitrary. We will see, however, that they do not matter for the relevant quantities to be derived from this model.

Density of States

- Knowing the energy levels, we can count how many energy levels are contained in an interval Δ***E*** at the energy ***E*** . This is best done in ***k***-space or **phase space**.



  - For the free electron gas, in phase space a surface of constant energy is a sphere, as schematically shown in the picture.
  - Any "state", i.e. solution of the Schroedinger equation with a specific ***k***, occupies the volume given by one of the little cubes in phase space, corresponding to the discrete states just discussed.
  - The number of little cubes fitting inside the sphere at energy ***E*** thus is the number of *all* electronic states ψ up to ***E***.
  - Since every state (characterized by its set of quantum numbers n_x, n_y, n_z) can accommodate *2 electrons* (one with **spin** up, one with spin down), the total number of electrons that can occupy states up to ***E*** is twice the number of little cubes; let this number (that includes the spin degeneracy) be $N_S(E)$.

  - Looking just at the energy, it is clear that at higher energies there are more states available. Counting the number of little cubes just in an energy interval ***E***, ***E*** + Δ***E*** corresponds to taking the difference of the numbers of cubes contained in a sphere with "radius" ***E*** + Δ ***E*** and ***E***.
  - We thus obtain the **density of states *D(E)*** as

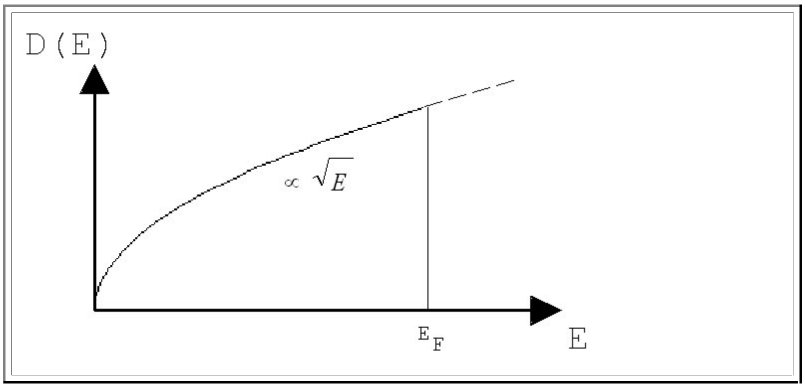
$$D(E) = \frac{1}{V} \cdot \frac{N_S(E + \Delta E) - N_S(E)}{\Delta E} = \frac{1}{V} \cdot \frac{dN_S}{dE} = \frac{1}{L^3} \cdot \frac{dN_S}{dE}$$

  - Note that ***D*(*E*)** is a density both with respect to ***E*** and to $V = L^3$ (= volume of the crystal).
- The final formula is

$$D(E) = \frac{1}{L^3} \cdot \frac{dN_S}{dE} = \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2}\right)^{3/2} \cdot E^{1/2}$$

  - The derivation of this formula and more to densities of states (including generating some numbers) can be found in the link.
- Some important points are:

- - ***D*(*E*)** is proportional to $E^{1/2}$.
  - For different (but physically meaningful) boundary conditions we obtain the same ***D*** (see the exercise **2.1** below).
  - The artificial length ***L*** disappears because we are only considering specific quantities, i.e. volume densities.
  - ***D*** is kind of a *twofold* density: It is first the density of energy states in an *energy interval* and second the (trivial) density of that number in *space* .
- If we fill the available states with the available electrons at a *temperature of* **0 K** (since we consider the free electrons of a material this number will be about **1** [or a few] per atom and thus is principally known) starting from ***E* = 0**, we find a special energy called **Fermi energy *E*$_F$** at the value where the last electron finds its place.



- In order to get (re)acquainted with the formalism, we do two simple exercises:

Exercise 2.1-1	**Exercise 2.1-2**
Solution of the free electron gas problem with fixed boundary conditions	Density of states as a function of dimensionality

Carrier Statistics

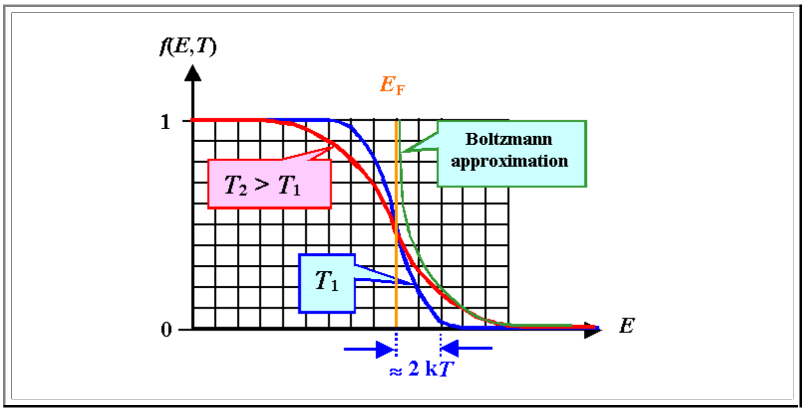
- We have the number of energy states for a given energy interval and want to know how many (charge) ***carriers*** we will find in the same energy interval *in thermal equilibrium* . Since we want to look at particles other than electrons too, but only at charged particles, we use the term "carrier" here.
- In other words, we want the distribution of carriers on the available energy levels satisfying three conditions:
  - The **Pauli exclusion principle**: There may be at most **2** carries per energy state (one with spin "up", one with spin "down"), not more.
  - The **equilibrium condition**: Minimum of the appropriate **thermodynamic potential**, here always the ***free Enthalpy G*** (also called **Gibbs energy**).
  - The **conservation of particles** (or charge) condition; i.e.*constant number of carriers* regardless of the distribution.
- The mathematical procedure involves a variation principle of ***G***. The result is the well-known **Fermi –Dirac** distribution **f(*E*,*T*)**:
  - **f(*E*,*T*) =** probability for occupation of (one!) state at ***E*** for the temperature ***T***

$$f(E, T) = \frac{1}{\exp\left(\frac{E - E_F}{kT}\right) + 1}$$

  - If you are not very familiar with the distributions in general or the Fermi distribution in particular, read up on it in the (German) link.
- This is the "popular" version with the **Fermi energy *E*$_F$** as a parameter. In the "correct" version, we would have the **chemical potential** μ instead of ***E*$_F$**.

  - Since the Fermi energy is a quantity defined independently of the equilibrium considerations above, equating $\boldsymbol{E}_\mathbf{F}$ with μ is only correct at $\boldsymbol{T} = \mathbf{0}$ **K**. Most textbooks emphasize that small differences may occur at larger temperatures, but do not explain what those differences are. We, like everybody else, will ignore these fine points and use the term "*Fermi Energy*" without reservations.
  - In all experience, many students (and faculty) of physics or materials science have problems with the concept of the "*chemical potential*". This is in part psychological (we want to do semiconductor physics and not chemistry), but mostly just due to little acquaintance with the subject. The link provides some explanations and examples which might help.

- The Fermi-Dirac distribution has some general properties which are best explained in a graphic representation.



  - It contains a convenient definition of the Fermi energy: The energy where exactly half of the available levels are occupied (or would be occupied if there would be any!) is the Fermi energy:

$$\mathrm{f}(E = E_F) \;=\; \frac{1}{2}$$

  - The width of "soft zone" is ≈ **4 k*T* = 1 meV** at **3 K**, and **103 meV** at **300K**.
  - For $\boldsymbol{E} - \boldsymbol{E}_\mathbf{F} \gg \mathbf{k}\boldsymbol{T}$ the **Boltzmann** approximation can be used:

$$\mathrm{f_B}(E,T) \;\approx\; \exp\left(-\frac{E - E_\mathrm{F}}{\mathrm{k}T}\right)$$

  - In this case the exclusion principle is not important because there are always plenty of free states around – the electrons behave akin to classical particles.

- This leads to the final formula for the incremental number or **density of electrons**, **d*n***, in the energy interval ***E*, *E* + Δ*E*** (and, of course, in thermodynamic equilibrium).

In words:	Formula
Density of electrons in the energy interval *E*, *E* + Δ*E* = **density of states *times* probability for occupancy *times* energy interval**	$\mathrm{d}n \;=\; D(E)\cdot \mathrm{f}(E,T)\cdot \mathrm{d}E$

- This is an extremely important formula, which is easily generalized for almost everything. The number (or density) of something is given by the density of available places times the probability of occupation.
  - This applies to the number of people found in a given church or stadium, the number of photons inside a "black box", the number of phonons in a crystal, and so on.
  - The tricky part, of course, is to know the probabilities or the **distribution function** in each case. However, if we do not consider church goers or soccer fans, but only physical particles (including electrons and holes, but also "**quasi-particles**" like phonons, excitons, ...), there are only *two* distribution functions (and the Boltzmann distribution as an approximation): The Fermi-Dirac distribution for Fermions, and the Bose-Einstein distribution for Bosons. Mother nature here made life real easy for physicists.

- Since all available electrons must be somewhere on the energy scale, we always have a normalization condition for the total electron density ***n***:

$$n = \int_0^\infty D(E) \cdot f(E,T) \cdot dE$$

Questions

Quick Questions to 2.1.1

2.1.2 Diffraction of Electron Waves

- ***Note:*** *It is too tiresome to underline all vectors, and we will simply stop doing it except if it is absolutely necessary*.

The Reciprocal Lattice

- Electron waves like all waves experience diffraction effects in periodic structures like crystals. This is best described in the *reciprocal lattice* of the crystal in question. There are several ways to construct a *reciprocal* lattice from a *space* lattice.
  - Remember that a *lattice* – in contrast to a *crystal* – is a *mathematical construct*. A lattice becomes becomes a crystal by putting a set of atoms – the **base** of the crystal – at every lattice point.
  - There are **14** different kinds of lattices – the **Bravais lattices** – with different symmetries that are sufficient to describe the lattice of any crystal.
  - If you are unsure about this topic, refer to the appropriate chapter in "Introduction to Materials Science I" (in German) or in "Defects" (in English)
- Lets look at three definitions of the reciprocal lattice, which are – of course – all equivalent but at different levels of abstraction. This is elaborated in the link.
  - **1.** The reciprocal lattice is the **Fourier** transform of the space lattice.
  - **2.** The reciprocal lattice with an elementary cell (***EC***) as defined by the base vectors $\boldsymbol{b}_{1,2,3}$ is obtained from the space lattice as defined by its base vectors $\boldsymbol{a}_{1,2,3}$ by the equations

$\boldsymbol{b}_1 = 2\pi \dfrac{\boldsymbol{a}_2 \times \boldsymbol{a}_3}{\boldsymbol{a}_x \times (\boldsymbol{a}_y \cdot \boldsymbol{a}_z)}$	$\boldsymbol{b}_2 = 2\pi \dfrac{\boldsymbol{a}_3 \times \boldsymbol{a}_1}{\boldsymbol{a}_x \times (\boldsymbol{a}_y \cdot \boldsymbol{a}_z)}$	$\boldsymbol{b}_3 = 2\pi \dfrac{\boldsymbol{a}_1 \times \boldsymbol{a}_2}{\boldsymbol{a}_x \times (\boldsymbol{a}_y \cdot \boldsymbol{a}_z)}$
$\boldsymbol{a}_x \times (\boldsymbol{a}_y \cdot \boldsymbol{a}_z)$ = **volume *V* of EC**		

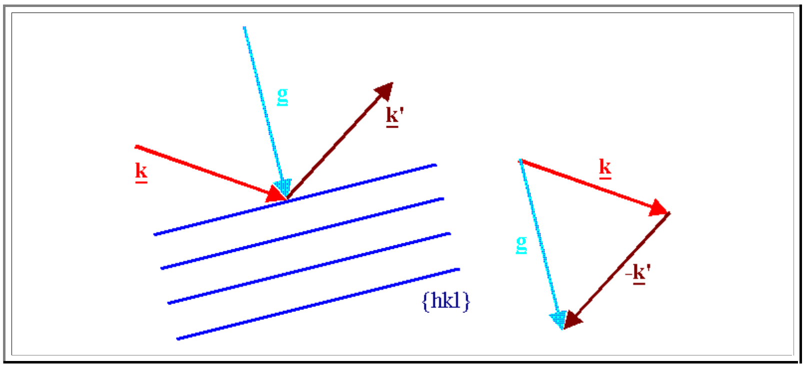
  - **3.** The base vectors of the reciprocal lattice can be constructed by drawing vectors perpendicular to the three **{100}** planes of the space lattice and taking their length as $2\pi/d_{hkl}$ with d_{hkl} **=** spacing between the *lattice* planes with **Miller indices {hkl}**.
  - *You will, of course, never confuse the spacing between lattice planes with the spacing between crystal planes, i.e. sheets of atoms, which may be something different.*

Bragg Condition

- If a wave impinges on a crystal – and it doesn't matter if it is an electromagnetic wave, e.g. **X**-rays, or an electron, or neutron "wave" – it will be reflected at a particular set of lattice planes **{hkl}** characterized by its reciprocal lattice vector ***g*** *only* if the so-called **Bragg condition** is met.
  - Let the wave vector of the incoming wave be ***k***, the wave vector of the reflected wave is ***k'***. The Bragg condition correlates the three vectors involved – ***k*** , ***k'***, and ***g*** – in the *simplest possible form*:

$$\boxed{\underline{\boldsymbol{k}} - \underline{\boldsymbol{k}}' = \underline{\boldsymbol{g}}}$$

  - There is no simpler relation correlating three vectors – mother nature again makes life as easy as possible for us!
- This Bragg condition is easily visualized:



- *If*, and only *if* the three vectors involved form a *closed triangle*, is the Bragg condition met. If the Bragg condition is *not* met, the incoming wave just moves through the lattice and emerges on the other side of the crystal (neglecting absorption).

So far we assumed implicitly that the diffraction (or, in more general terms, the **scattering**) of the incoming wave is **elastic**, i.e. the magnitude of ***k'*** is identical to that of ***k***, i.e. **|*k*| = |*k'*|** – which means that the diffracted wave has the same momentum, wavelength and especially energy as the incoming wave.

- This is not necessarily implied by Bragg's law – the equation ***k*** – ***k*** ' = ***g*** can also be met if ***k'*** differs from ***k*** by a reciprocal lattice vector; i.e. if scattering with a change of energy takes place. This will become important later.

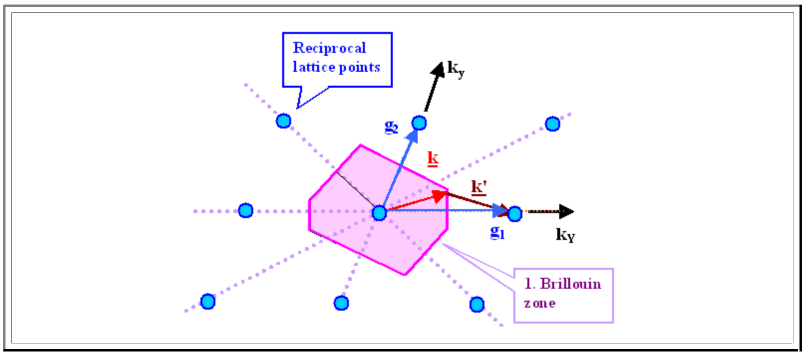
Brillouin Zones

In **X**-ray analysis of crystals we often deal with an (idealized) situation were *one* well-defined plane wave with fixed wavelength impinges on a crystal, and the question is which set of lattice planes will reflect the *one* incoming beam.

- This question can be answered most easily in a qualitative way by a geometric construction called the **Ewald construction** which can be found in the "Introduction to Materials Science II" Hyperscript.

However, if we consider the diffraction effects occurring to the free electrons contained in the crystal, we are looking at a (quasi) continuum of wave vectors: We have all possible directions and many wavelengths.

- The question is now a bit different: We want to know which *particular* wave vectors out of *many* (an infinite set, in fact) meet the Bragg condition for a given crystal lattice plane.
- This question can be answered most easily and semi-quantitatively by a geometric construction called the **Brillouin construction**. Lets look at it in a simplified two-dimensional form.



- If we construct **Wigner**–**Seitz** **cells** in the *reciprocal lattice* as shown by the pink lines, all wave vectors ending on the Wigner-Seitz cell walls will meet the Bragg condition for the set of lattice planes represented by the cell wall.

The Wigner–Seitz cells form a nested system of polyhedra which can be numbered according to size. These cells are called **Brillouin zones** (***BZ***); the smallest one is called the **1. BZ**, the next smallest one the **2. BZ**, and so on.

- All wave vectors that end on a **BZ** wall will fulfill the Bragg condition and thus are diffracted. Of course, the origin must be at the center of that **BZ**.
- Wave vectors completely in the interior of the **1. BZ**, or in between any two **BZ**s, will *never* get diffracted; they move pretty much as if the potential would be constant, i.e. they behave very close to the solutions of the free electron gas.

Of course, if we talk about diffraction in a crystal, we assumed implicitly that the potential is no longer constant, but *periodic* with the crystal.

- The statement above is thus not trivial, but a first important conclusion from diffraction geometry alone: We have good reasons (albeit no ironclad justification) to believe that the free electron gas model is a *decent approximation* for electrons with wave vectors *not* ending on (or close to) a Brillouin zone edge of the crystal in question.
- We thus only have to consider what happens to the electrons with wave vectors on or close to a **BZ** edge. In which properties do they differ from electrons of the free electron gas model?

Questions

Quick Questions to 2.1.2

2.1.3 Energy Gaps and General Band Structure

- Electron waves with wave vectors on or near a **BZ** edge are diffracted; all others aren't.
  - This means simply that electrons with wave vectors near or at a **BZ** edge – let's call them $\boldsymbol{k}_{\mathbf{BZ}}$ electrons – feel the periodic potential of the crystal while the others *do not* (in a first approximation).
  - In other words, $\boldsymbol{k}_{\mathbf{BZ}}$ electrons *interact* with the crystal, and this must express itself in their energies.

Origin of Energy Gaps

- Intuitively, we expect that "normal" electrons, not feeling any diffraction, pretty much obey the relation for the total energy ***E*** as before:

$$\boxed{E \;=\; E_{\text{kin}} \;=\; \frac{p^2}{2m} \;=\; \frac{(\hbar k)^2}{2\text{m}}}$$

  - For $\boldsymbol{k}_{\mathbf{BZ}}$ electrons, however, we must expect major modifications.
- What we will get in the most general terms is a **splitting of the energy value** if a given ***k*** ends exactly at the Brillouin zone edge, i.e. for a $\boldsymbol{k}_{\mathbf{BZ}}$ electron. Instead of $\boldsymbol{E(k) = (\hbar k)^2/2m_e}$ we obtain.

$$\boxed{E(k_{\text{BZ}}) \;=\; \frac{(\hbar k_{\text{BZ}})^2}{2\text{m}} \;\pm\; \Delta E}$$

  - In words: Electrons at the **BZ** edge can have *two* energies for the same wave vector and thus state. One value is somewhat lower than the free electron gas value, the other one is somewhat higher.
- Energies between these values are *unobtainable* for any electron – there is now an **energy gap** in the $\boldsymbol{E = E(k)}$ relation for all ***k*** vectors ending on a Brillouin zone wall.
- The time-honored way to visualize this energy gap is to look at a **one-dimensional** crystal – i.e. a *chain of atoms* , periodically spaced with the distance ***a***.
  - Since in this case, we have for the electron wave meeting the Bragg condition ...

$$\boxed{k' \;=\; -k \;=\; -k_{\text{BZ}}}$$

  - ... the electron wave will be reflected back on itself. Therefore, the solutions of the Schrödinger equation will be described by the possible superpositions of the two waves, and there are *two* possibilities to do that:

$$\psi^{+} \;=\; \left(\frac{1}{2L}\right)^{1/2} \cdot \left(\exp\,(ik_{\text{BZ}}x) \;+\; \exp\,(-ik_{\text{BZ}}x)\right)$$

$$\psi^{-} \;=\; \left(\frac{1}{2L}\right)^{1/2} \cdot \left(\exp\,(ik_{\text{BZ}}x) \;-\; \exp\,(-ik_{\text{BZ}}x)\right)$$

  - For the first Brillouin zone, we have $\boldsymbol{g_1 = 2\,\pi/a = k - k' = 2k_{\text{BZ}}}$, and so $\boldsymbol{k_{\text{BZ}} = \pi/a}$. Since this can be easily generalized for higher Brillouin zones, the same consequences will occur also there. To understand the basics, however, it is sufficient to consider just the edge of the first Brillouin zone.
- Both solutions are no longer propagating plane waves with $\psi \cdot \psi^* =$ **const.** throughout the crystal but *standing waves* …

$$|\psi^{+}|^2 = \frac{2}{L}\cos^2\left(\frac{\pi x}{a}\right)$$

$$|\psi^{-}|^2 = \frac{2}{L}\sin^2\left(\frac{\pi x}{a}\right)$$

  - ... with the maxima being *at* the coordinates of the atoms for the ψ^+ solution and *between* the atoms for the ψ^- solution.
  - In the first case the potential energy of the electrons is *lowered* , in the second case it is *raised* – there is an energy gap!

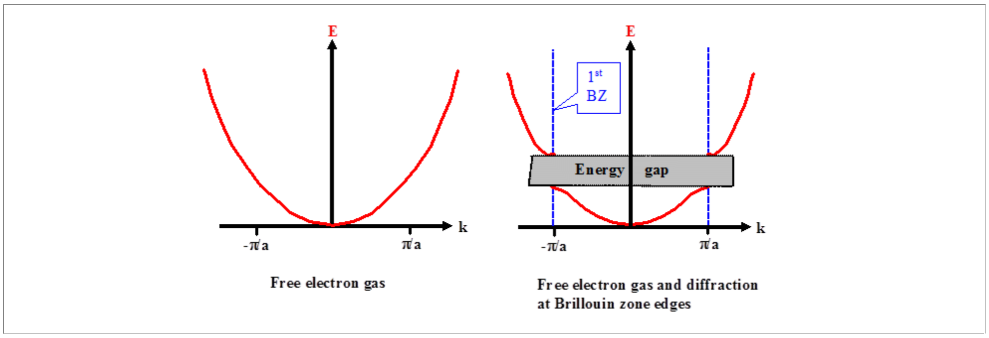
- Note that now a potential energy is involved, but only because we now implicitly assumed that the potential is no longer constant.
- While this is a relatively painless way to envision the occurrence of an energy gap, the three-dimensional case needs a few more considerations.
  - Waves with $\boldsymbol{k} \approx \boldsymbol{k}_{BZ}$, while diffracted, do not have to directly run back in themselves – after some more reflections, however, they will.
  - This leads to a splitting of the energy for *all* positions on the Brillouin zone edges; *the amount of splitting, however, may differ* .
  - A general relation yields for the energies of the $\boldsymbol{k}_{BZ}$ electron waves

$$E(\boldsymbol{k}_{BZ}) = \frac{(\hbar \boldsymbol{k}_{BZ})^2}{2m} \pm |U(\boldsymbol{g})|$$

  - With $U(\boldsymbol{g})$ = Fourier component of the periodic potential for the reciprocal lattice vector $\boldsymbol{g}$ relevant for the $\boldsymbol{k}_{BZ}$ vector considered.

Representation of Energy Gaps and Band Structures

- Bearing this in mind, we now can construct the $E(\boldsymbol{k})$ diagram in a principal way:



Free electron gas

Free electron gas and diffraction at Brillouin zone edges

  - In different directions we still would have an energy gap, but at different positions on the energy axis and with a different width. Nevertheless, this is already the first step to understand the *electronic band structure* of crystalline solids.
  - That is about as far as the free electron gas model *with diffraction added* (and therefore by necessity some unspecified periodic potential) will get us.

- For more insights we will actually have to solve the Schrödinger equation for some kind of *periodic* potential. This is difficult, even for very simple (unrealistic) periodic potentials; cf. the link.
  - For this we first need a halfway realistic potential – e.g. for a **Si** or a **GaAs** crystal – which we then use in the Schrödinger equation. The solutions will depend on the precise kind of potential and, as we must expect, they will not be easy to find (or even to express in closed form).

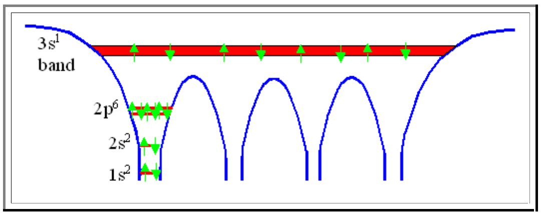
- However, since the potential is periodic, which means it doesn't matter if we look at it at ***r*** or at ***r*** + ***R*** with ***R* =** any *translation vector of the lattice* – it always looks the same – we may confidently expect that the solutions mirror somehow this property. After all, it should not matter much either, at which crystallographically equivalent crystal positions we look at the electrons.

This is exactly what the celebrated *Bloch theorem* asserts: No matter what kind of periodic potential is plucked into the Schrödinger equation, the solutions must have certain properties which can be specified in a very general way.

- We will deal with this in the next subchapter.

2.1.4 Periodic Potentials and Bloch's Theorem

- In the most simplified version of the free electron gas, the true three-dimensional potential was ignored and approximated with a constant potential (see the quantum mechanics script as well) conveniently put at **0 eV**.
  - The true potential, however, e.g. for a **Na** crystal, is periodic and looks more like this (including some electronic states):



- Semiconducting properties will not emerge without some consideration of the periodic potential – we therefore have to solve the Schrödinger equation for a suitable periodic potential. However, this is much easier said than done: There are several ways to do this (for real potentials always numerically), but they are all mathematically rather involved – and, thus, beyond the scope of this lecture.
  - Luckily, as stated before, it can be shown that *all* solutions must have certain general properties. These properties can be used to make calculations easier – as well as to obtain a general understanding of the the effects of a periodic potential on the behavior of electron waves.
  - The starting point is a potential $\boldsymbol{V(r)}$ determined by the crystal lattice that has the periodicity of the lattice, i.e.

$$\boxed{V(\boldsymbol{r}) \;=\; V(\boldsymbol{r} + \boldsymbol{T})}$$

  - With $\boldsymbol{T}$ **=** any translation vector of the lattice under consideration.
  - We then will obtain some wavefunctions $\psi(\boldsymbol{r})$ which are solutions of the Schrödinger equation for $\boldsymbol{V(r)}$. In addition, these wavefunctions have to fulfill the boundary conditions, since we are still dealing with a kind of "particle in a box" problem – the electrons are confined inside the crystal.

The Bloch Theorem

- The *Bloch theorem* in essence formulates a condition that *all* solutions $\psi(\boldsymbol{r})$, for *any* periodic potential $\boldsymbol{V(r)}$ whatsoever, have to meet. In one version it ascertains

$$\boxed{\psi(\boldsymbol{r}) \;=\; \mathrm{u}(\boldsymbol{r}) \cdot \exp\,(\mathrm{i} \cdot \boldsymbol{k} \cdot \boldsymbol{r})}$$

  - With $\boldsymbol{k}$ = any allowed wave vector for the electron that is obtained for a *constant* potential, and **u(*r*)** = some functions *with the periodicity of the lattice*, i.e.

$$\boxed{\mathrm{u}(\boldsymbol{r} + \boldsymbol{T}) \;=\; \mathrm{u}(\boldsymbol{r})}$$

- Any wavefunction meeting this requirement we will henceforth call a **Bloch wave**.
  - As before, we choose periodic boundary conditions; this ensures that no restiction on the translation vectors ***T*** needs to be considered.
- The Bloch theorem is quite remarkable, because, as said before, it imposes very special conditions on *any* solution of the Schrödinger equation, no matter what the form of the periodic potential might be.
  - We notice that, in contrast to the case of the constant potential, so far, $\boldsymbol{k}$ is just a wave vector in the plane wave part of the solution. Due to the periodic potential, however, its role as an index to the wave function is *not* the same as before – as we will shortly see.
- Bloch's theorem is a *proven* theorem with perfectly general validity. We will first give some ideas about the proof of this theorem and then discuss what it means for real crystals. As always with hindsight, Bloch's theorem can be proved in many ways; the links give some examples. Here we only look at general outlines of how to prove the theorem:
  - It follows rather directly from applying *group theory* to crystals. In this case one looks at symmetry properties that are invariant under translation.
  - It can easily be proved by working with *operator algebra* in the context of formal quantum theory mathematics (see the quantum mechanics script again).
  - It can be directly proved in *simple ways* – but then only for special cases or with not quite kosher "tricks".

- It can be proved (and used for further calculations) by expanding $\mathbf{V}(\mathbf{r})$ and $\psi(\mathbf{r})$ into a *Fourier series* and then rewriting the Schrödinger equation. This is a particularly useful way because it can also be used for obtaining specific results for the periodic potential. This proof is demonstrated in detail in the link, or in the book of Ibach and Lüth.

Bloch's theorem can also be rewritten in a somewhat different form, giving us a second version:

$$\boxed{\psi(\mathbf{r}+\mathbf{T}) = \psi(\mathbf{r})\cdot\exp(i\mathbf{k}\mathbf{T})}$$

- This means that *any* function $\psi(\mathbf{r})$ that is a solution to the Schrödinger equation of the problem, differs only by a phase factor $\exp(i\mathbf{k}\mathbf{T})$ between *equivalent positions* in the lattice.
- This implies immediately that the probability of finding an electron *is the same at any equivalent position in the lattice*, exactly as we expected, because

$$\boxed{|\psi(\mathbf{r}+\mathbf{T})|^2 = |\psi(\mathbf{r})|^2 \cdot |\exp(i\mathbf{k}\mathbf{T})|^2 = |\psi(\mathbf{r})|^2}$$

- This is so because $|\exp(i\mathbf{k}\mathbf{T})|^2 = 1$ for all $\mathbf{k}$ and $\mathbf{T}$.

That this *second version* of Bloch's theorem is equivalent to the first one may be seen as follows:

- If we write the wave function in the first form $\psi(\mathbf{r}) = u(\mathbf{r})\cdot\exp(i\mathbf{k}\mathbf{r})$ and consider its value at an equivalent lattice position $\mathbf{r} + \mathbf{T}$ we obtain

$$\psi(\mathbf{r}+\mathbf{T}) = \underbrace{u(\mathbf{r}+\mathbf{T})}_{=\,u(\mathbf{r})} \cdot \exp[i\mathbf{k}\cdot(\mathbf{r}+\mathbf{T})] = \underbrace{u(\mathbf{r})\cdot\exp(i\mathbf{k}\mathbf{r})}_{=\,\psi(\mathbf{r})} \cdot \exp(i\mathbf{k}\mathbf{T}) = \psi(\mathbf{r})\cdot\exp(i\mathbf{k}\mathbf{T}) \qquad \textit{q.e.d.}$$

Bloch's theorem has many more forms and does not only apply to electrons in periodic potentials, but for all kinds of waves in periodic structures, e.g. phonons. However, we will now consider the theorem to be proven and only discuss some of its implications.

- Most importantly, we will now clarify what index to apply to the wave function $\psi(\mathbf{r})$ and, thus, also to its lattice-periodic part $u(\mathbf{r})$.

Implications of the Bloch Theorem

One difference to the constant potential case is most crucial: If we know the wavefunction for *one* particular $\mathbf{k}$ value, we also know the wavefunctions for infinitely may other $\mathbf{k}$ values, too.

- This follows from the fact that in the "second version" of the Bloch theorem, $\psi(\mathbf{r}+\mathbf{T}) = \psi(\mathbf{r})\cdot\exp(i\mathbf{k}\mathbf{T})$, the vector $\mathbf{k}$ is not unique, since for any reciprocal lattice vector $\mathbf{g}$ it holds that $\exp(i\mathbf{g}\mathbf{T}) = 1$. Hence, when $\mathbf{k}$ is replaced by $\mathbf{k}' = \mathbf{k} + \mathbf{g}$, the Bloch theorem is also fulfilled:

$$\boxed{\psi(\mathbf{r}+\mathbf{T}) = \psi(\mathbf{r})\cdot\exp[i(\mathbf{k}+\mathbf{g})\,\mathbf{T}] = \psi(\mathbf{r})\cdot\exp(i\mathbf{k}\mathbf{T})}$$

- Obviously it is not clear which reciprocal space vector, $\mathbf{k}$ or $\mathbf{k}'$, is to be associated with this Bloch wave; this is why the index $\mathbf{k}$ was omitted so far. (That this non-uniqueness also holds for $u(\mathbf{r})$ can be seen from its Fourier series representation; cf. the relevant proof of the Bloch theorem.)
- The deeper reason behind this property is that, similar to $\mathbf{E}$ being an eigenvalue of the Hamilton operator $\mathbf{H}$ in the equation $\mathbf{H}\psi = \mathbf{E}\psi$, the exponential expressions $\exp(i\mathbf{k}\mathbf{T})$ are the eigenvalues of the *translation operator* $\mathbf{T}$ in the equation $\mathbf{T}\psi = \exp(i\mathbf{k}\mathbf{T})\psi$ (cf. the relevant proof of the Bloch theorem), and these eigenvalues do not change when $\mathbf{k}$ is replaced by $\mathbf{k}' = \mathbf{k} + \mathbf{g}$.

Therefore, in principle not a single vector $\mathbf{k}$ is to be attributed as an index to a specific wavefunction, but *all* $\mathbf{k}' = \mathbf{k} + \mathbf{g}$.

- All these points in $\mathbf{k}$-space are equivalent, because for any reciprocal lattice vector $\mathbf{g}$ it holds that $\psi_{\mathbf{k}+\mathbf{g}}(\mathbf{r}) = \psi_{\mathbf{k}}(\mathbf{r})$.
- On the other hand, this permits to restrict the $\mathbf{k}$ vector to the first Brillouin zone, because any $\mathbf{k}'$ not in the first Brillouin zone can always be written as $\mathbf{k}' = \mathbf{k} + \mathbf{g}$, with $\mathbf{k}$ now being in the first Brillouin zone.
- Keeping this in mind, we can now simply write $\psi_{\mathbf{k}}(\mathbf{r})$; further implications are discussed below.

In general, a Bloch wave $\psi_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r})\cdot\exp(i\mathbf{k}\mathbf{r})$ can be understood as a lattice-peridocally modulated plane wave.

- One way of looking at this "first version" of the Bloch theorem is to interprete the lattice-periodic function $u_{\mathbf{k}}(\mathbf{r})$ as a kind of *correction factor* that is used to generate solutions for periodic potentials from the simple solutions for constant potentials.

- We then have good reasons to assume that $\mathbf{u_k(\mathit{r})}$ for $\boldsymbol{k}$ vectors *not close to a Brillouin zone edge* will only be a minor correction, i.e. $\mathbf{u_k(\mathit{r})}$ should be close to **1**, then.

But in any case, the quantity $\boldsymbol{k}$, while still being the wave vector of the plane wave that is part of the Bloch wave function (and which may be seen as the "backbone" of the Bloch functions), has lost its simple meaning: Besides not being unique, there are explicit reasons why it can no longer be taken as a *direct* representation of the momentum $\boldsymbol{p}$ of the electron via $\boldsymbol{p} = \hbar \boldsymbol{k}$, or of its wavelength $\lambda = 2\pi/\boldsymbol{k}$:

- The momentum of the electron moving in a periodic potential, having the wave function $\psi_{\mathbf{k}}(\boldsymbol{r}) = \mathbf{u_k}(\boldsymbol{r}) \cdot \exp(\mathrm{i}\boldsymbol{kr})$, can be calculated from $-\mathrm{i}\hbar\nabla\psi_{\mathbf{k}}(\boldsymbol{r})$, thus it is not only related to $\boldsymbol{k}$ but also to $\nabla\mathbf{u_k}(\boldsymbol{r})$, which represents the influence of the lattice. As an example, for the standing waves resulting from (multiple) reflections at the Brillouin zone edges the momentum of the electron is actually *zero* (because the velocity is zero), while $\boldsymbol{k}$ is not.
- There is no unique wavelength to a plane wave modulated with some arbitrary periodic function. Its Fourier decomposition can have any spectrum of wavelengths, so which is the one to be relevant for being associated with $\boldsymbol{k}$?

To make this clear, sometimes the vector $\boldsymbol{k}$ for Bloch waves is called the "**quasi** wave vector".

Nevertheless, $\boldsymbol{k}$ is a quantum number related to the *translational symmetry of the lattice*. Thus, instead of associating it with the momentum of the electron which is related to the translational symmetry of free space, we may identify the quantity $\hbar\boldsymbol{k}$ with the so-called *crystal momentum* $\boldsymbol{P}$, being relevant for the movement of the electron in the periodic lattice.

- The crystal momentum $\boldsymbol{P}$ is something like the *combined* momentum of crystal and electron. While not being a "true" momentum (which should be expressible as the product of a distinct mass and a velocity), it still has many properties of momenta, in particular **it is conserved** during all kinds of processes (as we will see later on).
- This is a major feature for the understanding of semiconductors, as we will see soon enough!
- Only if $\boldsymbol{V} = \mathbf{0}$, i.e. if there is no periodic potential, then the crystal momentum is equal to the bare electron momentum; i.e. then the part of the crystal is zero.

Reduced- and Repeated-Zone Schemes

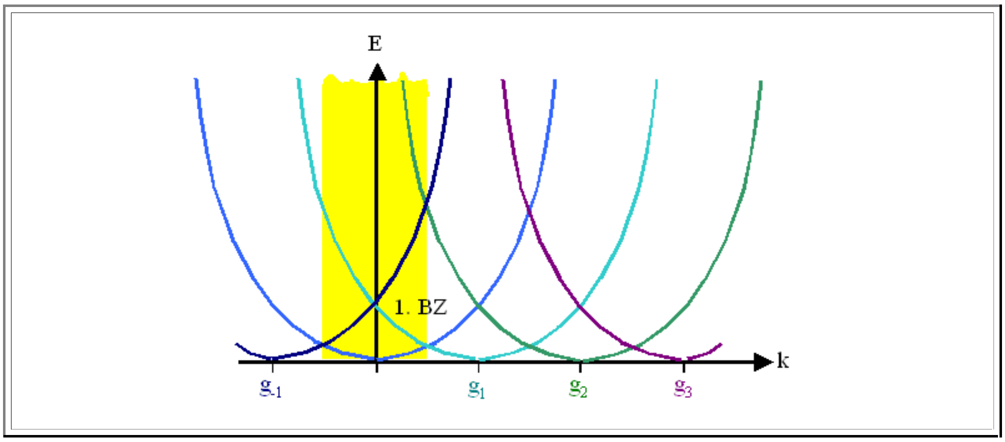
We now consider the far-reaching consequences of the fact that $\psi_{\mathbf{k+g}}(\boldsymbol{r}) = \psi_{\mathbf{k}}(\boldsymbol{r})$ for the energy value(s) associated with that wavefunction; with $\boldsymbol{g}$ = arbitrary reciprocal lattice vector.

- Writing this as $\psi(\boldsymbol{k} + \boldsymbol{g}, \boldsymbol{r}) = \psi(\boldsymbol{k}, \boldsymbol{r})$ and taking $\boldsymbol{k}$ as a continuous variable, this means that $\psi(\boldsymbol{k}, \boldsymbol{r})$ is periodic in $\boldsymbol{k}$-space.
- As a solution of the Schrödinger equation for the system, associated with $\psi(\boldsymbol{k}, \boldsymbol{r})$ there is a specific energy $\boldsymbol{E}(\boldsymbol{k})$ which necessarily is also periodic:

$$\boxed{E(k + g) = E(k)}$$

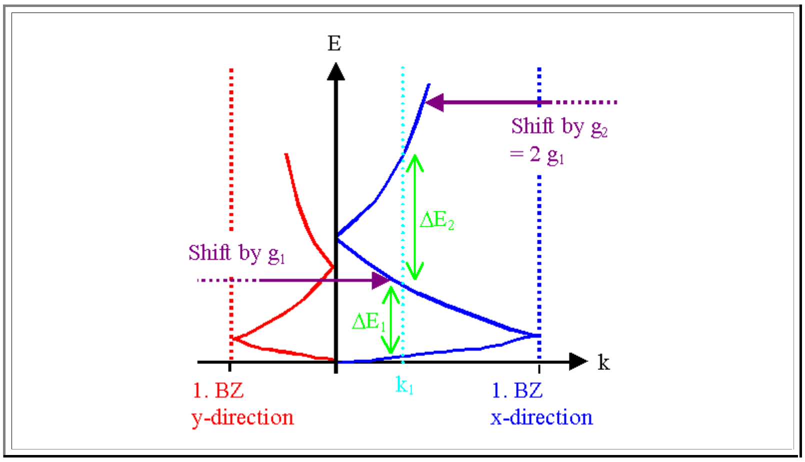
This is a major insight; it means that *any* reciprocal lattice point can serve as the origin of the $\boldsymbol{E}(\boldsymbol{k})$ function.

- Let's visualize this for the case of an infinitesimaly small periodic potential – we have the periodicity, but not a real potential. The $\boldsymbol{E}(\boldsymbol{k})$ function then is practically the same as in the case of free electrons, but starting at *every* point in reciprocal space:

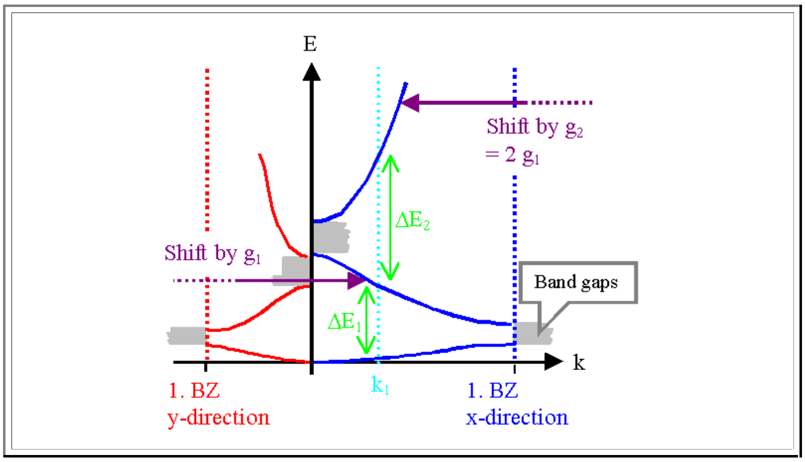


- As a consequence, we now have *many* energy values for *one* given $\boldsymbol{k}$, and in particular *all possible energy values are contained within the first Brillouin zone*, i.e. between $\mathbf{-0.5}\boldsymbol{g}_\mathbf{1}$ and $\mathbf{+0.5}\boldsymbol{g}_\mathbf{1}$ in the picture, which in total is an example of a **repeated-zone scheme** because the same holds for all Brillouin zones.

- It thus is sufficient to consider only the first Brillouin zone in graphical representations of the ***E*(*k*)** function – it contains all the information available about the system.
  - This is called a **reduced representation** of the band diagram (or **reduced-zone scheme**), which may look like this:



  - The branches outside the **1. BZ** have been shifted ("folded back") into the **1. BZ**, i.e. translated by the appropriate reciprocal lattice vector ***g***.
  - To make band diagrams like this one as comprehensive as possible, the symmetric branch on the left side is omitted; instead, the band diagram in a different direction in reciprocal space is shown.
- Altogether, this now looks like a specific electron (i.e., with a specific ***k***) could have many energies all at once – which is, of course, *not the case* .
  - Different energies, formerly distinguished by different ***k***-vectors, are still different energies, but in the first Brillouin zone they are distinguished by considering them to form different *bands* .
  - Every energy branch coming from larger ***k***-vectors carries an *index **n** denoting the band* ; this index specifies the original Brillouin zone that this branch originated from.
  - Formally, in the first Brillouin zone we have the function $E_n(\boldsymbol{k})$ associated with the Bloch wave $\psi_{n,k}(\boldsymbol{r})$.
- The identical construction, but now for the energy functions of a *periodic potential* as given before, now looks like this:



  - We now have **band gaps** – regions with unattainable energies – in all directions of the reciprocal lattice.
  - A numerical example for the Kroning–Penney Model is shown in this link.
- What does this mean for a particular electron, say one on the lowest branch of the blue diagram with the wave vector $\boldsymbol{k}_1$? It has a definite energy ***E*** associated with it.
  - But it also could have larger energies – all the values obtained for the same ***k*** but in higher branches of the band diagram.
  - For a transition to the next higher branch the energy ΔE_1 is needed. *It has to be supplied from the outside world*.
  - After the transition the electron has now a higher energy, but the wave vector is the same. *But wait*, in the reduced band diagram, we simply omitted a reciprocal wave vector, so its wave vector is actually $\boldsymbol{k}_1 + \boldsymbol{g}$. If we index the situation after the transition with "**2**", before with "**1**", we have the following equations:

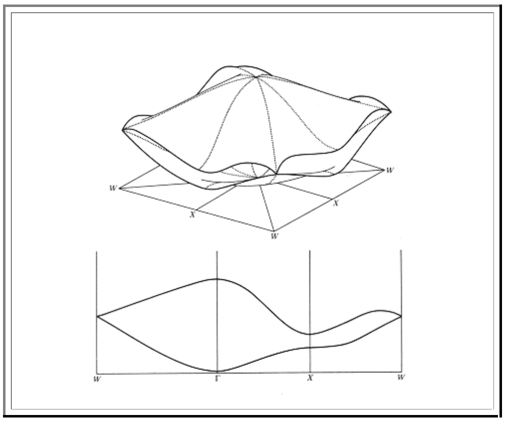
$$\begin{aligned} E_2 &= E_1 + \Delta E \\ \boldsymbol{k}_2 &= \boldsymbol{k}_1 + \boldsymbol{g} \\ |\boldsymbol{k}_1| &\neq |\boldsymbol{k}_2| \end{aligned}$$

- This is simply Braggs law, but now for *inelastic scattering*, where the magnitude of ***k*** may change – but only by a specified amount tied to a reciprocal lattice vector.
- Since we interpreted $\hbar\boldsymbol{k}$ as **crystal momentum** , we may consider *Braggs law to be the expression for the conservation of momentum in crystals*.
  - The reduced band diagram representation thus provides a very simple graphical representation of allowed transitions of electrons from one state, represented by (E_1, $\boldsymbol{k}_1$), to another state (E_2 , $\boldsymbol{k}_2$): the states must be on a vertical line through the diagram, i.e. straight up or down.
  - An alternative way of describing the states in the spirit of the reduced diagram is to use the same wave vector ***k*** for all states and a band index for the energy. The transition then goes from (E_n, ***k***) to (E_m, ***k***) with ***n***, ***m*** **=** number of the energy band involved.
- The possibility of working in a reduced band diagram, however, does not mean that wave vectors larger than all possible vectors contained in the **1. BZ** are not meaningful or *do not exist*:
  - Consider an electron "shot" into the crystal with a high energy and thus a large ***k*** – e.g. in an electron microscope. If you reduce its wave vector by subtracting a suitable large ***g*** vector without regard to its energy and band number, you may also reduce its energy – you move it, e.g., from a band with a high band number to a lower one. While this may happen physically, it will only happen via many transitions from one band to the next lower one – and this takes time!
  - Most of the time in normal applications the electron will keep its energy and its original wave vector. And it is this wave vector you must take for considering diffraction effects! An Ewald (or Brillouin) construction for diffraction will give totally wrong results for reduced wave vectors – think about it!

- If you feel slightly (or muchly) confused at this point, that is as it should be. Bloch's theorem, while relatively straightforward mathematically, is not easy to grasp in its implications to real electrons. The representation of the energy–wave vector relationship (the dispersion curves) in extended or reduced schemata, the somewhat unclear role of the wave vector itself, the relation to diffraction via Bragg's law, the connection to electrons introduced from the outside, e.g. by an electron microscope (think about it for minute!), and so on, are difficult concepts not easily understood at the "gut level".
  - While it never hurts to think about these questions, it is sufficient for our purpose to just accept the reduced band structure scheme and its implications as something useful in dealing with semiconductors – never mind the "small print" associated with it.
  - However, if you want to dig deeper: All these effects that are difficult to grasp are to some extent rooted in the formal quantum mechanics behind Bloch's theorem. It has to do with eigenvectors, eigenvalues and commutation of operators; so, if you know about that, all these effects come rather naturally.

2.1.5 Band Structures and Standard Representations

- Real crystals are three-dimensional and we must consider their band structure in three dimensions, too.
    - Of course, we must consider the reciprocal lattice, and, as always if we look at electronic properties, use the Wigner-Seitz cell (identical to the **1st** Brillouin zone) as the unit cell.
    - There is no way to express quantities that change as a function of three coordinates graphically, so we look at a two dimensional crystal first (which, incidentally, do exist in semiconductor physics).
- The qualitative recipe for obtaining the band structure of a two-dimensional lattice using the slightly adjusted parabolas of the free electron gas model is simple:
    - Construct the parabolas along major directions of the reciprocal lattice, interpolate in between, and fold them back into the first Brillouin zone. How this can be done for the free electron gas is shown in an illustration module.
    - An example – taken from "Harrison" – may look like this:



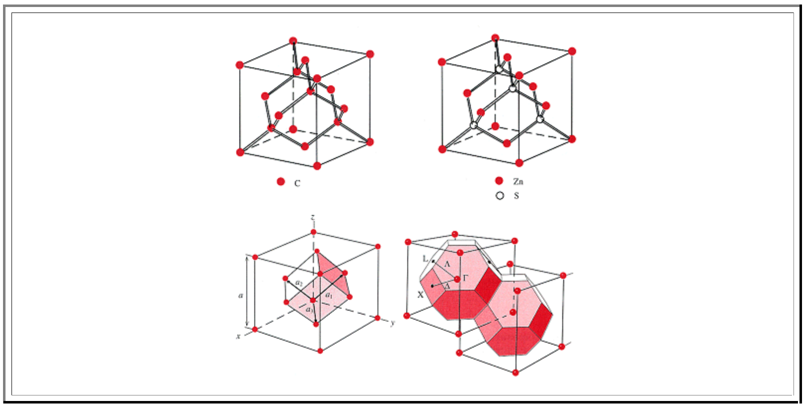
-
    - The lower part (the "cup") is contained in the **1st** Brillouin zone, the upper part (the "top") comes from the second **BZ**, but is now folded back into the first one. It thus would carry a different band index. This could be continued ad infinitum; but Brillouin zones with energies well above the Fermi energy are of no real interest.
    - The lower part shows tracings along major directions. Evidently, they contain most of the relevant information in condensed form. It is clear, e.g., that this structure has no band gap.
- It would be sufficient for most purposes to know the **E$_n$(*k*)** curves – the dispersion relations – along the major directions of the reciprocal lattice (***n*** is the band index) (see quantum mechanics script as well).
    - This is exactly what is done when real band diagrams of crystals are shown. Directions are chosen that lead from the center of the Wigner-Seitz unit cell – or the Brillouin zones in the more generalized picture – to special symmetry points. These points are labeled according to the following rules:
        - Points (and lines) inside the Brillouin zone are denoted with *Greek* letters.
        - Points on the surface of the Brillouin zone with *Roman* letters.
        - The center of the Wigner-Seitz cell is always denoted by a **Γ**
    - For cubic reciprocal lattices, the points with a high symmetry on the Wigner-Seitz cell are the intersections of the Wigner Seitz cell with the low-indexed directions in the cubic elementary cell.
    - We use the following nomenclature: (red for **fcc**, blue for **bcc**):

The intersection point with the **[100]** direction is called **X** (**H**); the line **Γ—X** is called Δ.

The intersection point with the **[110]** direction is called **K** (**N**); the line **Γ—K** is called Σ.

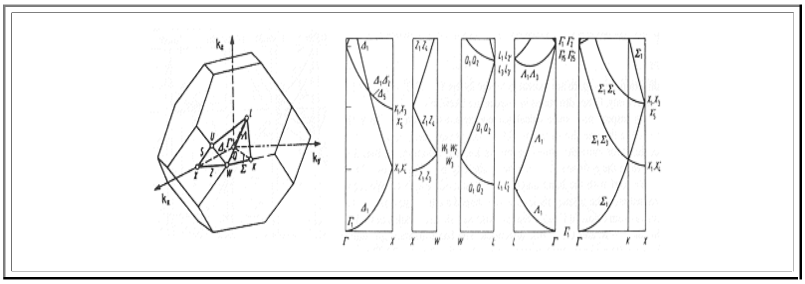
The intersection point with the **[111]** direction is called **L** (**P**); the line **Γ—L** is called Λ .

- The picture above already used this kind of labelling. Since the tracing of the dispersion curve can be done on different levels – corresponding to the **1st**, second, etc. Brillouin zone – the points are often indexed with the number of the Brillouin zone they use.
    - This may look like this:



- The top pictures show the elementary cell of the diamond lattice or of the ZnS type lattice; the lower left picture the Bravais lattice of the **fcc** type and the primitive (non-cubic) lattice which is an equally valid, if less symmetric, representation of the **fcc** lattice..
- The lower right picture shows the cubic reciprocal lattice of the cubic **fcc** lattice (which is a **bcc** lattice) and the Wigner-Seitz cells (identical with the first Brillouin zone) which also represent the reciprocal lattice

We now can draw the band diagrams along all kinds of lines – not only from **Γ** to some point on the Brillouin zone, but also from point to point, e.g., from **L** to **K** or to some other points not yet labeled. An example for the **fcc** structure and the *free electron gas approximation* is shown below.



- The first Brillouin zone with the proper indexing of the relevant points and some dispersion parabola along prominent directions are shown. The picture is taken from Hummel's book.
- The indexing of the various branches is a bit more complicated than in the illustration example for reasons explained below.

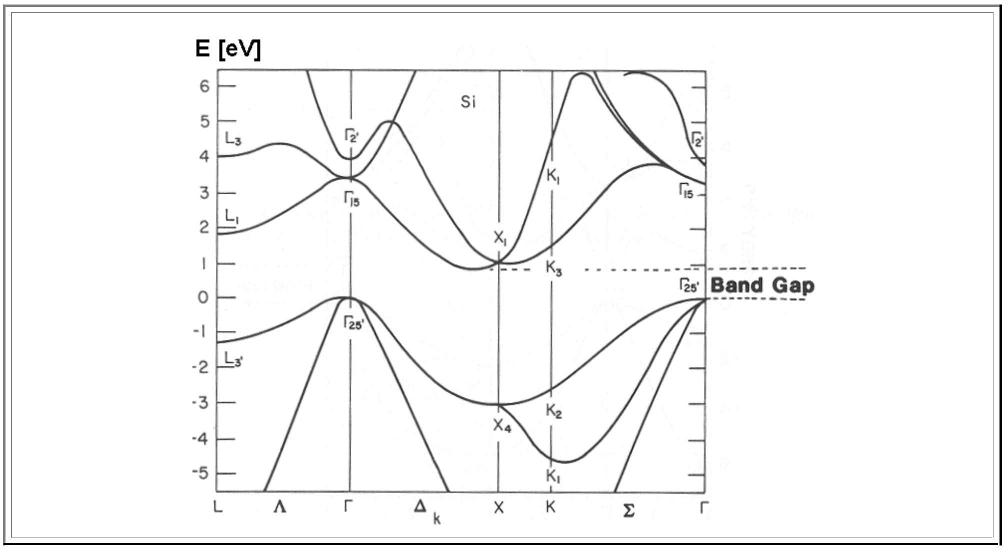
Contemplate this picture a bit and then ask yourself:

  - Do I find this picture alarming ? ("Gee, if even the most simple situation produces such a complicated structure, I'm never going to understand it)
  - Do I find this picture exciting? ("Gee, what a wealth of information one can get in a simple diagram if you pick a smart way of representation").

- Yes, it is a bit confusing at first. But do not despair: If you need it, if you work with it, you will quickly catch on!

It is standard praxis to join the single diagram at appropriate points and to draw band diagrams by showing two branches starting from **Γ** to major points and to continue from there as already practiced above.

- The band diagram of **Si**, e.g., then assumes its standard form:



  - The indexing of the major points in the Brillouin zone is more complex than described so far – it is more than just a band index. This reflects the fact that there is no unique choice of the **Γ** point, or that the the band structure allows certain symmetry operations without changing. The indexing follows rules of group theory displaying the symmetries, but shall not be described here.

- The band structure as shown in this standard diagram contains a tremendous amount of information; at this level it is, e.g., evident that:
  - **Si** has a band gap of about **1 eV**.
  - **Si** is an indirect semiconductor because the maximum of the valence band (at **Γ**) does not coincide with the minimum of the conduction band (to the left of **X**).

- There is, however, a lot more information encoded in this diagram, as we will see later.

- Of course, the question remains how the band structures of real-world materials can be obtained. However, concerning both measurementes and calculations, this is a rather involved subject of its own; we do not treat it here.
  - To get at least a coarse feeling about how involved such calculations are, you may have a look at some really advanced module where the theoretical basics and the available software are presented.

2.2 Basic Semiconductor Physics

2.2.1 Intrinsic Properties in Equilibrium

In this subchapter we deal with basic semiconductor properties and simple devices like **p-n** junctions on a somewhat simplified, but easy to understand base. It shall serve to give a good basic understanding, if not "gut feeling" to what happens in semiconductors, leaving more involved formal theory for later.

- However: Intrinsic semiconductors are theoretical concepts, requiring an absolutely perfect infinite crystal. Finite crystals with some imperfections may have properties that are widely different from their intrinsic properties.
- As a general rule of thumb: If you cannot come up with a material that is at least remotely similar to what it should be in its "intrinsic" state, it is mostly useless because then you cannot manipulate its properties by doping.
- That is the major reason, why we utilize so few semiconductors – essentially **Si**, **GaAs**, **GaP**, **InP**, **GaN**, **SiC** and their relatives – and tend to forget that there is a large number of "intrinsically" semiconducting materials out there. For a short list activate the (German) link.

Silicon crystals are pretty good and thus are closest to truly intrinsic properties. But even with the best **Si**, we are not really close to intrinsic properties, see exercise **3.1-1** for that. Nevertheless: *This chapter always refers to silicon, if not otherwise stated!*

- A few very basic aspects about semiconductors, including some specific expressions and graphical representations, will be taken for granted; in case of doubt refer to the link with an **alphabetical list of basic semiconductor terms**.

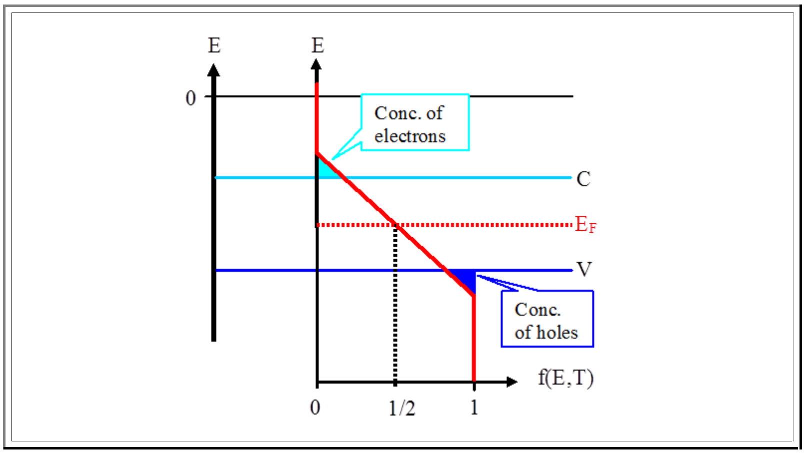
Fermi Energy and Carrier Density

In this first section we review the properties of **intrinsic semiconductors** . We make two simplifying assumptions at the beginning (explaining later in more detail what they imply):

- The semiconductor is "**perfect** ", i.e. it contains no crystal defects whatsoever.
- The effective density of states in the conduction and valence band, the mass, mobility, lifetime, and so on of electrons and holes are identical. (See below for any detail about these quantities.)

All we need to know for a start then is the magnitude of the band gap ***E*$_g$**. The Fermi energy then is exactly in the middle of the forbidden band; we can deduce that as follows:

- Namey, by just looking at a drawing schematically showing the density of electrons in the valence and conduction band where, for ease of drawing, the Fermi distribution is shown with straight lines instead of the actual curved shape.



- Note that in the standard literature (especially in the English language scientific literature), typically one doesn't sharply distinguish between *carrier density* and *carrier concentration*. If in doubt, look for the unit of measurement relevant in the given equation.

The density of electrons, n_e, in the conduction band is given exactly by

$$n_e = \int_{E_C}^{E'} D(E) \cdot f(E,T) \cdot dE$$

- With ***E* ' =** energy of the upper band edge.

- With the usual approximations:
  - Boltzmann distribution instead of Fermi distribution.
  - Substitution of an **effective density of states**, N_{eff} at the band edge instead of the true, energy-dependent density.
  - Integration from the lower band edge E_C to infinity

  we obtain

$$\boxed{n_e = N_{\text{eff}}{}^e \cdot \exp\left(-\frac{E_C - E_F}{kT}\right)}$$

- The light blue triangle in the picture symbolizes this density!

$N_{\text{eff}}{}^e$ (with the factor two for spin up/spin down included) can be estimated from the free electron gas model in a fair approximation to

$$\boxed{N_{\text{eff}}{}^e = 2\left(\frac{2\,\pi m kT}{h^2}\right)^{3/2}}$$

- How this is done and how some numbers can be generated from this formula (look at the dimensions in the formula above and start wondering) can be found in the link.

In an intrinsic semiconductor in thermal equilibrium, all electrons in the conduction band come from the valence band. The density of holes in the valence band, n_h , thus must be exactly equal to the density of electrons in the conduction band, or

$$\boxed{n_e = n_h = n_i = \text{intrinsic density}}$$

- The dark blue triangle in the picture then symbolizes the hole density.
- **Important:** This is how holes are *defined* , and for good reasons; as we will see (rather soon), **only** the empty *valence band* states can reasonably be considered as being occupied by holes (= *mobile* positive charge carriers).
- Given the assumptions made above and the symmetry of the Fermi distribution, the unavoidable conclusion is that the Fermi energy is exactly in the middle of the band gap.

Carrier Density and Conductivity

The carrier densities are decisive for the conductivity (or resistivity) of the material. If you are not familiar (or forgot) about conductivity, mobility, resistivity, and so on and how they connect to the average properties of an electron gas in thermal equilibrium, go through the following basic modules:

- Ohm's Law and Materials Properties
- Ohm's Law and Classical Physics

We thus have the density of mobile carriers in both bands and from that we can calculate the **conductivity** σ via the standard formula

$$\boxed{\sigma = e \cdot (\mu_e \cdot n_e + \mu_h \cdot n_h)}$$

- provided we know the **mobilities** μ of the electrons and holes, μ_e and μ_h, respectively.

Again, simplifying as much as sensibly possible, with $\mu_e = \mu_h = \mu$ we obtain

$$\boxed{\sigma = 2e\mu \cdot N_{\text{eff}}{}^e \cdot \exp\left(-\frac{E_C - E_F}{kT}\right) = 2e\mu \cdot N_{\text{eff}}{}^e \cdot \exp\left(-\frac{E_g}{2kT}\right)}$$

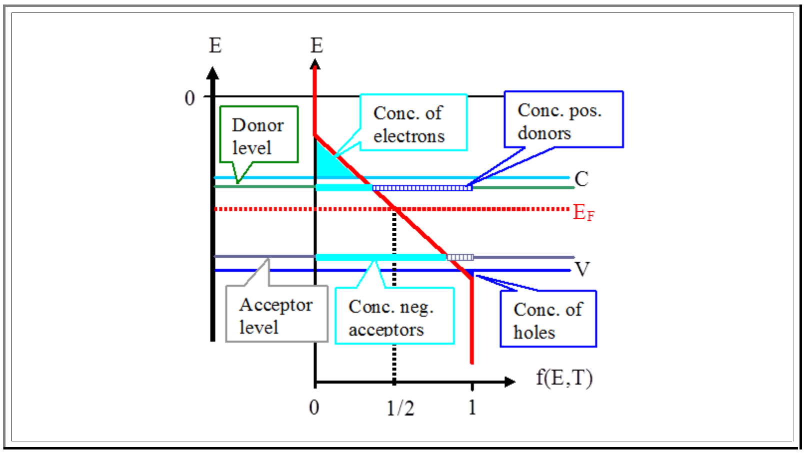
- because we have $E_C - E_F = E_g/2$ (with the fundamental band gap energy E_g) for the intrinsic case as discussed so far.

- This gives us already a good idea about the comparable magnitudes and especially the temperature dependences of semiconductors, because the exponential term overrides the pre-exponential factor which, moreover, we may expect not to be too different for *perfect intrinsic* semiconductors of various kinds.

2.2.2 Doping, Carrier Density, Mobility, and Conductivity

Again, we look at a "perfect" semiconductor, where **doping** has been achieved by replacing some lattice atoms by suitable doping atoms without – in the ideal world – changing anything else.

- We have now additional allowed energy levels in the band gap, belonging to the doping atoms, i.e. they are localized states.
- These levels may or may not be occupied by electrons; the Fermi distribution will give the probability for occupancy as before. In analogy to the intrinsic case, we now have the following highly stylized picture:



- As in the intrinsic case, in this drawing no distinction was made between *carrier density* (in cm^{-3}) and *carrier concentration* (in ppm, ppb, or the like).

Charge Neutrality Condition

While the picture may look complicated, it is actually just a means to keep track of the number of charges in the semiconductor. Since there is an over-all charge neutrality, we have

$$\boxed{\Sigma \text{ pos. charges} = \Sigma \text{ neg. charges}}$$

- This equation can be used to calculate the exact position of the Fermi energy (as we will see). Since almost everything else follows from the Fermi energy, let's look at this in some detail.

But wait – why are there *positive* charges at all? Aren't we just discussing the occupation of *electronic* levels/states? After all, electrons are negatively charged. And by which rule is it that some states in the above figure count the one way (negative), but some the other way (positive)?

- This rule is simple: Everything counts relative to the **individual** behavior at zero kelvin (*) where the conduction band is completely empty, the valence band is completely filled, all donor states are occupied, and all acceptor states are empty.
- Of course, if both donor and acceptor states were present together in a material, then at zero kelvin the electrons would not be distributed like that. And that's why it is emphasized that one has to look at the **individual** behavior of these states at zero kelvin, because charge neutrality is **defined** relative to the latter (for which it obviously holds).
- (* Yes: The units of measurement are spelled with small letters – because it's only the natural persons' names that are spelled with capital letters.)

First, let's count the *negative* charges which, following the above rule, means to look for those places where at zero kelvin *no* electrons are to be found.

- There are, ***first**, the electrons in the conduction band*. Their density, as spelled out before, was

$$\boxed{n_e = N_{\text{eff}}^{e} \cdot \exp\left(-\frac{E_C - E_F}{kT}\right)}$$

- More generally and more correctly, however, we have

$$\boxed{n_e = N_{\text{eff}}^{e} \cdot f(E_C, E_F, T)}$$

- The Fermi energy E_F is now included as a *variable* in the Fermi function, because the density of electrons depends on its precise value which we do not yet know. In this formulation the electron density comes out *always correct*, no matter where the Fermi energy is positioned in the band gap.
  - Then there are, ***second**, the negatively charged acceptor atoms* . Their density N_A^- is given by the density of acceptor states (which is just the *doping density* N_A) *times the probability* that the states are occupied, and that is given by the value of the Fermi distribution at the energy of the acceptor state, $\mathbf{f}(E_A, E_F, T)$. We have accordingly

$$\boxed{N_A^- = N_A \cdot \mathrm{f}(E_A, E_F, T)}$$

  - *Note:* The Fermi distribution in this case should be slightly modified to be totally correct. The difference to the straight-forward formulation chosen here is slight, however, so we will keep it in this simple way. More about thisin an advanced module.
- Now let's count the *positive* charges which, following the above rule, means to look for those places that at zero kelvin are *completely filled* with electrons.
  - ***First*** , we have the *holes* in the valence band. Their number is given by the number of electrons that do *not* occupy states in the valence band; in other words we have to multiply the effective density of states with the probability that the state is *not* occupied.
  - The probability that a state is *not* occupied is just **1** *minus* the probability that it is occupied, or simply $\mathbf{1 - f}(E, E_F, T)$. This gives the density of holes to

$$\boxed{n_h = N_{eff}{}^h \cdot \left(1 - \mathrm{f}(E_V, E_F, T)\right)}$$

  - ***Second*** , we have the *positively charged donors*, i.e. the donor atoms that lost an electron. Their density N_D^+ is equal to the density of states at the donor level (which is again identical to the density of the donors themselves) times the probability that the level is *not* occupied; so, we have

$$\boxed{N_D^+ = N_D \left(1 - \mathrm{f}(E_D, E_F, T)\right)}$$

- Charge neutrality thus demands

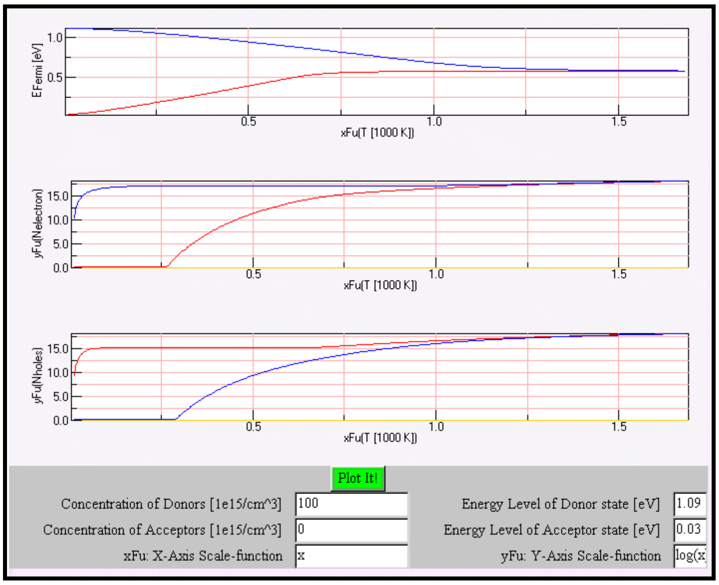
$$\boxed{N_{eff}{}^e \cdot \mathrm{f}(E_C, E_F, T) + N_A \cdot \mathrm{f}(E_A, E_F, T) = N_{eff}{}^h \cdot \left(1 - \mathrm{f}(E_V, E_F, T)\right) + N_D \cdot \left(1 - \mathrm{f}(E_D, E_F, T)\right)}$$

  - If we insert the expression for the Fermi distribution

$$\boxed{\mathrm{f}(E_n, E_F, T) = \frac{1}{\exp\left(\dfrac{E_n - E_F}{kT}\right) + 1}}$$

  - where E_n stands for $E_{C,V,D,A}$, we have, for a given material with a given doping, *one* equation for the *one* unknown quantity E_F !
- Solving this equation for any given semiconductor and any density of (ideal) donors and acceptors will not only give us the exact value of the Fermi energy E_F for any temperature T, i.e. $E_F(T)$, but *all the carrier densities* as specified above.
  - Unfortunately, this is a messy transcendental equation – it has no direct solution that we can write down.
  - Before the advent of cheap computers, this was a problem – you had to do case studies and use approximations: High or low temperatures, only donors, only acceptors and so on. This is still very useful, because it helps to understand the essentials.
- However, here we could use a program (written by **J. Carstensen**) that solves the transcendental equation and provides all functions and numbers required.
  - Unfortunately, the relevant JAVA applet doesn't work anymore, so the link to this illustration module is deactivated.
- Therefore, here we show only a screen shot with the result for the typical case of **Si** with

- **1.** An *acceptor* density of **10^{15} cm^{-3}** (red line).
- **2.** A *donor* density of **10^{17} cm^{-3}** (blue line).



Concentration of Donors [1e15/cm^3] 100

Concentration of Acceptors [1e15/cm^3] 0

xFu: X-Axis Scale-function x

Energy Level of Donor state [eV] 1.09

Energy Level of Acceptor state [eV] 0.03

yFu: Y-Axis Scale-function log(x

One of many important points to note about carrier densities is the simple, but technologically supremely important fact that the majority carrier density for many semiconductors in a technically useful temperature interval is practically identical to the dopant density.

- This is then the temperature regime were we can hope to have only a weak temperature dependence of electronic properties and thus of devices made from **Si**.

The equations used for charge neutrality also make possible to deduce an extremely important relation, the **mass action law** (for electrons and holes), as follows:

- First we consider the product

$$n_e \cdot n_h = N_{eff}^{e} \cdot f(E_C, E_F, T) \cdot N_{eff}^{h} \cdot [1 - f(E_V, E_F, T)]$$

- Then we insert the formula for the Fermi distribution and note that

$$1 - f(E_V, T) = 1 - \frac{1}{\exp\left(\frac{E_V - E_F}{kT}\right) + 1} = \frac{1}{1 + \exp\left(\frac{E_F - E_V}{kT}\right)}$$

- Using the Boltzmann approximation for the Fermi distribution, we get the famous and very important mass action law (try it yourself!):

$$n_e \cdot n_h = N_{eff}^{e} \cdot N_{eff}^{h} \cdot \exp\left(-\frac{E_C - E_V}{kT}\right) = n_i^{\,2}$$

We are now in a position to calculate the density of majority and minority carriers with very good precision if we use the complete formula, and with a sufficient precision for the appropriate temperature range where we can use the following very simple relations:

$$n_{\text{maj}} = N_{\text{dop}}$$

$$n_{\text{min}} = \frac{n_i^2}{n_{\text{maj}}} = \frac{n_i^2}{N_{\text{dop}}}$$

- We know that the conductivity σ of the semiconductor is given by

$$\sigma = e \cdot (n_e \cdot \mu_e + n_h \cdot \mu_h)$$

  - With **μ =** mobility of the carriers.

- We now have the densities; obviously we now have to consider the *mobility* of the carriers.

Mobility

- Finding simple relations for the mobility of the carriers is *just not possible*. Calculating mobilities from basic material properties is a far-fetched task, much more complicated and involved than the carrier density business.
  - However, the carrier densities (and their redistribution in contacts and electrical fields) is far more important for a basic understanding of semiconductors and devices than the carrier mobility.
  - At this point we will therefore only give a cursory view of the essentials relating to the mobility of carriers.
- The mobility μ of a carrier in an operational sense is defined as the proportionality constant between the average **drift velocity v_D** of an ensemble of carriers in the presence of an electrical field E:

$$v_D = \mu \cdot E$$

  - The average (absolute) velocity v of a carrier and its drift velocity v_D must not be confused; for a detailed discussion consult the links to two basic modules:
    - Ohm's Law and Materials Properties
    - Ohm's Law and Classical Physics
  - The simple linear relationship between the drift velocity and the electrical field as a driving force is pretty universal – it is the requirement for ohmic behavior (look it up in the link) – but not always obeyed. In particular, the drift velocity may *saturate* at high field strengths, i.e. increasing the field strength does not increase v_D anymore. We come back to that later.
- Here we only want to get a feeling for the order of magnitudes of the mobilities and the major factors determining these numbers.
  - As we (should) know, the prime factor influencing mobility is the average time between scattering processes. In fact, the mobility **μ** may be written as

$$\mu = \frac{e \cdot \tau_s}{m}$$

  - With τ_s **=** mean scattering time. We thus have to look at the major scattering processes in semiconductors.
- There are three important mechanisms:
- The *first* (and least important one) is *scattering at crystal defects* like dislocations or (unwanted) impurity atoms.
  - Since we consider only "perfect" semiconductors at this point, and since most economically important semiconductors are pretty perfect in this respect, we do not have to look into this mechanism here.
  - However, we have to keep an open mind because semiconductors with a high density of lattice defects are coming into their own (e.g. **GaN** or **$CuInSe_2$**) and we should be aware that the mobilities in these semiconductors might be impaired by these defects.
- *Second*, we have the *scattering at wanted impurity atoms*, in other word at the (charged) *dopant atoms*.

  - This is a major scattering process which leads to decreasing mobilities with increasing doping density. The relation, however, is non-linear and the influence is most pronounced for larger doping densities, say beyond **10^{17} cm^{-3}** for **Si**.
  - Examples for the relation between doping and mobilities can be found in the illustration.
  - As a rule of thumb for **Si**, increasing the doping level by **3** orders of magnitude starting at about about **10^{15} cm^{-3}** will decrease the mobility by one order of magnitude, so the change in conductivity will be about only two orders of magnitude instead of three if only the carrier density would change.
  - The scattering at charged dopant atoms *decreases* with increasing temperature. (This results mainly from the scattering being Rutherford scattering via Coulomb interaction, for which the scattering cross section reduces with increasing temperature.)
- *Third*, we have *scattering at phonons* – the other important process. **Phonons** are the (quantum-mechanical) "particles" corresponding to the thermally stimulated lattice vibrations and thus strongly depend on temperature.
  - This part scales with the density of phonons, i.e. it *increases* with increasing temperature (with about $T^{3/2}$). It is thus not surprising that it dominates at high temperatures (while scattering at dopant atoms may dominate at low temperatures).
- Scattering at phonons and dopant atoms together essentially dominate the mobilities.
  - The different and opposing temperature dependencies almost cancel each other to a certain extent for medium to high doping levels (see the illustration), again a very beneficial feature for technical applications where one doesn't want strongly temperature dependent device properties.

Conductivity

- With a relatively simple law for the carrier densities and no direct equations for the mobility, but relatively simple behavior of the mobility, conductivity data can now be compiled.
  - Some data giving the relationship between doping and resisitivity (**=1/**conductivity) or the temperature dependence of the conductivity are shown in the illustration.
- The illustrations show different behavior for **n-** and **p**-type **Si**, which could not be predicted in our present simplified treatment; it is due to different mobilities of the different carriers.
  - Our present assumptions of equal parameters (mass, mobility, life time, and so on) for electrons and holes, while justifiable in a **1st** degree approximation, are too simple. Specific differences between holes and electrons exist and will be dealt with in subsequent chapters. They are responsible for the differences in mobility, conductivity, and so on between holes and electrons.

2.2.3 Lifetime and Diffusion Length

- So far we looked at (perfect) semiconductors in perfect equilibrium. The density of holes and electrons was given by the type of the semiconductor (as signified by the band gap), the doping, and the temperature.
  - The only other property of interest (introduced in passing) was the carrier mobility – a quantity that is determined in a comparatively complex way by properties of the semiconducting materials in question.
  - The carrier density so far was constant and did not change in time – we perceived the system statically.

Concept of Life Time

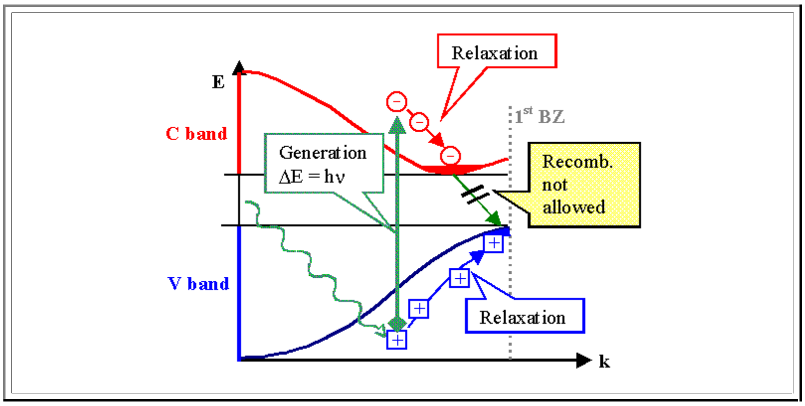
- In reality, however, we have a **dynamic equilibrium**: Electron–hole pairs are generated all the time and they recombine all the time, too – but their average density in equilibrium stays constant.
  - The easiest way to include this dynamic equilibrium in the formal representation of semiconductors is to introduce the concept of the **minority carrier life time**, or life time in short.
  - Here we look at it in a extremely simplified way – the idea is to just get the fundamentals right. In the next chapter we will delve a little deeper into the subject.
- Let's consider a **p**-doped semiconductor. The majority carriers then are holes in the valence band; in the appropriate temperature range, their density is essentially given by the density of the acceptors put into the material.
  - The electrons in the conduction band are the minority carriers, their density n_e is given via the mass action law ($n_h \cdot n_e = n_i^2$) by equating the majority carrier density with the density of the doping atoms:

$$\boxed{n_e = n_{min} = \frac{n_i^2}{n_h} = \frac{n_i^2}{N_A}}$$

- Now consider some light impinging on the material with an energy larger than the band gap, so it will be absorbed by generating electron–hole pairs. Let's assume that $G_{h\nu}$ electron hole pairs will be generated every second (the index "**h**ν" refers to the photons via their energy to distinguish this **generation rate** from others yet to come).
  - The density of both electrons and holes will now increase and no longer reflect the equilibrium values.
  - The deviation from equilibrium is much more pronounced for the minority carriers (which here are the electrons). If, for example, the density of the electrons is **0.1 %** of the hole density; an increase of the hole density of **0.1 %** due to the light-generated holes would increase the electron population by **100 %**.
  - Consequently, in non-equilibrium conditions, we are mostly interested in what happens to the minority carriers – the majority carriers then will be automatically taken care of, too, as we will see.
- Evidently, the density of minority carriers cannot grow indefinitely while light shines on the semiconductor. Some of the excess electrons will disappear again due to recombination with holes.
  - If the **recombination rate** is proportional to the density of the minorities, an equilibrium will be reached eventually, where the additional rate of recombination equals the generation rate $G_{h\nu}$, and the density will then be constant again at some higher level.
  - If we now were to turn off the light, the minority carrier density will decrease (in the usual exponential fashion) to its thermal equilibrium value. The average time needed for a decrease of **1/e** is the **minority carrier life time** τ.
- This is a simple but nonetheless correct way to think about life times and we are now using this concept to arrive at a few more major properties and relations.

Direct and Indirect Semiconductors

- First, let's consider the rough magnitude of the life time τ. A recombination process implies that the (quasi) wave vectors of both, electron and hole, change during this process – simply because the particles annihilate each other, so also the (quasi) wave vectors get annihilated in the end.
  - Any change in (quasi) wave vector is always subjected to the requirement of conservation of the (crystal) momentum. In the simplest case of a band–band transition, this requires a vertical line in the reduced band diagram.
  - This is only possible if there are occupied electron and hole states directly on top of each other. Let's visualize this by looking at the electron–hole pair generation process by a photon *more closely*:



- Shown is the band diagram for an indirect semiconductor, i.e. the minimum of the conduction band is not directly over the maximum of the valence band.
- A photon with the energy **hν** lifts an electron from somewhere in the valence band to a position directly overhead in the conduction band, giving all its energy to the electron. That the new state of the electron lies "directly overhead" is due to the fact that the momentum of the photon is neglibly small.
- Instant exercise: Show that this is true by estimating the magnitude of the photon's wavevelength, λ, for a photon energy of 1 eV. Why is it sufficient to know this value? (Hint: We consider band–band transitions within the first Brillouin zone.)
- A hole and an electron deep in the valence band or conduction band, respectively, are created which immediately (within picoseconds), by releasing phonons, give up their surplus energy relative to the maximum of the valence band or minimum of the conduction band, respectively, and thus come to an energetic rest at the extrema of the bands.
- Their wave vectors now are different – direct recombination is not allowed; it would violate Braggs (generalized) law or the conservation of crystal momentum. Recombination needs a third partner (e.g. lattice defects), and life times will be *large* (typically microseconds if not milliseconds) and depends somehow on the density of suitable third partners.
- Did you look closely? Yes? Closely enough to notice that this picture contains a very basic mistake (that doesn't influence what has been discussed, however)? Good! If you didn't notice, you may want to activate the link.

For direct semiconductors the diagram would be very similar, except that the extrema of the bands are now on top of each other.

- Recombination now is easily possible, the energy liberated will be in the form of a photon – *recombination thus produces light*.
- Life times in the case of direct semiconductors thus tend to be *short* – typically nanoseconds – and are dominated by the properties of the semiconductor itself.

What exactly determines the life time in indirect semiconductors like **Si**? In our simplified view of "perfect" crystals, would recombination simply be impossible?

- There are several mechanisms that make recombination possible in real crystals. Generally, a third partner is needed to provide momentum conservation (and is thus also involved in the energy transfer).
- Usually, this third partner is a defect of some kind. Most notorious are certain atomic defects, often interstitial atoms as, e.g., **Fe**, **Ni**, **Cu**, **Au**, and many others. But the doping atoms and coarser defects like dislocations and grain boundaries also help recombination along. Last, but also least, it is also possible that phonons get involved.

In summary, the life time of indirect semiconductors is dominated by defects, by impurities, by anything that makes the crystal imperfect. It is thus a property that can vary over many orders of magnitude – some examples can be found in the link. In very perfect indirect semiconductors, however, it is a very large time (for electrons), and can easily be found in the millisecond range.

Generalization

In an important generalization of what has been said so far, we realize that a minority carrier does not "know" how it was generated. Generation by a thermal energy fluctuation in thermal equilibrium or generation by a photon in non-equilibrium – it's all the same!

- The minority carrier, once it was generated, will recombine (on average) after the life time τ. This is valid for all minority carriers (at least as long as their density is not too far off the the equilibrium value).
- This implies that the minority carriers will disappear within fractions of a second after they were generated!

However, since we have a constant density of minority carriers in thermal equilibrium, we are forced to introduce a **generation rate *G*** for them that is exactly identical to their **recombination rate *R*** in equilibrium.

- In other words, the carrier densities in the valence and conduction bands are not in static, but in **dynamic equilibrium** (see the thermodynamics script as well).

- Their (average) density stays constant as long as ***G = R***. Think of your bank account. Its average balance will be constant as long as the withdrawal rate is equal to the deposition rate.

From the values of the densities we can not make any statement as to the recombination and generation rate – your bank balance stays constant if you withdraw and deposit **1 $** a week or **1 million $**!

- If we know the life time τ, however, we can immediately write down the recombination rate:
- ***R*** is the number of recombinations per second (and per **cm**^{3}), i.e. $R = n_{\min}/\tau$ with $n_{\min}$ = density (or number; as always we use these quantities synonymously even so this is not strictly correct) of minority carriers.
- With the relations from above we obtain the following expression:

$$\boxed{G = R = \frac{n_i^2}{\tau \cdot N_{\text{Dop}}}}$$

With N_{Dop} = density of the doping atoms.

This equation is a good approximation for the temperature range where the doping atoms are all ionized but little band–band transitions take place, i.e., in the temperature range where the semiconductor is usually used. And do not forget: it is only valid in *thermodynamic equilibrium*.

Diffusion Length

Minority (and majority) carriers are not sitting still in the lattice (they are only "sitting still" in ***k***-space!), but move around with some average velocity.

- In the free electron gas model this velocity could be directly calculated from the momentum $mv = \hbar \cdot k$, so $v = \hbar k/m$.
- But this is the **phase velocity** and this relation does not apply to the holes and electrons at the edges of the valence and conduction band, i.e. in the region of the dispersion relation where the differences between the free electron gas model and the real band structure is most pronounced and the **group velocity** can be very low if not zero.
- For the purpose of this subchapter let's simply assume that the carriers move randomly through the lattice (being scattered all the time) and that this movement can just as well be described by a **diffusion coefficient *D*** as always.
- We already had a glimpse that there is a connection between the mobility **µ** and the random movement of the carriers. Here we will take this for granted (we will come back to this issue later) and take note of this simple, but far-reaching relation, called **Einstein relation**. This relation connects the mobility and the **diffusion coefficient *D*** via

$$\boxed{D = \mu \cdot \frac{kT}{e}}$$

- The diffusion coefficient ***D***, of course, is the proportionality constant that connects the diffusion current ***j*** of particles (not necessarily an electrical current!) to their density gradient $\nabla\rho(x,y,z)$ (with ρ = density of the particles) via **Fick's first law**:

$$\boxed{j(x,y,z) = -D \cdot \nabla\rho(x,y,z)}$$

Bearing this in mind, we now can relate the average distance $(x^2 + y^2 + z^2)^{1/2}$ that a particle moved away from its position **(0,0,0)** at $t = 0$ after it diffused around for a time ***t*** via another Einstein relation by

$$\boxed{\langle r \rangle = \left(D \cdot t\right)^{1/2} = \left(\frac{kT \cdot \mu \cdot t}{e}\right)^{1/2}}$$

- A minority carrier with a specific life time τ thus will be found at an average distance from the point were it was generated given by

$$\boxed{\langle r\rangle = \left(D \cdot \tau\right)^{1/2} =: L}$$

- This specific distance we call the **diffusion length *L*** of the minority carriers.

The diffusion length ***L*** is just as good a measure of the dynamics of the carrier system as the life time τ; we can always switch from one to the other via the Einstein relations.

- Long life times correspond to long diffusion lengths. In today's **Si**, diffusion lengths of **mm** can be achieved – a tremendous distance in the world of electrons and holes.

Real Semiconductors

So far we have explicitly or implicitly only considered "perfect" semiconductors – i.e., semiconductors without any unintentional lattice defects. But now it is easy to consider real semiconductors. Real semiconductors contain unintentional lattice defects and these defects can have two major detrimental effects if they introduce energy levels in the band gap.

- They may influence or even dominate the *carrier density*, i.e. they act as unwanted dopants. In many cases of semiconducting materials without any technical applications so far, the carrier density is pretty much determined by defects and is practically unalterable (it usually also comes in only one kind of conductivity type).
- They influence or dominate the *minority carrier life time*. This may be a major problem, e.g. for solar cells where large life times are wanted in rather imperfect materials.
- In addition, they will also influence the *mobility* (usually decreasing it).

Technically, there are several options for dealing with real semiconductors.

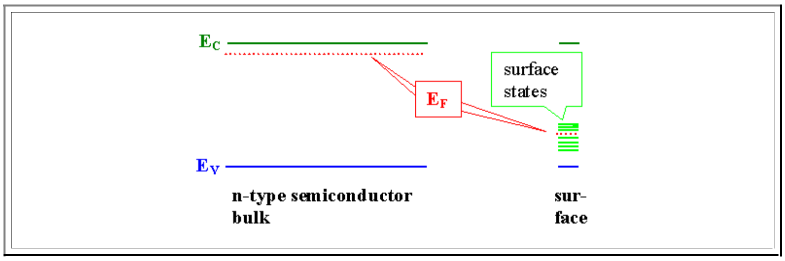
- Make the material very perfect. This is the recipe for **Si** microelectronics and most of the **III-V** compound devices.
- Live with the defects and render them impotent as far as possible, e.g. by **hydrogen passivation**.
- Find semiconductors where the defect levels are not in the band gap (after passivation). An example may be **$CuInSe_2$**, which is usable for solar cells despite lots of defects, and **GaN**, which is the standard material for blue light-emitting diodes.

2.2.4 Simple Junctions and Devices

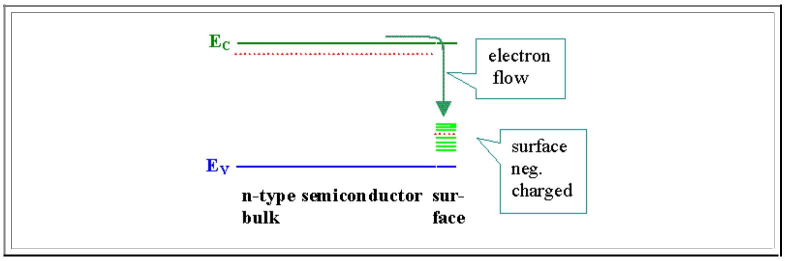
- In this section we will look at some **junctions** in a cursory manner with the goal to get a basic understanding for current flow and the driving forces behind it.
  - We will see that it is possible (with only a little cutting of corners) to come to a complete quantitative description of the current–voltage relationship of a **p–n** junction; even including the (usually omitted) *generation current* part, without doing a lot of math.
  - We will do this in three steps: First we look at a *hypothetical* junction, then at the *classical* ***p–n*** *junction*, then at a *real* **p–n** junction with recombination and generation currents in the space charge region.
- But don't deceive yourself! These are *difficult* subjects, even if they are made to look relatively easy here.
  - We will delve somewhat deeper into these subjects later. However, within the scope of this course, there will not be enough time to go into junction theory thoroughly and in depth.
  - Here we only attempt to give enough background knowledge, so you can understand the rigid treatment of device physics that can be found in many books.

Equilibrium Between a Semiconductor and its Surface as a Model

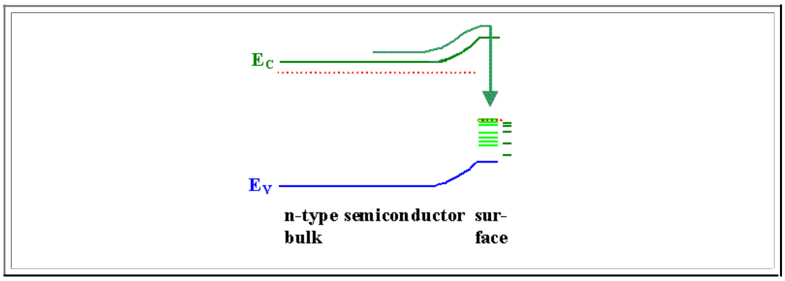
- First, we must realize that the surface is a *defect* – beyond it, there is no periodic lattice anymore; the bonds of the surface atoms have some other arrangements than the bonds in the bulk.
  - The dispersion relation of electrons in the two-dimensional surface "crystal" thus must be expected to be different from that of the bulk.
  - In the simplest approximation conceivable, we may picture the surface with the same band diagram as the bulk (but in only two spacial dimensions), but with plenty of additional states in the band gap.
- We have then the following representation:



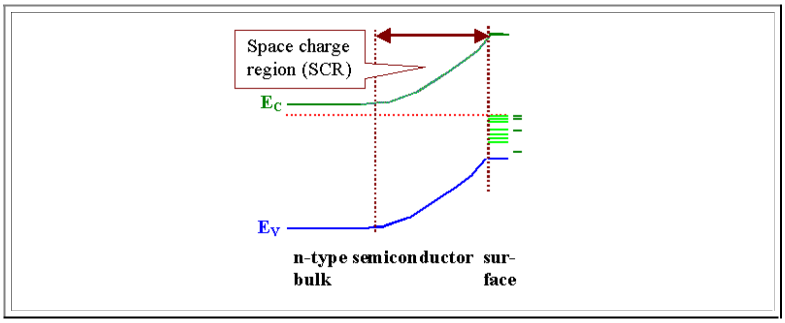
  - The surface introduces some unspecified states in the band gap of an **n**-type semiconductor, shown here arbitrarily as acceptor states. The Fermi energy will adjust itself as shown: Since states in the bandgap are available that are lower than the donor states (not shown for graphic simplicity), electrons will move down to these states – so the Fermi energy also comes down.
- We will now consider what would be happening if we were to bring the bulk and the surface in close contact. Shown here are three graphical representations of this (fictitious!) process:
  - Upon contacting, the electrons from the conduction band of the bulk will flow to the still empty states of the surface. This will put an additional negative charge at the surface. (Note that before that, the surface was not charged.)



  - In response to the negative surface charge, the electrostatic potential at the surface goes up and extends (decreasingly) into the bulk – we have an **internal potential** ***V*(*x*)**. The simplest way to visualize the **band bending** going with this potential is to remember that the electrons from the bulk are repelled by the negative surface charge. Since work is needed to move them to the surface, this is an "uphill" process. (Remember that the band diagram gives the energy of the electrons.)



- As long as the total energy can be reduced by moving electrons from the bulk to the surface, the process will continue. (Remember that energy minimization is the way towards equilibrium.) Equilibrium is reached as soon as it makes no difference any more if one more electron is in the bulk or on the surface, i.e., if we change the electron amount by Δn_e. This means that equilibrium is given by $\mathbf{0 = d}G \mathbf{= (\partial} G/\partial n_e\mathbf{)_{bulk} \cdot \Delta} n_e\mathbf{(bulk) + (\partial} G/\partial n_e\mathbf{)_{surface} \cdot \Delta} n_e\mathbf{(surface)}$.



- From the obvious relation $\Delta n_e\mathbf{(bulk) = -\Delta} n_e\mathbf{(surface)}$ it follows that in equilibrium, we have $\mathbf{(\partial} G/\partial n_e\mathbf{)_{bulk} = (\partial} G/\partial n_e\mathbf{)_{surface}}$. Since $\mathbf{(\partial} G/\partial n_e\mathbf{)}$ = chemical potential μ = Fermi energy E_F, this means that:

$$E_F\mathbf{(bulk)} \;=\; E_F\mathbf{(surface)}$$

This is a crucial point; let's make it again:

In thermal equilibrium, the Fermi energy E_F is the same everywhere.

- If you are not too sure about this, read up about the concept of the chemical potential in the basic module.

This automatically leads to an imbalance of charge, we now have *local violations of charge equilibrium*.

- The area with the band bending will not contain free electrons, but still the positively charged donor atoms (= "ionized"), it thus contains *charges fixed in space* and is called a **space charge region** (***SCR***).
- The **SCR** also contains an electrical field $\underline{E}$ which is given directly by the derivative of the electrostatic potential V by $\underline{E} = -\nabla V$ (or, for the one-dimensional case, by $E_x = -\mathrm{d}V/\mathrm{d}x$), or it can be obtained via the Poisson equation.

The Space Charge Region

The easiest way to think about this electrical field is to consider the *field lines*, which start at a positive charge and end at a negative charge.

- The negative charges are the surplus electrons sitting in the surface states (and thus also in real space at the *surface*), the positive charges are the "ionized" donor atoms in the *bulk*.
- This view immediately leads to a "quick and dirty" formula for the width d of the space charge region: We consider the **capacitance C_{SCR}** of the **SCR**.
- Since the positive charges are spread homogeneously through the volume of the **SCR**, we approximate the capacity of the **SCR** by a plate capacitor with *half* the distance d between the plates, i.e., $d_{cap} = d/2$. In other words, we sort of put all the positive charge on a fictitious plate at half the width of the **SCR**.
- The capacitance then becomes

$$C_{SCR} = \frac{2\,\epsilon_r \cdot \epsilon_0 \cdot A}{d} = \frac{Q}{U}$$

- With ***A*** **=** area of the capacitor plates, ***Q*** **=** charge on the plates and ***U*** **=** potential difference between the plates.
- The charge on the plates is equal to the number of ionized donors in the volume of the **SCR**. The volume is ***d* · *A*** and the number of charges is just the density of donors N_D (assuming that all are ionized) times the elementary charge **e** times the volume ***d* · *A***, so we have

$$Q = e \cdot N_D \cdot d \cdot A$$

- Substituting this in the equation from above, we obtain

$$\frac{2\,\epsilon_r \cdot \epsilon_0 \cdot A}{d} = \frac{e \cdot N_D \cdot d \cdot A}{U}$$

- This gives us one of the more important semiconductor device equations in its simplest form – and this is the *correct* equation despite the somewhat questionable assumption of $d_{cap} = d/2$:

$$d = \left(\frac{2\,\epsilon_r\,\epsilon_0\,U_{bi}}{e\,N_D}\right)^{1/2}$$

- The voltage U_{bi} is the difference of the values of the internal potential between the bulk and the surface; it is called the **built-in potential**. Here, of course, it is simply the difference of the Fermi energies expressed as a potential, i.e.

$$U_{bi} = \frac{\Delta E_F}{e}$$

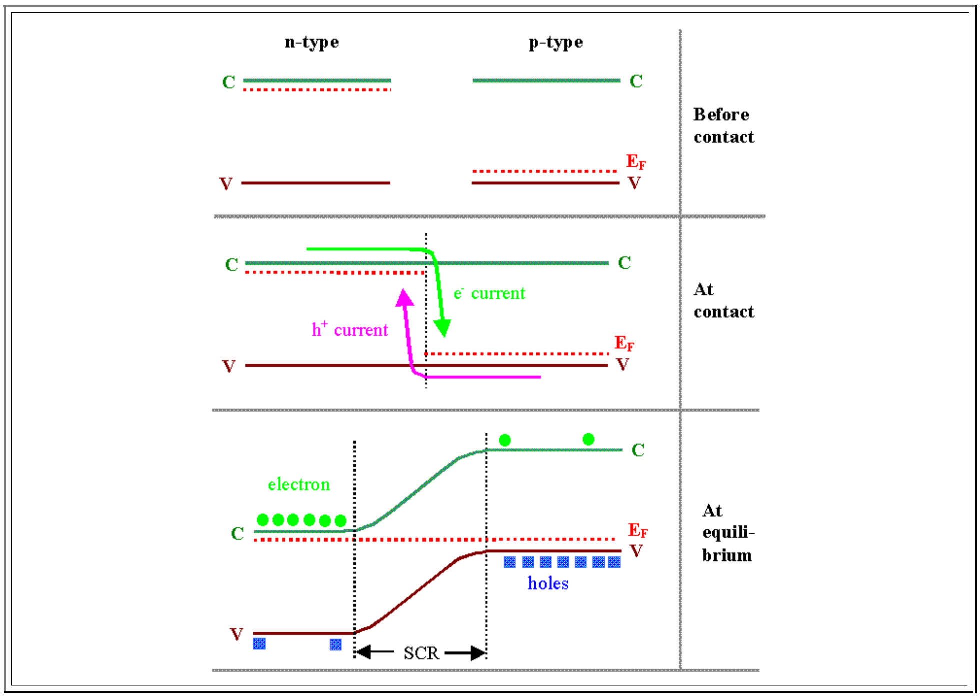
- We immediately can generalize: If, in addition to the "built-in" potential $\Delta E_F/e$, an additional external potential U_{ex} is added from the outside by simply connecting the material to a voltage source at U_{ex}, the total voltage becomes $U = \Delta E_F/e + U_{ex}$, and the width of the space charge region is

$$d = \frac{1}{e} \cdot \left(\frac{2\,\epsilon_r \cdot \epsilon_0 \cdot (\Delta E_F + e \cdot U_{ex})}{N_D}\right)^{1/2}$$

- Note that, since the built-in potential U_{bi} is taken to be positive, a positive U_{ex} will increase the width of the space charge region – by increasing the potential difference between bulk and surface.
- It is easy to to obtain the same equation by integrating the **Poisson equation** for this case. This is done in an illustration module.
- This example illustrates nicely the approach we take in this chapter: We start from the most simple consideration of the case and try to deduce proper relations and formula by analogies – cutting corners a little if necessary (but only as long the results are still correct).

"Ideal" p–n Junction

- We now construct a **p–n** junction exactly along the recipe given above:
- Draw the band diagrams of both parts.
- Join the two parts, move electrons to the materials with the lower Fermi energy, holes opposite.
- Build up space charges and shift the potentials accordingly until the *Fermi energy is the same everywhere*.
- These steps are illustrated below:



- As a graphical aid, which will be useful in cases to come, the carriers are schematically indicated as circles (electrons) and squares (holes). Many such symbols being present is meant as an indication for majority carriers, whereas few ones being present indicates the minorities.

- The only differences to the first situation is that we now have a space charge region on both sides of the junction.
  - The field lines are now pointing from the positively charged donors on the **n**-type side to the negatively charged acceptors on the **p**-doped side.
  - The width of the space charge region is now a little more involved to calculate (but there is nothing new); it comes out to

$$d = \frac{1}{e} \cdot \left(2 \cdot \epsilon_r \cdot \epsilon_0 \cdot \left(\frac{1}{N_A} + \frac{1}{N_D} \right) \cdot \left(\Delta E_F + e \cdot U_{ex} \right) \right)^{1/2}$$

- Now let's look at the various *currents* flowing in the conduction and valence bands without an external voltage.
  - We know that the net current is zero – we are in an equilibrium condition as long as we do not apply an external voltage (or shine some light on it).
  - We also know that we have a dynamic equilibrium as in the case of the recombination/generation business before. The net current is zero because the local currents cancel each other. This implies that the electron current flowing uphill (from left to right) is identical in magnitude and opposite in sign to the one flowing from right to left (downhill).
  - The same reasoning applies, of course, to the holes in the valence band.

- The partial currents discerned above have several specific names. The current component *flowing uphill in energy* is called . . .
  - **Diffusion current**, because the driving force behind this current is the density gradient in the carrier density. This always leads to a current component given by Fick's first law to

$$j_{Diff} = -D\, e\, \nabla n$$

  - With $\boldsymbol{n}$ **=** density of the carriers in question. (The electric charge **e** has to be given explicitly in this equation because $\boldsymbol{n}$ only gives the number of particles present.)
  - It is also alternatively called **recombination current** because all the electrons (or holes) flowing to the other side become excess minority carriers there and must disappear by recombination, which can be depicted as a current flowing between the bands.
  - Looking a little ahead, a **p–n** junction is a diode and currents in diodes are classified as either forward or reverse current. Well, the current component in question here is responsible for the **forward current** in the diode and therefore is also addressed under that name. We will use mostly this name and abbreviate this current component with j_F.

- The partial current *flowing downhill in energy* is called . . .
  - **Field current**, because it is the current that the electrical field in the junction produces; i.e. the carriers flow in the general direction of the field lines with the respective proper signs. The diffusion current, in contrast, flows against the force exerted by the field (which will slow down these carriers!)

- **Drift current**, because it is the current that results from a drift – caused by the electrical field – superimposed on random diffusion movements.
- **Generation current**, because the (minority) carriers that were swept down the energy hill (or up in the case of the holes) get replaced by increased generation, thereby keeping the density constant.
- **Reverse current** in the diode nomenclature explained above. We will use mostly this name and abbreviate this current component with j_R.

Simple Current–Voltage Characteristics

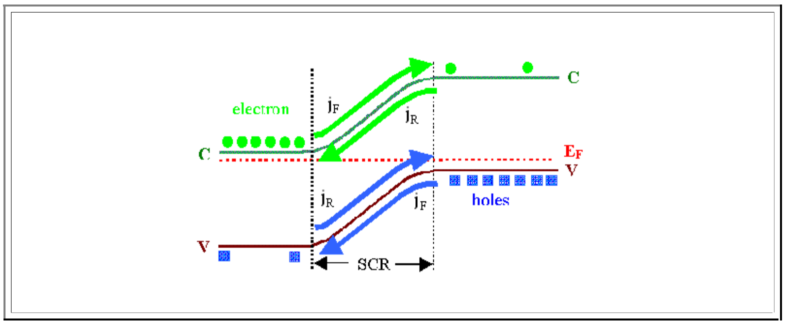
- In equilibrium, without an external voltage U_{ex} and without illumination, we know that the net current is zero, or
  - $j_F = -j_R$, or, to be more precise,

$$j_F(U_{ex}=0) \;=\; -\,j_R(U_{ex}=0)$$

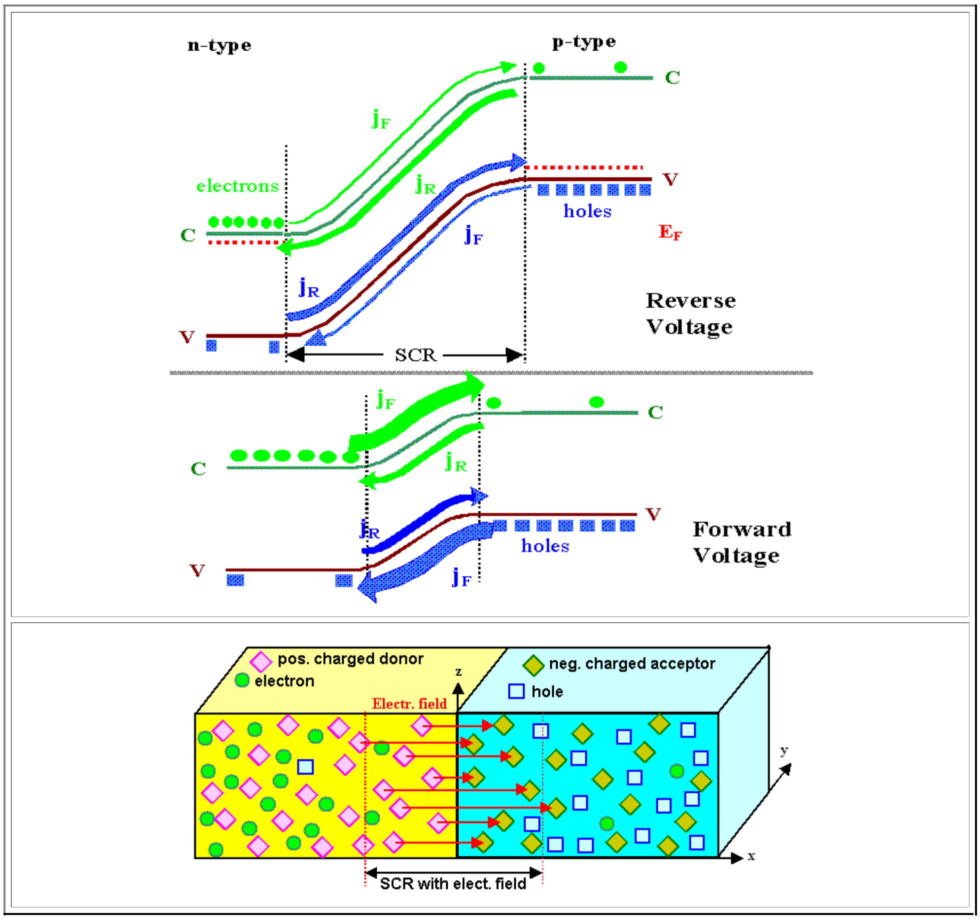
- To make life a little easier, we now drop some matter-of-course indices, *and the signs*; i.e. we only look at the *magnitudes of the current components*. It is easy enough to sort out the signs in the end again; in case of doubt refer to the link. In this shorthand notation we have

$$j_F(0) \;=\; j_R(0)$$

  - If we draw these currents schematically into the band diagram from above, it looks like this:



  - Of course, the currents do not flow on both sides of the band edge; this is simply a drawing means.
- What happens for finite external voltages? First let's look at the band diagrams:
  - An external potential U_{ex} is added, which will raise or lower the equilibrium potential, depending on its sign; for sake of simplicity we only move the **p**-side band diagram. But do we have to move it up or down?
  - It depends: We have two possibilities for choosing the polarity of the external voltage, either increasing the already existing electrical field or weakening it. (Remember that the field lines are pointing from the **n**-side to the **p**-side; if you have forgotten why this is so, look again at the explanation given above.)
  - Thus, the potential on the **p**-side goes *up* for the negative pole of the voltage supply on the **p**-side (always think about if electrons are repelled or attracted if you are unsure about how a potential moves bands), and *down* for the positive pole on the **p**-side.
  - We no longer have an equilibrium situation – the Fermi energy is no longer the same everywhere; we must leave it undefined across the junction. (Later on we will have a look at how the Fermi energy behaves inside the **SCR**.)
  - Far away from the junction, however, nothing (or very little) has changed. We still may consider these parts of the semiconductor to be in equilibrium (or at least very close to it).
- The band diagram for the two possible basic cases then look like this:



- The diffusion currents are shown increased (thicker arrow) if the energy barrier was lowered, and decreased if it was raised. The drift currents are unchanged.

The situation now is rather simple. The potential step is either increased or decreased. Let's first look what happens to the *forward currents* $\boldsymbol{j_F(U_{ex})}$.

- At zero volts the electron and hole forward currents have a certain value that is certainly determined by the height of the energy barrier that the carriers have to overcome, following *Boltzmann statistics* (as an approximation to the Fermi statistics, of course).
- In other words, the equation for this current includes an **exp[–*E*/(k*T*)]** term. Changing the energy barrier ***E*** by Δ ***E*** simply means to multiply the current at zero volts with **exp[–Δ*E*/(k*T*)]**.
- Since the magnitude of $\boldsymbol{j_F}$ for zero external voltage is just $\boldsymbol{j_R(U_{ex}=0)}$, the forward current current $\boldsymbol{j_F(U_{ex})}$ at *any* voltage $\boldsymbol{U_{ex}}$ is (and this is true both for the electron and the hole forward current)

$$j_F(U_{ex}) \;=\; j_R(U_{ex}=0)\cdot\exp\left(-\frac{\Delta E}{kT}\right) \;=\; j_R(U_{ex}=0)\cdot\exp\left(-\frac{e\cdot U_{ex}}{kT}\right)$$

- Note that, due to the minus sign in the exponent (which we inherited from the Boltzmann distribution), increasing the energy barrier leads to a decrease of the forward current; we will come back to this later.
- Imagine this as a lot of (drunken) bicyclists with various random momentums driving around randomly at the foot of a hill. Some of them on occasion will make it up the hill because their momentum was large enough and they were heading in the right direction. The fraction of bicyclists making it will simply change exponentially with the Boltzmann factor relative to their "current" at some reference value if the hill is raised or lowered.

Now to the *reverse current* $\boldsymbol{j_R(U_{ex})}$.

- It corresponds to (drunken = randomly moving) bicyclists that drive around for a certain time (corresponding to the life time of the minority carriers) on the plateau on top of the hill before falling off their bikes (recombining).
- However, everybody who by accident makes it to the edge of the hill will invariably careen down, i.e., produce a reverse current.
- Clearly, it only matters how many carriers happen to make it to the edge of the potential drop, not how deep it is. In other words: *The reverse current does not depend on the external voltage*, or

$$j_R(U_{ex}) \;=\; j_R(U_{ex}=0) \;=\; j_R$$

The *total current* is simply the *difference* between forward and reverse current for the electrons and the holes, so we have

$$\boxed{\begin{aligned} j(U_{ex}) &= \Big(j_F(U_{ex}) - j_R\Big)_e + \Big(j_F(U_{ex}) - j_R\Big)_h \\ j(U_{ex}) &= \left(j_R{}^e + j_R{}^h\right) \cdot \left(\exp\left(-\frac{e \cdot U_{ex}}{kT}\right) - 1\right) \end{aligned}}$$

- Note that we did *not* assume that the forward or reverse current of the holes must be identical to the forward or reverse current of the electrons, respectively. Of course, if everything were symmetrical, we would have $j_R{}^e = j_R{}^h$, but we want to keep it a as general as possible even at that level since many real devices employ wildly different electron and hole currents across the junction.

This is the famous **diode equation**, and this is all there is to it for straight-forward **p–n** junctions – except that the technically relevant diode voltage U_D, being related to the technical current flow direction, is taken as *positive* for the current flowing in *forward* direction; thus, we have $U_D = -U_{ex}$.

- This just means that, in order to make the forward current flowing, the external voltage must be applied such that it effectively reduces the built-in potential difference.

Combining the reverse currents of electrons and holes in a single constant $j_0 := j_R{}^e + j_R{}^h$, we arrive at the final form of the ideal diode equation:

$$\boxed{j(U_D) = j_0 \cdot \left(\exp\left(\frac{e \cdot U_D}{kT}\right) - 1\right)}$$

All that is left to do is to consider the reverse currents j_0 a bit more closely.

- Thinking again about drunken bicyclists, but now driving around on the top of the hill (and thus representing electrons), we might be tempted to assume that $j_R{}^e$ should be *proportional to their numbers*, i.e., to the minority carrier *density* on the plateau.

This, however, is *wrong*.

- So let's sober up and think a bit harder: You only can extract the bicyclists *once*! Yet if you want a *constant* current over time, the best you can do is to take all carriers for the current that are *(i) generated per time unit and (ii) making it to the edge*.
- In other words, you *need to refill the ranks of bicyclists* and the current will be *proportional to the generation rate* ***G*** (to what comes out of the bars per time interval), which we know is equal to the *recombination rate* ***R*** in undisturbed semiconductors and given by

$$\boxed{G = R = \frac{n_{min}}{\tau} = \frac{n_i{}^2}{\tau \cdot N_{Dop}}}$$

- With τ = lifetime.

Thinking a bit harder yet, we realize that minority carriers generated way back from the junctions will not contribute to the current. They will "fall off their bike" (recombine) long before they have a chance to come close to the drop-off. Obviously we only must consider carriers within a certain distance of the junction, and this distance is, of course, the *diffusion length* ***L***.

- This gives us

$$\boxed{j_R = \text{const.} \cdot e \cdot L \cdot G = \frac{\text{const.} \cdot e \cdot L \cdot n_{min}}{\tau} = \frac{\text{const.} \cdot e \cdot L \cdot n_i{}^2}{\tau \cdot N_{Dop}}}$$

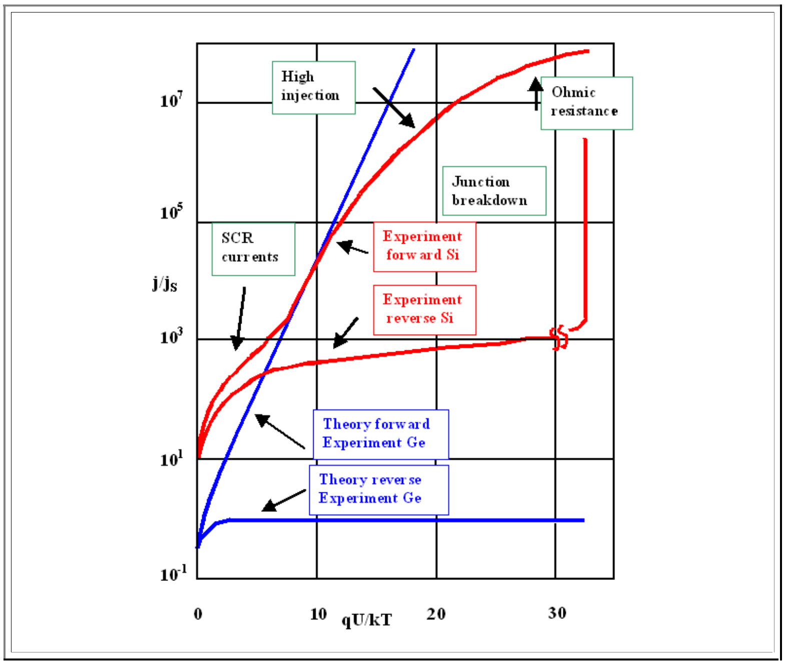
- The elementary charge **e** is needed to make an electrical current out of a particle current. And as it turns out by more involved calculations, the *proportionality constant is =* ***1***.

This gives us the complete diode equation exclusively in terms of *primary material properties*:

$$\boxed{j(U_D) = \left(\frac{e \cdot L \cdot n_i{}^2}{\tau \cdot N_A}\left(\exp\left(\frac{e \cdot U_D}{kT}\right) - 1\right)\right)_{\text{electrons}} + \left(\frac{e \cdot L \cdot n_i{}^2}{\tau \cdot N_D}\left(\exp\left(\frac{e \cdot U_D}{kT}\right) - 1\right)\right)_{\text{holes}}}$$

This is a remarkable achievement – obtained without involved calculations and cutting corners only once (**proportionality const. = 1**). *But how good is it*? Only the experiment can tell.

- If we measure the current–voltage characteristic of an "ideal" **p–n** junction, we will find the following curves:



- Generally, for **Ge** (or other semiconductors with relatively small band gaps) the measured characteristics is remarkably close to the one predicted by our formula. For **Si**, however (and other semiconductors with large band gaps), *it is not a good formula*, particularly for for the reverse current. We have large deviations from theory, labeled with black lettering. So, let's see what went wrong (more details in the link) in each case.

First, our theory has the usual (trivial) omissions. We did not include any ohmic resistance which can be easily added by putting a resistor in series to the junction. The result is the ohmic behavior for larger voltages as seen at larger currents.

Second, everything will break down at high field strength; this is true for a junction, too. So for large reverse voltages (easily in the range **100 V . . . 1,000 V**), the junction will go up in smoke while drawing a large current called "junction breakdown".

- While these points would apply to a **Ge** junction too (they are not shown in the ideal characteristics), the remaining deviations for **Si** involve a major oversight in our present theory:

Third: **We did not include carrier generation (and recombination) in the space charge region!**

- In the bicyclist picture, we did not take into account that there are bars along the slope of the hill, too, which will emit bicyclists with a certain momentum and direction – they may either go up or down the hill and thus add to the current of particles moving in either direction.
- This adds four more current components (forward and reverse for holes and electrons) summarily called **generation currents from the SCR**.

How large are these **SCR**-caused current components?

- The answer comes from one of the more involved problems in semiconductor physics, it is not easy to obtain for real semiconductors (we will do it later). However, there is an *easy way of thinking about it* that even comes up with the usually given formula resulting from serious (but still approximate) computing. But we sure will have to cut a few corners!
- Here we go:

Current–Voltage Characteristics with Generation Currents from the Space Charge Region

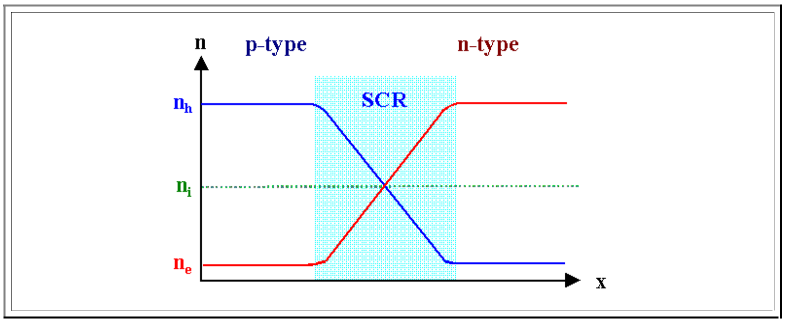
We know the maximum current $j_{\max}$ that could emerge from the space charge region: It is the generation rate of carriers times the width of the **SCR** in complete analogy to the discussion above. More than that cannot flow out (in one direction for the maximum) per unit time.

- Every generation event produces a hole and an electron, so we have

$$\boxed{j_{\max} \;=\; 2e \cdot G_{SCR} \cdot d}$$

- With d = width of the **SCR**, and G_{SCR} = generation rate inside the space charge region.

How large is G_{SCR}? This needs to be considered in detail since it wouldn't help to simply use the known equation $G = n_{\min}/\tau$, because $n_{\min}$ is *not constant* across the **SCR**; *very schematically*, it rather looks like this:



- Inside the **SCR**, this makes ***G*** a strong function of ***x*** – how are we going to handle this?
  - Well, let's take a kind of *average value* for the carrier density, and what suggests itself is the *intrinsic carrier density* $\boldsymbol{n_i}$, which is the value at the point where the two curves cross each other because of the mass action law stating $n_e \cdot n_h = n_i^2$.
  - Within this (questionable!) approximation the maximum current generated in the **SCR** then becomes

$$j_{max} = \frac{2e \cdot n_i \cdot d}{\tau}$$

- And we know more: If the external voltage is zero, the total current is zero and this means that half of the **SCR** current must be forward and the other half reverse. We have

$$j_F(0) = j_R(0) = \frac{e \cdot n_i \cdot d}{\tau}$$

- If we now change the barrier height by eU_{ex}, the forward current will change as before. However, we cannot simply multiply by $\exp[-eU_{ex}/(kT)]$ as before!
  - While a carrier generated at the bottom of the hill experiences the full added potential, a carrier generated further up sees less or even no potential if it originates all the way uphill.
  - So again, let's be sloppy and assume an *average additional energy barrier*, average between everything and nothing – and this will be $eU_{ex}/2$.
  - Again assuming that j_R does not depend on the *barrier height* (but slightly on ***U*** since the width ***d*** of the **SCR** is voltage dependent), this gives us:

$$j_F(U_{ex}) = \frac{e \cdot n_i \cdot d}{\tau} \cdot \exp\left(-\frac{eU_{ex}}{2kT}\right)$$

$$j_R = \frac{e \cdot n_i \cdot d}{\tau}$$

  - The two components no longer add up to j_{max}, but we don't have to worry about this. We only would be very wrong for *large* forward currents, but in this case the *bulk* forward current is always much larger anyway – so it does not really matter much for the diode behavior.
- The total current from the space charge region then becomes $j_{SCR} = j_F - j_R$. Written out and, as above, using $U_D = -U_{ex}$, we have

$$j_{SCR}(U_D) = \frac{e \cdot n_i \cdot d}{\tau} \cdot \left(\exp\left(\frac{eU_D}{2kT}\right) - 1\right)$$

  - This happens to be *exactly the same formula* (give or take a factor of **2**) *that we would obtain with the "proper" theory*.

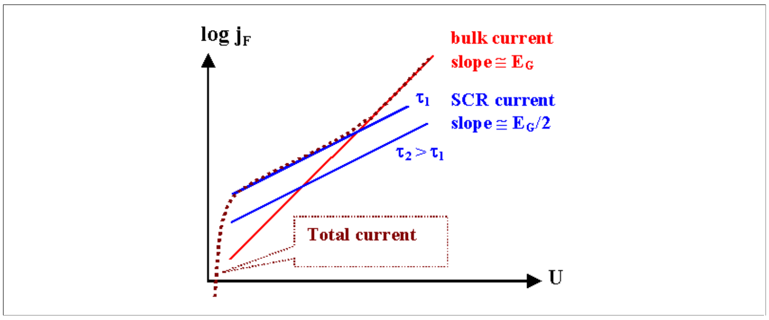
- We now can write down the diode equation in all its splendor:

$$j_{\text{total}}(U_D) = \underset{\text{e}^- \text{ bulk}}{\frac{e\cdot L\cdot n_i^2}{\tau\cdot N_A}\left(\exp\left(\frac{eU_D}{kT}\right) - 1\right)} + \underset{\text{h}^+ \text{ bulk}}{\frac{e\cdot L\cdot n_i^2}{\tau\cdot N_D}\left(\exp\left(\frac{eU_D}{kT}\right) - 1\right)} + \underset{\text{h}^+\text{, e}^- \text{ SCR}}{\frac{e\cdot n_i\cdot d}{\tau}\left(\exp\left(\frac{eU_D}{2kT}\right) - 1\right)}$$

  - We may use the same abbreviation for the bulk reverse current as above and additionally abbreviate the SCR one in a similar manner, then we see more clearly that we effectively have two diodes parallel, the first (prefactor j_{01}) representing the bulk currents and the second (prefactor j_{02}) representing the SCR currents:

$$j(U_D) \;=\; j_{01}\cdot\left(\exp\left(\frac{e\cdot U_D}{kT}\right) - 1\right) + j_{02}\cdot\left(\exp\left(\frac{e\cdot U_D}{2kT}\right) - 1\right)$$

- Let's see what it means in forward direction:
  - We may neglect the **–1** in the **SCR** part of the forward current; it then adds a component that increases with **exp[eU_D / (2kT)]**, i.e., with *half the slope* of the bulk current (in an Arrhenius diagram). The half-slope component will always "win" at small voltages but pale to insignificance at higher voltages as shown in the illustration:



  - The combined characteristic looks very much like the measured behavior shown above; the sharp drop of the current close to U **= 0 V** is due to the **–1** in the diode equation (i.e., the reverse current j_R) which we neglected for larger voltages.
  - We also see that the exact value of the **SCR** forward current indeed only matters for small voltages, as claimed above.
- What do we get in *reverse direction*?
  - First, *the reverse current now is voltage dependent* because the width of the space charge region and thus j_R increases with a $U^{1/2}$ law.
  - Second, the *total reverse current now is larger*. Assuming a symmetric junction ($N_D = N_A = N_{Dop}$) and identical life times (and so on) for electrons and holes, it is easy to calculate the relation **j_R(bulk) / j_R(SCR)**; we have

$$\frac{j_R(\text{bulk})}{j_R(\text{SCR})} = \frac{\dfrac{e\cdot L\cdot n_i^{\,2}}{\tau\cdot N_{Dop}}}{\dfrac{e\cdot d\cdot n_i}{\tau}} = \frac{L\cdot n_i}{N_{Dop}\cdot d}$$

- The decisive factor is n_i. It decreases exponentially with increasing band gap E_g.
  - This answers the question why **Ge** junctions follow the simple theory, while **Si** junctions are far off: If $n_i \gg N_{Dop}$, then **j_R(bulk) >> j_R(SCR)** and the **SCR** contribution will not be felt.
  - The generation current from the **SCR** thus is much more important in semiconductors with larger band gaps. The characteristics from above show this rather clearly.
  - Whereas the **SCR** part may be safely neglected for **Ge** (**E_g = 0.6 eV**, it is about (**10^2 . . . 10^3**) times larger than the bulk diffusion current in **Si**.
  - The **SCR** currents should be *absolutely dominating* in large band-gap semiconductors. Not only is n_i rather small, but **$L = (D\tau)^{1/2}$** is very small, too, since these materials are often direct semiconductors.

2.3 Elements of Advanced Theory

2.3.1 Effective Masses

- So far we have treated electrons and holes as identical (except for their charge and concentration) and we also always assumed thermal equilibrium.
  - On occasion, we were somewhat inconsistent about the issues. We stated that **p–n** junctions with external voltages are actually *not* in equilibrium, or we kept the electron and hole parts of the currents separate, but we did not really make much of this – for such situations, we used equilibrium formulas, or did not justify why we do not mix electron and hole currents in the junction.
  - In this subchapter we delve deeper into some "advanced" subjects (they are still pretty elementary) to get at least these aspects straight.
  - First, we will see how we can save large parts of the simplicity of the free electron gas model by assigning *effective masses* to the carriers.
- In the free electron gas model an electron had the mass **m**$_e$ [always written straight and not in *italics* because it is not a variable but a constant of nature (disregarding relativistic effects for the moment), and from now on without the subscript $_e$] – and that was all to it.
  - If a force ***F*** acts on it, e.g. via an electrical field ***E***, in *classical mechanics* **Newtons laws** applies and we can write

$$\underline{F} = -\,e\cdot\underline{E} = m\cdot\frac{d^2\underline{r}}{dt^2}$$

  - With $\underline{r}$ = position vector of the electron.
  - An equally valid description is possible using the momentum $\underline{p}$ which gives us

$$\underline{p} = m\cdot\frac{d\underline{r}}{dt}$$

$$\underline{F} = -e\cdot\underline{E} = \frac{d\underline{p}}{dt}$$

- *Quantum mechanics* might be different from classical mechanics, so let's see what we get in this case.
  - The essential relation to use is the identity of the *particle velocity* with the the **group velocity** $\underline{v}_{group}$ of the wave package that describes the particle in quantum mechanics. The equation that goes with it is

$$\underline{v}_{group} = \frac{1}{\hbar}\cdot\nabla_k\,E(\underline{k})$$

  - Let's see what we get for the free electron gas model. We had the following expression for the energy of a quasi-free particle:

$$E(\underline{k}) = E_{kin} = \frac{\hbar^2\cdot k^2}{2m}$$

  - For $\underline{v}_{group}$ we then obtain

$$\underline{v}_{group} = \frac{1}{\hbar}\cdot\nabla_k\,\frac{\hbar^2\cdot k^2}{2m} = \frac{\hbar\cdot\underline{k}}{m} = \frac{\underline{p}}{m} = \underline{v}_{classic}$$

  - Since $\hbar\underline{k}$ was equal to the momentum $\underline{p} = m\underline{v}_{classic}$ of the particle, we have indeed $\underline{v}_{group} = \underline{v}_{classic} = \underline{v}_{phase}$ as it should be.

- *In other words*: As long as the $E(\underline{k})$ curve is a parabola, *all* the energy may be interpreted as *kinetic energy* for a particle with a (constant) mass **m**.
- Contrariwise, if the dispersion curve is *not a parabola*, not all the energy is kinetic energy (or the mass is not constant? ??).

How does this apply to an electron in a periodic potential? .

- We still have the wave vector $\underline{k}$, but $\hbar\underline{k}$ is no longer identical to the momentum of an electron (or hole), but is considered to be a **crystal momentum**.
- $E(\underline{k})$ is no longer a parabola, but a more complicated function.

Since we usually do not know the exact $E(\underline{k})$ relation, we seem to be stuck. However, there are some points that we still can make:

- Electrons at (or close to) the Brillouin zone in each band are diffracted and form standing waves, i.e. they are described by superpositions of waves with wave vector $\underline{k}$ and $-\underline{k}$. Their *group velocity* is necessarily close to zero!
- This implies that $\nabla_k E(\underline{k})$ at the **BZ** is close to zero too, which demands that the dispersion curve is almost horizontal at this point.
- The most important point is: We are not interested in electrons (or holes) far away from the band edges. Those electrons are just "sitting there" (in $\underline{k}$-space) and not doing much of interest; only electrons and holes at the band edges (characterized by a wave vector $\underline{k}_{ex}$) participate in the generation and recombination processes that are the hallmark of semiconductors.

We are therefore only interested in the properties of these electrons and holes – and consequently only that part of the dispersion curve that defines the *maxima* or *minima* of the valence band or conductance band, respectively, is important.

- The thing to do then is to expand the $E(\underline{k})$ curve around the points $\underline{k}_{ex}$ of the extrema into a Taylor series, written, for simplicity's sake, as a scalar equation and with the terms after k^2 neglected.

$$E_n(\underline{k}) \;=\; E_{V,C} \;+ \underline{k}\cdot \left.\frac{\partial E_n}{\partial k}\right|_{k_{ex}} + \frac{k^2}{2}\cdot\left.\frac{\partial^2 E_n}{\partial k^2}\right|_{k_{ex}} + \;....$$

- Since we chose the extrema of the dispersion curve, we necessarily have

$$\left.\frac{\partial E_n}{\partial k}\right|_{k_{ex}} = 0 \qquad\qquad E_n(k_{ex}) \;=\; E_V \;\;\text{or}\;\; E_C$$

- i.e. we are looking at the top of the valence and the bottom of the conductance band.

This leaves us with

$$E_n(k) \;=\; E_{V,C} \;+\; \frac{k^2}{2}\cdot\left.\frac{\partial^2 E_n}{\partial k^2}\right|_{k_{ex}}$$

If we now look at the conduction band and consider *only* the deviation from its bottom (the zero point of the energy scale usually is at the top of the valence band), we have *the same quadratic relation in **k** as for the free electron gas*, provided we change the definition of the mass as follows:

$$\left.\frac{\partial^2 E_C}{\partial k^2}\right|_{k_{ex}} \;=:\; \frac{\hbar^2}{m^*}$$

- which permits to rewrite the Taylor expansion for the conduction band as follows

$$E_C(k) \;-\; E_g = \frac{\hbar^2}{2m^*}\cdot k^2$$

- And this is the same form as the dispersion relation for the free electron gas!

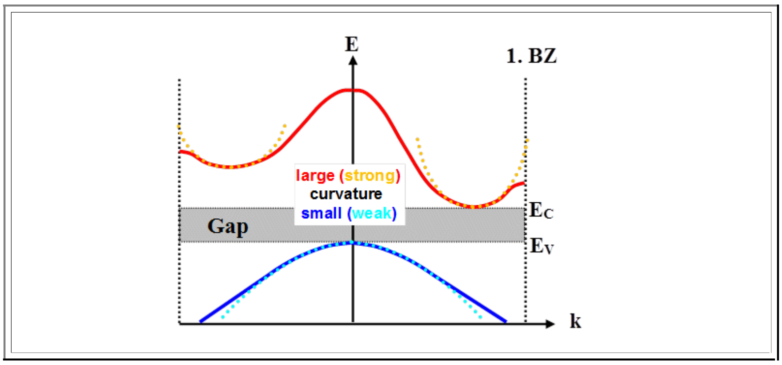
However, since $\partial^2 E_{V,C}/\partial k^2$ may have *arbitrary* values, the *effective* mass m^* of the quasi-free particles will, in general, be different from the regular *electron rest mass* **m**.

  - We therefore used the symbol ***m**** which we call the **effective mass** of the carrier and write it in *italics*, because it is no longer a constant but a variable. It is defined by

$$m^* \; = \; \hbar^2 \cdot \frac{1}{\partial^2 E_n \,/\, \partial k^2 \,|_{k_{ex}}}$$

  - The decisive factor for the effective mass is thus the *curvature of the dispersion curve at the extrema*, as expressed in the second derivative. Large curvatures (= large second derivative = *small* radius of curvature) give small effective masses, small curvatures (= small second derivative = *large* radius of curvature) give large ones.

- Let's look at what we did in a simple illustration and then discuss what it all means.



  - Shown is a band diagram not unlike **Si**. The true dispersion curve has been approximated in the extrema by the parabola resulting from the Taylor expansion (dotted lines). The red ones have a larger curvature (i.e. the radius of an inscribed circle is small); we thus expect the effective mass of the electrons to be smaller than the effective mass of the holes.

- *The effective mass has nothing to do with a real mass*; it is a mathematical contraption. However, if we know the dispersion curves (either from involved calculations or from measurements), we can put a number to the effective masses and find that they are not too different from the real masses.

  - This gives a bit of confidence to the following interpretation (which can be fully justified theoretically):

  - If we use the effective mass ***m**** of electrons and holes instead of their real mass **m**, we may consider their behavior to be identical to that of electrons (or holes) in the free electron gas model.

- This applies *in particular* to their response to forces. In this case, the deviation from the real mass takes care of the influence of the lattice on the movement of the particle.

- Taken to the extremes, this may even imply *zero or negative effective masses*. (Yes, even zero effective mass is possible, but this is a special case of its own and will not be discussed here; for example, graphene has some ***k***-points where $m^* = 0$.)

  - As one can see from the dispersion curves, the effective mass of the electrons in the valence band, m^*_V, is always negative. (This means that a force in $+x$ direction will cause such an electron to move in $-x$ direction.)

  - Since this is counter-intuitive, usually one doesn't consider the electrons in the valence band, but just the unoccupied places – i.e., the holes; however, in their proper definition (which we have disregarded so far), holes have a positive mass by setting

$$m^*_h := -m^*_V$$

  - Note (again) that "holes" only exist in the valence band; this is due to their proper definition as quasi-particles contributing to the electrical conductivity. Thus, all other empty states – in the conduction band, or on additional levels (e.g. of dopant atoms) – are *not* holes; this is clear since they do not contribute to the electrical conductivity.

- We will not go into more detail but give some (experimental) values for effective masses:

Holes: *m**/m							
IV; IV-IV		**III-V**		**II-VI**		**IV-VI**	
C	0.25	AlSb	0.98	CdS	0.80	PbS	0.25
Si	0.16 (0.49)	GaN	0.60	CdSe	0.45	PbTe	0.20
Ge	0.04 (0.28)	GaSb	0.40	ZnO	?		
SiC (α)	1.00	GaAs	0.082	ZnS	?		
		GaP	0.60				
		InSb	0.40				
		InAs	0.40				
		InP	0.64				

Electrons : *m**/m							
IV; IV-IV		**III-V**		**II-VI**		**IV-VI**	
C	0.2	AlSb	0.12	CdS	0.21	PbS	0.25
Si	0.98 (0.19)	GaN	0.19	CdSe	0.13	PbTe	0.17
Ge	1.64 (0.082)	GaSb	0.042	ZnO	0.27		
SiC (α)	0.6	GaAs	0.067	ZnS	0.40		
		GaP	0.82				
		InSb	0.014				
		InAs	0.023				
		InP	0.077				

- If there are two values it simply means that two dispersion curves (from different branches in ***k***-space) have an extremum at the same point.
  - We already get a feeling that it may make a difference if you work with electrons or holes in a certain device whenever you consider its frequency limit: As soon as the carriers can no longer follow rapidly changing forces (alternating electrical fields at high frequencies), the device will not work anymore.
  - Small effective masses mean small (apparent) inertia or high mobilities **μ**. Looking at one of the many formulas for mobility, $\mu = e \cdot \tau_S / m$, the mobility goes up if we insert the (smaller) effective mass. It may thus be wise to use **p**-doped **Si** if high frequencies matter (and everything else does not matter).
- We also notice that **Ge** has the smallest effective mass for holes!
  - This means that its holes respond more quickly to the accelarating force of an electrical field and that means that they also can change direction more quickly than holes in other semiconductors if the electrical field changes its direction.
  - In other words: If speed only depends on the effective mass of the carriers, **Ge** still "works" at high frequencies when the other semiconductors have given up!
  - And that is why we now (**2004**) see a sudden revival of **Ge**, which after its brief period of glory in the end of the fifties / beginning of the sixties of the **20**th century, when it was the one and only semiconductor used for single transistors, was all but extinct.

2.3.2 Quasi Fermi Energies

So far we implicitly defined (thermal) equilibrium as a *total* equilibrium involving *three* components if you think about it:

- **1.** Equilibrium of the *electrons* in the *conduction band*. This means their density was given (within the usual approximations) by

$$n_e = N_{\text{eff}}^{e} \cdot \exp - \frac{E_C - E_F}{kT}$$

- **2.** Equilibrium of the *holes* in the *valence band*. This means their density was given (within the usual approximations) by

$$n_h = N_{\text{eff}}^{h} \cdot \exp - \frac{E_F - E_V}{kT}$$

- **3.** Equilibrium between the electrons and holes, i.e. *between the bands*. This means that the *Fermi energy is the same for both bands* (and positioned somewhere in the band gap).

Definition of the Quasi Fermi Energy

If an equilibrium is disturbed, it takes a certain time before it is restored again; this is described by the *kinetics* of the processes taking place. In a strict sense of speaking, the Fermi energy is not defined without equilibrium, but only after it has been restored. This restoring process occurs in the bands and between the bands:

- *In the bands*, *local* equilibrium (in ***k***-space) between the carriers will be obtained after there was time for some collisions, i.e. after some multiples of the scattering time. This process is known as **thermalization** and occurs typically in *picoseconds*.
- *Between the bands*, equilibrium will be restored by generation and recombination events and this takes a few multiples of the carrier life time, i.e. at least *nanoseconds* if not *milliseconds*.

This means that we can have a *partial (or local) equilibrium* in the bands long before we have equilibrium between the bands. This **local equilibrium** implies that:

- Non-equilibrium means something changes in time. Changes in the properties of the particle ensemble considered (i.e. electrons and holes) in *local* equilibrium are only due to "traffic" *between* the bands while the properties of the particles in the band do not change anymore. The term "local" of course, does not refer to a coordinate, but to a band.
- The carrier densities therefore do not have their total or global equilibrium value as given, e.g., by the mass action law, but their local equilibrium density can still be given in terms of the equilibrium distribution by

$$n_e = N_{\text{eff}}^{e} \cdot \exp - \frac{E_C - E_F{}^{e}}{kT}$$

$$n_h = N_{\text{eff}}^{h} \cdot \exp - \frac{E_F{}^{h} - E_V}{kT}$$

- with the only difference that the *Fermi energy now is different for the electrons and holes*. Instead of *one* Fermi energy $\boldsymbol{E_F}$ for the whole system, we now have *two* **Quasi Fermi energies**, $\boldsymbol{E_F{}^{e}}$ and $\boldsymbol{E_F{}^{h}}$.

For the product of the carrier densities we now obtain a somewhat modified mass-action law

$$n_e \cdot n_h = N_{\text{eff}}^{e} \cdot N_{\text{eff}}^{h} \cdot \exp - \frac{(E_C - E_V) + (E_F{}^{h} - E_F{}^{e})}{kT} = n_i^2 \cdot \exp - \frac{E_F{}^{h} - E_F{}^{e}}{kT}$$

- For this we used the by now basic relation

$$N_{\text{eff}}^{e} \cdot N_{\text{eff}}^{h} \cdot \exp -\frac{E_C - E_V}{kT} = n_i^2$$

- The name "*Quasi Fermi energy*" is maybe not so good, there is nothing "*quasi* " about it. Still, that's the name we and everybody else will use. Sometimes it is also called "**Imref**" (Fermi backwards), but that doesn't help much either.
  - Rewriting the equations from above gives a kind of definition for the Quasi Fermi energies:

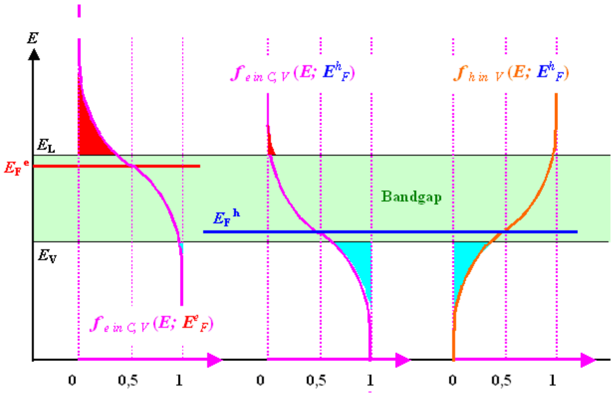
$$E_F{}^{e} = E_C - kT \cdot \ln \frac{N_{\text{eff}}^{e}}{n^{e}}$$

$$E_F{}^{h} = E_V + kT \cdot \ln \frac{N_{\text{eff}}^{h}}{n^{h}}$$

- Quasi Fermi energies are extremely helpful for the common situation where we do have non-equilibrium, but only *between* the bands – and that covers most of semiconductor devices under conditions of current flow (due to an applied voltage) or under illumination. We will make frequent use of Quasi Fermi energies!

Carrier Densities and Quasi Fermi Energies

- If we calculate carrier densities in non-equilibrium with the Quasi Fermi energies, we have to be careful to use the right Quasi Fermi energy in the Fermi-Dirac formula or in the Boltzmann approximation.
  - After all, we now have *two* (Quasi) Fermi energies, one "regulating" the density of electrons in the conduction band, and the other one doing the same for the holes in the valence band. That was already implied above, here we want to make this topic a bit clearer; we also introduce a new distribution function as a kind of short-hand.
  - You really must now write $\mathbf{f}(E, E_F{}^{e}, T)$ or $\mathbf{f}(E, E_F{}^{h}, T)$ instead of simply $\mathbf{f}(E, E_F, T)$ or $\mathbf{f}(E)$ because, due the different arguments, the meaning of these two expressions is now different. This is illustrated below with the two curves on the left and should be obvious.
  - In the pictures we even have some redundancy by writing $\mathbf{f}_{\text{e in C}}(E, E_F{}^{e}, T)$ and so on. This is not necessary, but helps in the beginning to avoid mix-up.



- The density of electrons or holes in the conduction or valence band, respectively, would now be

$$n_e = N_{\text{eff}}^{e} \cdot [\mathrm{f}_{\text{e in C}}(E, E_F{}^{e}, T)] \approx N_{\text{eff}}^{e} \cdot \exp -\frac{E_C - E_F{}^{e}}{kT}$$

$$n_h = N_{\text{eff}}^{h} \cdot [1 - \mathrm{f}_{\text{e in V}}(E, E_F{}^{h}, T)] \approx N_{\text{eff}}^{h} \cdot \exp -\frac{E_F{}^{h} - E_V}{kT}$$

- The red or blue triangles above symbolize the density of electrons in the conduction band or holes in the valence band, respectively, as before.

The right-hand side is identical (of course) to what we had above and shows a kind of symmetry not contained in the formulation with the Fermi distribution, where we have $\mathbf{f}(E, E_F{}^h, T)$ and $\mathbf{1 - f}(E, E_F{}^h, T)$.

- This can be remedied easily by simply setting $\mathbf{1 - f}(E, E_F{}^h, T) =: \mathrm{f}_{\mathrm{h\ in\ V}}(E, E_F{}^h, T)$ with $\mathrm{f}_{\mathrm{h\ in\ V}}$ being the probability of finding holes on the available states in the valence band.
- This is the curve shown on the right-hand side in the picture above.

If we use that definition, we obtain more symmetry at the cost of more heavily indexed functions. It's a matter of taste.

- However, later we will encounter situations where proper bookkeeping of electrons and holes is complicated and essential. Then it might be easier to keep the situation symmetric, to use $\mathrm{f}_{\mathrm{h\ in\ V}}$ for the holes in the valence band, and to express *all* carrier densities in the valence band with $\mathrm{f}_{\mathrm{h\ in\ V}}$, while in the conduction band we use $\mathrm{f}_{\mathrm{e\ in\ C}}$.

2.3.3 Shockley-Read-Hall Recombination

- In this chapter we take a closer look at the generation and recombination of carriers. Even the simple treatments given so far – cf. the formulas for the **p-n** junction – made it clear that *generation and recombination are the major parameters that govern device characteristics and performance*.
  - First, we will treat in more detail the band-to-band recombination in *direct semiconductors*, next the recombination via defects in indirect semiconductors, and for this we introduce and use the "**Shockley -Read- Hall Recombination**" or **SRH** model.
  - However, we will just sort of scratch the subject. In an advanced module some finer points to recombination are treated; here we will stick to fundamentals.
- First a few basic remarks. Generally, we do not only have to maintain energy and momentum conservation for any generation/recombination process, we also have to assure that we keep the minimum of the free enthalpy, or in other words, we have also to consider the **entropy** of these processes. These requirements transform into the conditions
- **1.** $\underline{\boldsymbol{k}} - \underline{\boldsymbol{k}}' = \underline{\boldsymbol{g}}$ as an expression of the (crystal) *momentum conservation*.
- **2.** $E^{e} - E^{h} = \Delta E^{\text{something else}}$ for energy conservation.
  - We have $E_C - E_V = \Delta E^{\text{something else}}$ because the electrons and holes recombining are always close to the band edges for *energy conservation*.
  - $\Delta E^{\text{something else}}$ refers to the unavoidable condition that "something else" has to *provide* the energy needed for generation, or must *take away* the energy released during recombination.
- **3.** Now we look at the *entropy*. Recombination *reduces* the entropy of the system (empty bands are more orderly than bands with a few wildly moving holes and electrons). The "something else" that takes energy out of the system may in addition take some entropy out of it, too. However, no easy law can be formulated.
- The first two points determine if a recombination/generation event – which we from now on are going to call an **R/G**-event – can take place *at all*, i.e. if it is allowed; the third point comes in – in principle – when we discuss the *probability* of an allowed **R/G**-event to take place. This insight, however, will only be used in an indirect way in what follows.
- The major quantities are the **recombination rate *R*** and the **generation rate *G***.
  - The recombination rate ***R*** is the more important one of the two. It is related to the carrier density $n_{e,h}$ by

$$\boxed{\frac{dn_{e,h}}{dt} = -R}$$

  - It is *always* directly given by the rate at which the carrier density *decreases* (the minus sign thus makes ***R*** a positive quantity) and it does not matter which carrier type we are looking at because $dn_e/dt = dn_h/dt$ as long as the carriers disappear in pairs by recombination.
  - Note that the equilibrium condition of constant carrier density does *not* mean that there is no dynamics anymore in the charge carrier population (i.e. that the carriers remain where they are): When ***n*** remains constant, this just means that as many carriers recombine as are generated, since ***n*** is an average quantity. (That this is not unlike the drift velocity of electrons which can be zero despite large thermal velocities of the individual electrons.)

Recombination and Generation in Direct Semiconductors

- If we first look at *recombination in direct semiconductors*, we need holes and electrons at the *same position in the band diagram*; i.e. in ***k**-space*. However, that does not imply that they are at the *same position in real space*. For a recombination event they have to find each other; i.e., we also need them to be at about the same position in real space.
  - The recombination rate ***R*** thus must be proportional to the two densities, n_e and n_h, because the probability of finding a partner scales with the carrier density. We thus can write down the recombination rate ***R*** as

$$\boxed{R = r \cdot n_e \cdot n_h = r \cdot N_{\text{eff}}^{\,h} \cdot N_{\text{eff}}^{\,e} \cdot \exp -\frac{E_C - E_F^{\,e}}{kT} \cdot \exp -\frac{E_F^{\,h} - E_V}{kT}}$$

  - With ***r*** = proportionality constant, having the dimensions of volume/time; we will come back to this later. We also only assume only *local* equilibrium as evidenced by the use of quasi Fermi energies.
- We can rewrite this equation as follows

$$R \;=\; r \cdot N_{\text{eff}}^{\,h} \cdot N_{\text{eff}}^{\,e} \cdot \exp \frac{-E_C + E_F^{\,e} - E_F^{\,h} + E_V}{kT} \;=\; r \cdot N_{\text{eff}}^{\,h} \cdot N_{\text{eff}}^{\,e} \cdot \exp - \frac{E_C - E_V}{kT} \cdot \exp - \frac{E_F^{\,h} - E_F^{\,e}}{kT}$$

- Using our old relation for the intrinsic carrier density $\boldsymbol{n_i}$

$$n_i^{\,2} \;=\; N_{\text{eff}}^{\,h} \cdot N_{\text{eff}}^{\,e} \cdot \exp - \frac{E_C - E_V}{kT}$$

- we finally obtain

$$R \;=\; r \cdot n_i^{2} \cdot \exp \frac{E_F^{\,e} - E_F^{\,h}}{kT}$$

- Note again that we have not invoked *total* equilibrium, but only *local* equilibrium in the bands – we use the quasi Fermi energies $\boldsymbol{E_F^{\,e,h}}$. That is essential; after all it is recombination and generation that restore equilibrium between the bands and the **SRH** theory only makes sense for non-equilibrium.

- If we were to consider *total* thermal equilibrium, we know that the generation rate ***G*** must be identical to the recombination rate ***R***; both quasi Fermi energies are identical (= $\boldsymbol{E_F}$) and $\boldsymbol{R = r \cdot n_i^2}$ applies.
  - Note that we did *not* assume intrinsic conditions; the Fermi energy can have any value, i.e. the semiconductor may be doped.
  - In essence, we see the following:

The recombination rate in non-equilibrium depends very much on the actual carrier density!

- So far it was easy and straigth-forward. Now comes an important point.
  - In *contrast to the recombination rate* ***R***, the **generation rate *G*** does *not* depend (very much) on the carrier density; it is just a reflection on the thermal energy contained in the system and therefore *pretty much constant*. In other words, under most conditions we have

$$G = G_{\text{therm}} \;\approx\; \text{constant} \neq R(n)$$

  - We may, from the above consideration, equate ***G*** under *all* conditions with the recombination rate for equilibrium, i.e.

$$G = G_{\text{therm}} \;=\; r \cdot n_i^{2}$$

- In non-equilibrium, which will be the normal case for devices under operation, the *difference* **(*R* – *G*)** is no longer zero, but has some value

$$U \;=\; R \;-\; G$$

  - Since ***R*** is mostly (but not always) larger than ***G*** under non-equilibrium conditions, ***U*** is the *net* rate of recombination (or, on special occasions, the *net* generation rate).
  - Using the expressions derived so far, we obtain

$$U \;=\; R \;-\; G_{\text{therm}} \;=\; r \cdot (n_e \cdot n_h - n_i^{2}) \;=\; r \cdot n_i^{2} \cdot \left(\exp \frac{E_F^{\,e} - E_F^{\,h}}{kT} \;-\; 1 \right)$$

  - This equation tells us, for example, how fast a non-equilibrium carrier density will decay, i.e. how fast full equilibrium will be reached, or, if we keep the non-equilibrium density fixed for some reason, what kind of **recombination current** we must expect.

- This is so because U, the difference between recombination and generation, *times* the charge is nothing but a net *current flowing from the conduction band to the valence band* (for positive U).

Determining the Proportionality Constant *r*

- We still need to determine the proportionality constant r.
  - This is not so easy, but we can make a few steps in the right direction. We assume in a purely classical way that an electron (or hole) moves with some average velocity **v** through the lattice, and *whenever it encounters a hole (or electron), it recombines*.
  - The problem is the word "*encounters* ". If the particles were to be small spheres with a diameter d_p, "encountering" would mean that parts of such a sphere would be found in the cylinder with diameter d_p formed by another moving sphere because that would cause a physical contact.
  - Our particles are not spheres, but for the purpose of scattering theory we treat them as such, except that the diameter of the cylinder that characterizes its "scattering size" is called **scattering cross section** σ and has a numerical value that need *not* be identical to the particle size.
- *One* electron now covers a volume **v** · σ per second and *all* N_e electrons of the whole sample (a number, not a density) probe per second the volume

$$\boxed{V_{\text{probed}} = N_e \cdot \mathbf{v} \cdot \sigma}$$

- Any time an electron encounters a hole in the volume it probes, it recombines. The absolute recombination rate R_{abs} then is simply the number of encounters per second, occurring in the whole sample.
  - How many holes are "hit" per second? In other words, how many are to be found in the volume probed? That is easy: The *number* N_h of holes encountered in the volume probed by electrons, and thus the recombination rate is

$$\boxed{N_h = n_h \cdot V_{\text{probed}} = n_h \cdot N_e \cdot \mathbf{v} \cdot \sigma = R_{\text{abs}}}$$

  - Here, n_h is simply the *density* of holes in the sample. You many wonder if this is correct, considering that the holes move around, too, but simply realize that the density of holes is nevertheless constant.
- The formula is a bit unsatisfying, because it contains the volume *density* of holes, but the *absolute* number of electrons.
  - That is easily remedied, however, if we express N_e, the *number* of electrons, by their *density* n_e via $N_e = n_e \cdot V$ with V = sample volume. Using the latter for normalizing the absolute recombination rate to the sample volume, this gives us

$$\boxed{R = \frac{R_{\text{abs}}}{V} = \begin{matrix}\textbf{Recombinations per}\\ \textbf{s and cm}^{3}\end{matrix} = n_e \cdot n_h \cdot \mathbf{v} \cdot \sigma}$$

  - In other words, if we use the *density* of the electron and holes, we obtain a recombination rate *density* , i.e. recombination events per **s** and per $\mathbf{cm^3}$ – as it should be. *As always, we are going to be a bit sloppy about keeping densities and numbers apart. But there is no real problem: Just look at the dimensions you get, and you know what it is.*
- A comparison with the formula from above yields

$$\boxed{r = \mathbf{v} \cdot \sigma}$$

- This leaves us with finding the proper value for σ. Whereas this is difficult (in fact, the equation above is more useful for determining σ from measurements of R than to calculate r), we are still much better off than with r alone:
  - Whereas we had *no idea* about a rough value for r, we do know something about **v** (it is the group velocity of the carriers considered), and we know at least the rough order of magnitude for σ: We would expect it to be in the general range of atomic dimensions (give or take an order of magnitude).
- You might wonder now why we assume that any "meeting" of the elctrons and holes leads to recombination, given that we have to preserve momentum, too. You are right, but remember:
  - We are treating *direct* semiconductors here! Since we only consider the mobile electrons and holes, we only consider the ones at the band edges – and those have the same **k**-vector in the reduced band diagram!

Useful Approximations and the Lifetime τ

- We now consider non-equilibrium, but describe it in terms of deviations from equilibrium. Then it is sensible to rewrite the carrier densities (or numbers, *take whatever you like*) in terms of the equilibrium density $\boldsymbol{n_{e,h}}$**(equ)** *plus/minus some delta*:

$$\boxed{n_{e,h} \;=\; n_{e,h}(\text{equ}) \;+\; \Delta n_{e,h}}$$

  - This is one of the decisive "tricks" to get on with the basic equations, because it permits to specify particular cases (e.g. $\Delta n_{e,h} \ll n_{e,h}(\text{equ})$ or whatever), and then resort to approximations. We will encounter this "trick" fairly often.

- We obtain after some shuffling of the terms form the equation for the net recombination rate ***U***

$$\boxed{U \;=\; v\cdot\sigma\cdot\Big(\{n_e(\text{equ}) + \Delta n_e\}\cdot\{n_h(\text{equ}) + \Delta n_h\} - n_e(\text{equ})\cdot n_h(\text{equ})\Big)}$$

- So far everything is still correct. But now we consider a first *special*, but still rather general case:
  - We assume that $\Delta n_e = \Delta n_h = \Delta n$, i.e. that only additional electron–hole *pairs* were created in non-equilibrium. We then may simplify the equation to

$$\boxed{U \;=\; r\cdot\Big(\Delta n\cdot\{n_e(\text{equ}) + n_h(\text{equ})\} + \Delta n^2\Big)}$$

  - We can simplify even more. For the *extrinsic* case where *one* carrier density – let's say for example n_h – is far larger than n_e or Δn (i.e. we have a **p**-doped semiconductor), we may neglect the terms $\Delta n\cdot n_e(\text{equ})$ and Δn^2 and obtain

$$\boxed{U \;\approx\; r\cdot n_h\cdot\Delta n}$$

- ***U*** was the difference between the recombination and the generation rate. We are now looking at an approximation where only some Δn in the density of the minority carriers is noticeably different from equilibrium conditions (where we always have $U = 0$).
  - We thus may write

$$\boxed{U \;=\; R(\text{equ}) + R(\Delta) - G(\text{equ}) = R(\Delta)}$$

  - Here, $R(\Delta)$ denotes the *additional* recombination due to the **excess minorities**. Remembering the basic definition of ***R*** we see that now we have

$$\boxed{\frac{d(\Delta n_e)}{dt} \;=\; -U \;=\; -r\cdot n_h\cdot\Delta n_e \;=\; -v\cdot\sigma\cdot n_h\cdot\Delta n_e}$$

- This is a differential equation for $\Delta n_e(t)$, it has the simple solution

$$\boxed{\Delta n_e(t) \;=\; \Delta n_e(t=0)\cdot\exp -\frac{t}{\tau}}$$

  - The quantity demanded by the general solution is, of course, the **life time** of the minority carriers. We now have a formula for this prime parameter, it comes out to be

$$\boxed{\tau = \frac{1}{v\cdot\sigma\cdot n_h} \;=\; \frac{1}{v\cdot\sigma\cdot n_{maj}}}$$

- The last equality generalizes for both types of carriers – it is always the density of the *majority* carriers that determine the lifetime of the *minority* carriers. This is clear enough considering the "hit and recombine" scenario that we postulated at the beginning

- Substituting $\boldsymbol{r \cdot n_{\mathrm{h}}}$ with $\mathbf{1/\tau}$ in the equation for $\boldsymbol{U}$ yields

$$U = \frac{\Delta n}{\tau}$$

  - In other words: The recombination rate in excess of the recombination rate in equilibrium is simply given by the excess density of minority carriers divided by their life time.

- In yet other words:
  - The *net current* flowing from the band containing the minority carriers to the other band is given by $\boldsymbol{U}$ (times the elementary charge, of course, and times the total sample volume), because $\boldsymbol{U}$ gives the net amount of carriers "flowing" from here to there! And that is the definition of a current!

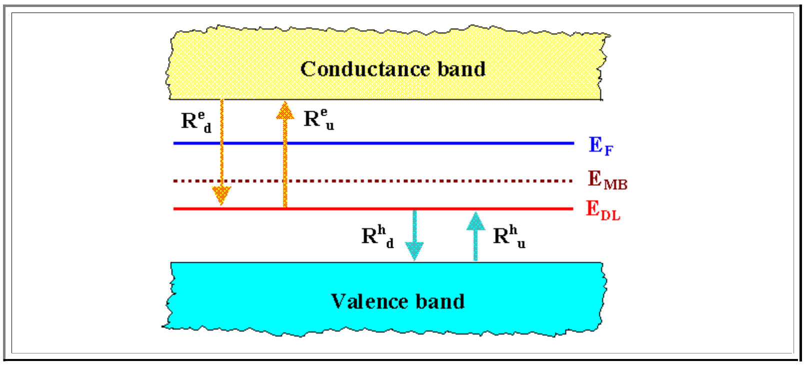
- This result not only justifies our earlier approach, it gives us the *minority carrier life time* in more basic quantities including (at least parts) of its temperature dependence via the thermal velocity **v** and the majority carrier density $\boldsymbol{n_{\mathrm{h}}}$ – the $\boldsymbol{T}$-dependence of which we already know.
  - Since $\mathbf{1/}\boldsymbol{n_{\mathrm{h}}}$ is more or less proportional to the resistivity, we expect τ to increase linearly with the resistivity which it does as illustrated before, at least for resistivities that are not too low.
  - A rough order of magnitude estimate gives indeed a good value for many *direct* semiconductors:

$$\begin{aligned} \sigma &\approx 10^{-15}\,\mathrm{cm}^2 \\ \mathrm{v} &= 10^{7}\,\mathrm{cm/s} \qquad \Rightarrow \qquad \tau \approx 10^{-9}\,\mathrm{s} = 1\,\mathrm{ns} \\ n_{\mathrm{h}} &= 10^{17}\,\mathrm{cm}^{-3} \end{aligned}$$

Recombination and Generation in Indirect Semiconductors

- In *indirect* semiconductors, *direct* recombination is theoretically impossible or, being more realistic, very improbable.
  - In general, a recombination event needs a *third partner* to provide conservation of energy and crystal momentum.
  - Under most (but not all) circumstances, this third partner is a *lattice defect*, most commonly an impurity atom, with an energy state "*deep*" in the band gap, i.e. not close to the band edges.
  - Recombination then is determined by these "**deep states**" or **deep levels**, and is no longer an intrinsic or just doping dependent property.

- How the recombination and generation depends on the *properties of deep levels* is the subject of the proper Shockley–Read–Hall theory (what we did so far was just a warming-up exercise). It is a lengthy theory with long formulas; here we will just give an outline of the important results. More topics will be covered in an even more advanced module.
  - First we look at the situation in a band diagram that shows the relevant energy levels plus the mid-band energy level $\boldsymbol{E_{\mathrm{MB}}}$, which will come in handy later on.



- Besides the energy level of the "deep level" defect, we now need *four* **transition rates** instead of just *one* recombination rate:
  - $R^e{}_d$, the rate with which *electrons* from the conductance band transit from the conduction band to the deep level, or more simply put, *occupy* the deep level with the energy E_{DL} – in short: the rate with which they are going *down* to the deep level.
  - $R^e{}_u$, the rate with which *electrons* occupying the deep level state go *up* to the conduction band.
  - $R^h{}_u$, the rate with which *holes* from the valence band go *up* to the deep level – or better: the *electrons* in the deep level go *down* to the valence band, and finally
  - $R^h{}_d$, the rate with which *holes* from the deep level go *down* to the valence band – or better: *electrons up* to the deep level.

The equilibrium density of electrons (and holes) on the deep level is, as always, given by the Fermi distribution. We have

- $n^-{}_{DL} = N_{DL} \cdot \mathrm{f}(E_{DL}, T)$ = density of negatively charged deep levels with *one* electron sitting on it, and
- $n^0{}_{DL} = N_{DL} \cdot [1 - \mathrm{f}(E_{DL}, T)]$ = density of deep levels with *no* electron sitting on it. N_{DL} , of course, is the *density* of deep level states, e.g. the density of impurity atoms. It's written with capital ***N*** (otherwise used for absolute numbers) to avoid confusion with the carrier densities.
- To make life easier, we assumed that the deep level is normally neutral, i.e. does not contain an unalterable fixed charge, and it can only accommodate one additional electron.

We may now write down formulas for the transition rates in direct analogy to the consideration of the recombination rate in direct semiconductors as given above. For $R^e{}_d$ we have

$$\boxed{R^e{}_d \;=\; r \cdot n_e \cdot n^0{}_{DL} \;=\; v \cdot \sigma^e \cdot n_e \cdot N_{DL} \cdot [1 - \mathrm{f}(E_{DL}, T)]}$$

- With σ^e = **scattering cross section** (also called **capture cross section**) of the deep level for electrons.

For the other transition rate $R^e{}_u$ we have to think a little harder. For the electron *trapped* at the deep level to go up to the conduction band it needs a free place up there, hence:

$$\boxed{R^e{}_u \;=\; r' \cdot (N_{\mathrm{eff}}{}^e \;- n_e) \cdot n^-{}_{DL}}$$

- With r' = some proportionality constant, *principally different* from r, and $N_{\mathrm{eff}}{}^e - n_e$ = density of free places (which, please remember, aren't *holes*!) in the conduction band.
- Since n_e is much smaller than $N_{\mathrm{eff}}{}^e$, we may approximate this equation by

$$\boxed{R^e{}_u \;\approx\; r' \cdot N_{\mathrm{eff}}{}^e \cdot N_{DL} \cdot \mathrm{f}(E_{DL}, T)}$$

We have not invoked some cross section and thermal velocity here, because the electron now is localized and doesn't move around. We also used a different proportionality constant r' because the situation is not fully symmetric to the reverse process. It is common to call the quantity $e^e = r' \cdot N_{\mathrm{eff}}{}^e$ the **emission probability** for electrons from the deep level.

- The *emission probability* contains the information about the generation of carriers from the deep level; in this it is comparable to the *generation rate* from the valence band for the simple recombination theory considered above.

Now , if we assume that the transitions of conduction band electrons to the deep level and their re-emission to the conduction band are in *local* equilibrium (which does not necessarily entail *total* equilibrium), we have $R^e{}_u = R^e{}_d$.

- From this we get – *after a minimal shuffling of the terms* – for the emission probability e^e in *local* equilibrium:

$$\boxed{e^e \;=\; \frac{v \cdot \sigma^e \cdot n_e \cdot [1 - \mathrm{f}(E_{DL}, T)]}{\mathrm{f}(E_{DL}, T)}}$$

Again, as in the case of the generation rate ***G*** for direct semiconductors, we may assume *that the emission probability* e^e *is pretty much constant* and this is a crucial point for what follows.

Since we want to find quantities like life times as a function of the density and energy level of the deep level, it is useful to use the mid-band energy level as a reference, and to rewrite the equation for e^e in terms of this mid-band level E_{MB} via the relations

$$\boxed{\begin{gathered}\frac{1\;-\;\mathrm{f}(E_{\mathrm{DL}},T)}{\mathrm{f}(E_{\mathrm{DL}}\,,T)} = \exp - \frac{E_{\mathrm{F}}\;-\;E_{\mathrm{DL}}}{kT}\\[2ex] E_{\mathrm{MB}} \;=\; \frac{E_{\mathrm{C}}-E_{\mathrm{V}}}{2}\\[2ex] n_{\mathrm{i}} \;=\; N_{\mathrm{eff}}^{\mathrm{e}}\cdot\exp - \frac{E_{\mathrm{C}}-E_{\mathrm{MB}}}{kT}\end{gathered}}$$

- These equations may need a little thought. The first one came up before in a similar way, the second simply defines mid band-gap, and the last one uses the fact that the Fermi energy for intrinsic semiconductors is in mid band gap (at least in a good approximation).

Using these equations, we first rewrite the formula for the density of electrons in the conduction band and obtain

$$\boxed{n_{\mathrm{e}} \;=\; N_{\mathrm{eff}}^{\mathrm{e}}\cdot\exp - \frac{E_{\mathrm{C}}-E_{\mathrm{F}}}{kT} = N_{\mathrm{eff}}^{\mathrm{e}}\cdot\exp - \frac{E_{\mathrm{C}}-E_{\mathrm{MB}}}{kT}\cdot\exp - \frac{E_{\mathrm{MB}}-E_{\mathrm{F}}}{kT} \;=\; n_{\mathrm{i}}\cdot\exp\frac{E_{\mathrm{F}}-E_{\mathrm{MB}}}{kT}}$$

- Putting everything together, we get for the emission probability

$$\boxed{e^{\mathrm{e}} \;=\; v\cdot\sigma^{\mathrm{e}}\cdot n_{\mathrm{i}}\cdot\exp\frac{E_{\mathrm{DL}}-E_{\mathrm{MB}}}{kT}}$$

- This is the best we can do to describe the traffic of electrons between the deep level and the conduction band.

Next, we do the matching calculation for the *transitions rates with the valence band*, $\boldsymbol{R}^{\mathrm{h}}{}_{\mathrm{u}}$ and $\boldsymbol{R}^{\mathrm{h}}{}_{\mathrm{d}}$.

- Except, *we* won't do it. Too boring – everything is quite similar. As the final result for the ***emission probability for the holes***, $\boldsymbol{e}^{\mathrm{h}}$, we obtain exactly what we should expect anyway:

$$\boxed{e^{\mathrm{h}} \;=\; v\cdot\sigma^{\mathrm{h}}\cdot n_{\mathrm{i}}\cdot\exp\frac{E_{\mathrm{MB}}-E_{\mathrm{DL}}}{kT}}$$

The Net Interband Recombination

We captured the electron and hole traffic betwen a deep level and the conduction or valence band, respectively, with these equations – always for *local* equilibrium of the deep level with the respective band. Now we consider the *interband* generation and recombination rates, $\boldsymbol{G}$ and $\boldsymbol{R}$.

- This is exactly the same thing as the money traffic form one major bank to another one via an intermediate bank. Each bank can deposit and withdraw money from all three accounts, while the total amount of all the money must be kept constant. If it would be *your* money, you sure like hell would want to and be able to keep track of it. So let's do it with electrons and holes, too.

With $\boldsymbol{G}$ we still denote the rate of electron–hole *pair generation* taking place directly between the bands; by thermal or other energies, e.g. by illumination. It is thus the rate with which electrons and holes are put directly into the conduction or valence band, *no matter what goes on between the deep level and the bands*.

- We may, for some added clarity, decompose $\boldsymbol{G}$ into $\boldsymbol{G}_{\mathrm{perfect}}$, the generation always going on even in a hypothetical perfect semiconductor, and $\boldsymbol{G}_{\mathrm{ne}}$ for whatever is added in non-equilibrium (e.g. the generation by light). We have $\boldsymbol{G} = \boldsymbol{G}_{\mathrm{perfect}} + \boldsymbol{G}_{\mathrm{ne}}$.
- After all, before we put in "*our* " deep levels or switched on the light, the hypothetically perfect crystal already must have had some generation and recombination, too (for which $\boldsymbol{R}_{\mathrm{perfect}} = \boldsymbol{G}_{\mathrm{perfect}}$ must hold). However, we can expect that $\boldsymbol{R}_{\mathrm{perfect}}$ is rather small in a perfect indirect semiconductor, which makes $\boldsymbol{G}_{\mathrm{perfect}}$ rather small, too.

- The rate of change of the electron and hole density in their bands is then the *sum total* of all processes withdrawing and depositing electrons or holes, i.e.

$$\frac{\mathrm{d}n_e}{\mathrm{d}t} = G_{\text{perfect}} + G_{ne} - R_{\text{perfect}} + R^e{}_u - R^e{}_d = G_{ne} - (R^e{}_d - R^e{}_u)$$

$$\frac{\mathrm{d}n_h}{\mathrm{d}t} = G_{\text{perfect}} + G_{ne} - R_{\text{perfect}} + R^h{}_d - R^h{}_u = G_{ne} - (R^h{}_u - R^h{}_d)$$

  - Note that $G_{\text{ perfect}} - R_{\text{ perfect}} = 0$ by definition.

- *Local* equilibrium between the bands and the deep level, still not necessarily implying *total* equilibrium, now demands that both $\mathrm{d}n_e/\mathrm{d}t$ and $\mathrm{d}n_h/\mathrm{d}t$ must be zero.
  - That means that the density of electrons in the conduction band and the density of the holes in the valence band do not change with time anymore. However, that does not mean that they have their *global* equilibrium value, only that we have a so-called **steady state** (in global non-equilibrium) which, on the time scales considered, appears to keep things at a constant value.
  - As an example, a piece of semiconductor under constant illumination conditions will achieve a *steady state* in global non-equilibrium conditions. The carrier densities in the bands will be constant, but not at their equilibrium values if light generates electron–hole pairs all the time.
  - This gives us the simple equation

$$R^e{}_d - R^e{}_u = G_{ne} = R^h{}_u - R^h{}_d$$

- Essentially, this says that the total electron or hole traffic or current [= difference of the partial rates (times elementary charge)] from the conduction or valence band, respectively, to the deep level are identical and equal to the extra band-to-band generation current produced in non-equilibrium for the given material and situation.
  - But steady state also implies that there must be an additional recombination exactly equal to G_{ne} and that is of course exactly what the terms $R^e{}_d - R^e{}_u$ or $R^h{}_u - R^h{}_d$ denote: They are identical to the additional recombination rates needed for balancing the additional generation G_{ne}, or simply
  - We thus have

$$R^e{}_d - R^e{}_u = R - R_{\text{perfect}} = R - G_{\text{perfect}} =: U_{DL}$$

  - The quantity U_{DL} is exactly analogous to the *difference* **($R - G$)** defined for direct semiconductors.

- U_{DL} is also the difference between the recombination to a deep level and the emission from it. For the example considered so far (additional generation via illumination) it must be positive, there is more recombination than generation
  - However, our treatment is completely general; U_{DL} can have any value – if it is negative, we would have more generation via deep levels than recombination.
  - Of course, U_{DL} makes only sense for global *non-equilibrium* conditions, because for global equilibrium U_{DL} must be zero!

- All we have to do now is to express the R^e's with the formulas from above. Inserting the equations for the various R's, the emission probabilities, and setting $\sigma^e = \sigma^h = \sigma$ for the sake of simplicity, we get, after some shuffling of the terms, the *final equation*

$$U_{DL} = \frac{v \cdot \sigma \cdot N_{DL} \cdot (n_e \cdot n_h - n_i^2)}{n_e + n_h + 2n_i \cdot \cosh \dfrac{E_{DL} - E_{MB}}{kT}}$$

  - The *cosh* (= hyperbolic cosine) comes from the sum of the two exponential functions. Its value is **1** for $E_{DL} = E_{MB}$; it increases symmetrically for deviations of E_{DL} from the mid-level energy E_{MB}.
  - *A chain hanging down from two posts has exactly a* ***cosh(x)*** *shape* – that's the way to memorize the general shape of a **cosh** curve. If you want to look more closely at the **cosh** function, activate the link.

The equation for U_{DL} is quite similar to the one we had for direct semiconductors, as far as the numerator is concerned. We will explore a little more what it implies.

- For *global* equilibrium, the mass action law $n^e \cdot n^h = n_i^2$ applies, and $U_{DL} = 0$. In other words, there is no *net* recombination, i.e. recombination *in excess* of what is always going on.
- Without deep levels $U_{DL} = 0$! The recombination rate then is fixed and simply $R_{perfect}$.
- The recombination rate – everything else being constant – is directly proportional to the *density of the deep levels* and their scattering cross section (or *capture cross section* as it is called in this case).
- Since the recombination rate is *highest for deep levels exactly in mid-band* (look at the **cosh** function), defects with levels near mid-band are more efficient in recombining carriers than those with levels farther off the mid-band position.

Approximations and the Lifetime τ

As before, let's look at some special case. Again, we write the carrier densities as $n_{e,h} = n_{e,h}(\text{equ}) + \Delta n$ assuming equal Δ 's for electron and holes.

- This gives us

$$U = v \cdot \sigma \cdot N_{DL} \cdot \frac{[n_e(\text{equ}) + \Delta n] \cdot [n_h(\text{equ}) + \Delta n] - n_i^2}{n_e(\text{equ}) + n_h(\text{equ}) + 2\,\Delta n + 2n_i \cdot \cosh[(E_{DL} - E_{MB})/(kT)]}$$

- Looking at a **p**-doped semiconductor and only considering the large densities n_h as in the example before, we obtain

$$U = \frac{v \cdot \sigma \cdot N_{DL} \cdot n_h(\text{equ}) \cdot \Delta n_e}{n_h(\text{equ}) + 2n_i \cdot \cosh[(E_{DL} - E_{MB})/(kT)]}$$

- Since n_i is also much smaller than $n_h(\text{equ})$, we may neglect the whole **cosh** term, too – as long as $\cosh[(E_{DL} - E_{MB})/(kT)]$ is not large, i.e. for deep levels around mid-band.
- As a consequence, $n_h(\text{equ})$ cancels and we are left with

$$U = v \cdot \sigma \cdot N_{DL} \cdot \Delta n_e$$

Again, as before, the change in excess minority carrier density is given by $d(\Delta n_e)/dt = -U$, giving

$$\frac{d(\Delta n_e)}{dt} = -v \cdot \sigma \cdot N_{DL} \cdot \Delta n_e$$

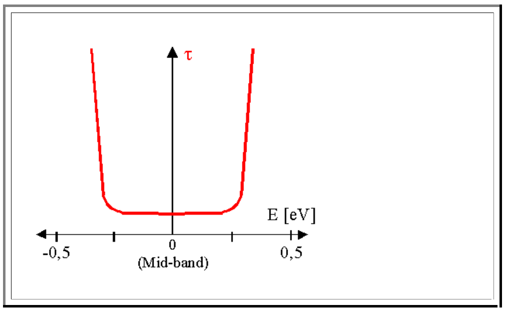
- The solution of the differential equation now becomes trivial and we have

$$\Delta n_e(t) = \Delta n_e(t = 0) \cdot \exp -\frac{t}{\tau}$$

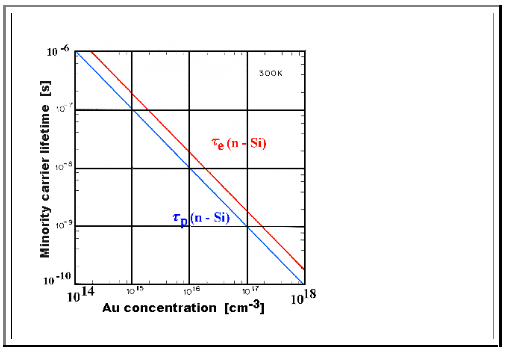
- with **τ = minority life** time or better **recombination life time** in *indirect* semiconductors defined by

$$\tau = \frac{1}{v \cdot \sigma \cdot N_{DL}}$$

- This is the same equation as before *except that the density of the majority carriers (holes in the valence band for the example) now is replaced by the density of (mid-band) deep levels*.
  - That this formula is a useful approximation is shown in the two illustrations below:



Dependence of the life time on the deep level position relative to the mid level – it is fairly constant (and small) as long as the deep level is approximately in mid band.



Dependence of life time on deep level density – it is linear as predicted. (The red curve is for p-Si, of course.)

- The picture on the right illustrates a sad fact hidden in all these equations: it doesn't take *much* dirt (or **contamination**, to use the proper word) to considerably degrade the life time. Interstitial gold atoms obviously are felt at **10^{14} cm^{-3}** , i.e. at concentrations well below **ppb**.
  - More to Shockley–Read–Hall recombination can be found in an advanced module.

2.3.4 Useful Relations

- There is no way to cover all relevant semiconductor physics within the scope of this course. This subchapter provides some important or useful relations needed for the understanding of the topics.
  - It also serves as the "gate" to a number of modules providing additional information.
  - This subchapter therefore is more open than the other ones; it will fill out and sprout a network in connection with the lecture course that cannot be predicted by now.

Einstein Relation

- We have encountered the Einstein relation before. It is of such fundamental importance that we give *two* derivations: one in this paragraph, another one in an advanced module.
- First, we consider the internal current (density) in a material with a *gradient* of the carrier density (n_e or n_h).
  - Fick's first law then tells us that the diffusion-driven *particle* current $j_{p,diff}$ is given by

$$\boxed{j_{p,diff} \;=\; -\,D_{e,h} \cdot \nabla\, n_{e,h}}$$

  - If the particles are carrying a charge q, this *particle* current is also an **electrical current** (which obviously is a **diffusion** current, then), given by

$$\boxed{j_{e,h} \;=\; q \cdot j_{p,diff} \;=\; -\,q \cdot D_{e,h} \cdot \nabla\, n_{e,h}}$$

  - Considering only the one-dimensional case for electrons (i.e. $q = -\mathrm{e}$; holes behave in exactly the same way with $q = +\mathrm{e}$), we have

$$\boxed{j_e\,(x) \;=\; \mathrm{e} \cdot D_e \cdot \frac{\mathrm{d}n_e(x)}{\mathrm{d}x}}$$

- Since there can be no *net* current in a piece of material just lying around (which nevertheless might still have a density gradient in the carrier density, e.g. due to a gradient in the doping density), the carriers displaced by diffusion always generate an electrical field that will drive the other carriers back.
  - Any field $E(x)$ (written in mauve to avoid confusion with energies) now will cause a (so far one-dimensional) current given by

$$\boxed{j \;=\; \sigma \cdot E(x) \;=\; q \cdot n(x) \cdot \mu \cdot E(x)}$$

  - With σ = conductivity, μ = mobility.
  - Note that the result is always the technical current density, which is positive for positive charge carriers. Yet this equation also works for electrons because for them, effectively, two minus signs cancel: one from their negative charge and the other from their direction of movement opposite to the electric field. This means that in a strict sense, their mobility should be negative. However, in this equation one only considers positive charges and positive mobilities – also for electrons. Therefore, to use ths equation in full generality, we write it as

$$\boxed{j \;=\; \mathrm{e} \cdot n(x) \cdot \mu \cdot E(x)}$$

- The *total* (one-dimensional) current in full generality is then

$$\boxed{j_{total}(x) \;=\; \mathrm{e} \cdot n(x) \cdot \mu \cdot E(x) \;-\; q \cdot D \cdot \frac{\mathrm{d}n(x)}{\mathrm{d}x}}$$

  - We will need this equation later.

- For our case of *no net current* and only *fields caused by the diffusion current*, both currents have to be equal in magnitude:

$$e \cdot n(x) \cdot \mu \cdot E(x) \;=\; q \cdot D \cdot \frac{dn(x)}{dx}$$

  - This is an equation that comes up repeatedly; we will encounter it again later when we derive the Debye length.
  - Note that in this equation, the sign on the right-hand side depends on the type of charge carriers (since $q = \pm e$). This is balanced on the left-hand side by the direction of the electric field.

- Now we are stuck. *We need some additional equation in order to find a correlation between **D** and **μ***.
  - This equation is the *Boltzmann distribution* (here used as an approximation to the *Fermi distribution*), because we have *equilibrium* in our material.
  - However, we also know that, in this equilibrium situation, we have spatially varying charge carrier densities and electric fields. We know such a situation from the p–n junction in equilibrium. There, this was only possible due to the band bending, i.e. that the band edges were functions of the lateral position. This we also consider here.
  - Just to derive the relation to the electric field in the above equation, for the moment we just consider the case of electrons as majority carriers. For their local density it holds that

$$n(x) \;=\; N_{eff} \cdot \exp - \frac{E_C(x) - E_F}{kT}$$

  - Differentiation of the Boltzmann distribution gives us

$$\frac{dn}{dx} \;=\; - N_{eff} \cdot 1/(kT) \cdot \frac{dE_C(x)}{dx} \cdot \exp - \frac{E_C(x) - E_F}{kT}$$

$$\frac{dn}{dx} \;=\; - n(x) \cdot 1/(kT) \cdot \frac{dE_C(x)}{dx}$$

  - The slope of the conduction band comes directly from the spatially varying electric potential ***V* (*x*)**; to convert the electric potential to the absolute energy of a charge carrier, the elementary charge **e** is needed as an additional factor. From the p–n junction we know that the sign (i.e., the direction) of the electric field is identical to that of the slope of the conduction band. Thus, altogether we have

$$E(x) \;=\; - \frac{dV(x)}{dx} \;=\; 1/e \cdot \frac{dE_C(x)}{dx}$$

  - Using this relation, the current balance from above becomes

$$e \cdot n(x) \cdot \mu \cdot E(x) \;=\; q \cdot D \cdot \frac{dn(x)}{dx}$$

$$e \cdot n(x) \cdot \mu \cdot E(x) \;=\; - q \cdot n(x) \cdot D \cdot e/(kT) \cdot E(x)$$

$$D \;=\; \mu kT/e$$

  - In words: Equilibrium between diffusion currents and electrical currents for charged particles demands a simple, but far reaching relation between the diffusion constant ***D*** and the mobility **μ**.

- Distinguishing again between electrons and holes gives as the final result the famous **Einstein–Smoluchowski relations**:

$$D_e = \frac{\mu_e \cdot kT}{e}$$

$$D_h = \frac{\mu_h \cdot kT}{e}$$

- You may want to have a look at a different derivation in an advanced module.

Non-Equilibrium Currents

- In the consideration above we postulated that there is *no net current flow*; in other words, we postulated *total equilibrium*. Now let's consider that there *is* some *net current flow* and see what we have to change to arrive at the relevant equations.
- In order to be close to applications, we treat the *extrinsic* case and, since we do not assume equilibrium per se, we automatically do not assume that the carrier densities have their *equilibrium* values n_e**(equ)** and n_h**(equ)**, but *arbitrary* values that we can express by some Delta to the equilibrium value. We thus start with

$$n_e = n_e\,(\text{equ}) + \Delta n_e$$

$$n_h = n_h(\text{equ}) + \Delta n_h$$

  - Since carriers above the equilibrium density are often created *in pairs* we have for this *special, but rather common* case

$$\Delta n_e = \Delta\, n_h = \Delta n$$

$$\Delta\, n = n_e - n_e(\text{equ}) = n_h - n_h(\text{equ})$$

  - *This is a crucial assumption!*

- This allows us to concentrate on *one kind of carrier*, let's say we look at **n**-type **Si** with electrons as the majority carriers. We now *focus on holes* as the minority carriers since we always can compute the electron density n_e by

$$n_e = n_e(\text{equ}) + \Delta n_e = n_e(\text{equ}) + n_h - n_h\,(\text{equ})$$

  - We now must consider **Fick's second law** or the **continuity equation** (it is the same thing for special cases, but the continuity equation is more general).

- For the net (mobile) charge density ρ (which is the *difference* of the electron and hole density, $\rho = e \cdot (n_h - n_e)$, in *contrast to the total particle density*, which is the *sum* !) we have

$$\frac{\partial \rho}{\partial t} = -\,\text{div}\,(j_{\text{total}})$$

  - With $j_{\text{total}} = j_e + j_h$ = sum of the electron and hole currents.

- In the simplest form we have for the holes

$$\frac{\partial n}{\partial t} = -\,(1/e) \cdot \text{div}\,(j_h)$$

- The factor **1/e** is needed to convert an electrical current j to a particle current j_{part} via $j = q \cdot j_{\text{part}}$, with $q = \pm e$. Here, as always, we have to pick the right sign for the elementary charge **e** (negative for electrons, positive for holes).

- This is simply the statement that the *charge is conserved*. It would be sufficient that no holes disappear or are created in any differential volume **d*V*** considered, i.e. **div j_h = 0**, to satisfy that condition.

- But this is, of course, a condition that we know *not* to be true.
  - In all semiconductors, we have constant generation and recombination of holes (and electrons) as discussed before. In *in equilibrium*, of course, the generation rate ***G*** and the recombination rate ***R*** are equal, so they cancel each other in a balance equation and need not be considered – since **div j_h = 0** is correct *on average*.
  - We are, however, considering *non-equilibrium* , so we must primarily consider the recombination of the *surplus* minority carriers given by

$$\boxed{\Delta n_h = n_h - n_h(\mathrm{equ})}$$

  - Why? Because, as stated before, the generation essentially does *not* change, so it still balances against the recombination rate of the equilibrium density, and only the recombination rate of the surplus minorities, $R^\Delta = [n_h - n_h(\mathrm{equ})]/\tau$ needs to be considered (τ is the minority carrier life time).
  - $R^\Delta = [n_h - n_h(\mathrm{equ})]/\tau$ is the rate with which carriers disappear by recombination, we thus must *subtract* it from the carrier balance as expressed in the continuity equation, and obtain

$$\boxed{\frac{\mathrm{d}n}{\mathrm{d}t} = -\,\frac{n_h - n_h(\mathrm{equ})}{\tau} - (1/e)\cdot \mathrm{div}\,(j_h)}$$

- The current ***j*** can always be expressed as the sum of a *field* current and a *diffusion* current as we did above by

$$\boxed{j_{h,\mathrm{total}}(x) = e\cdot n(x)\cdot \mu \cdot E_x(x) - e\cdot D_h \cdot \frac{\mathrm{d}n_h(x)}{\mathrm{d}x}}$$

- Inserting this equation in our continuity equation yields

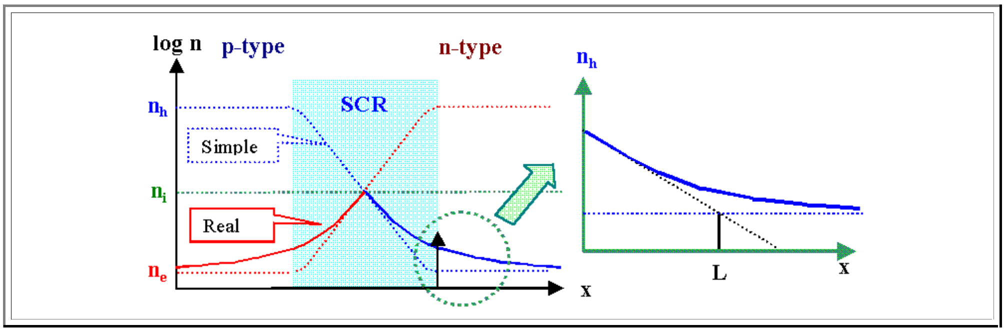
$$\boxed{\frac{\partial n_h(x)}{\partial t} = -\,\frac{n_h(x) - n_h(\mathrm{equ})}{\tau} - n_h(x)\cdot \mu \cdot \frac{\partial E(x)}{\partial x} - E(x)\cdot \mu \cdot \frac{\partial n_h(x)}{\partial x} + D\cdot \frac{\partial^2 n_h(x)}{\partial x^2}}$$

  - This is an *important*, if not so simple equation. It is not so simple, because the electrical field strength ***E*(*x*)** at ***x*** is a function of the carrier density ***n_h*(*x*)** at ***x***, which is what we want to calculate! We have used the symbols for partial derivatives ("∂ ") to emphasize that it is in reality a three-dimensional equation.

- We will now look at some applications of this equation.

Pure Diffusion Currents

- Consider the minority carrier situation just outside of the space charge region of a biased **p–n** junction.
  - If it is forwardly biased, a lot of majority carriers are flowing to the respective other side where they become minority carriers. They will eventually disappear by recombination, but the minority carrier density right at the edge of the space charge region will be larger than in equilibrium and will decrease as we go away from the junction.
  - This is now shown in the illustration used before in the simple model of the **p–n** junction, but the *realistic* minority carrier situation is now included.



The region outside the space charge region, *while now showing a density gradient of the minority carrier density*, is essentially field free or at least has only a small electrical field strength.

- If we let $E_x = 0$ and consequently $\partial E_x(x)/\partial x = 0$, too, the current equation from above reduces to

$$\frac{\partial n_h}{\partial t} = -\frac{n_h - n_h(\text{equ})}{\tau} + D \cdot \frac{\partial^2 n_h(x)}{\partial x^2}$$

- Since $\partial n_h/\partial t = \partial[n_h(\text{equ}) + \Delta n_h]/\partial t = \partial \Delta n_h/\ \partial t$, and correspondingly $\partial^2 n\ _h(x)/\partial x^2 = \partial\ ^2\Delta n_h(x)/\partial\ x^2$, we have

$$\frac{\partial \Delta\ n_h}{\partial t} = -\frac{\Delta\ n_h}{\tau} + D \cdot \frac{\partial^2 \Delta\ n_h(x)}{\partial\ x^2}$$

- If we consider *steady state* , we have $\partial \Delta n_h/\partial\ t = 0$, and the solution of the differential equation is now mathematically easy.
- But how can steady state be achieved in practice? How can we provide for a *constant* , *non-changing* density of minority carriers *above* equilibrium?
  - For example by having a defined source of (surplus) holes at $x = 0$. In the illustration this is the (constant) hole current that makes it over the potential barrier of the **p–n** junction.
  - But we could equally well imagine holes generated by light a $x = 0$ at a constant rate. The surplus hole density then will assume some distribution in space which will be constant after a short initiation time - i.e. we have steady state and a simple differential equation:

$$D \cdot \frac{\partial^2[\Delta\ n_h(x)]}{\partial x^2} - \frac{\Delta\ n_h(x)}{\tau} = 0$$

- The solution (for a one-dimensional bar extending from $x = 0$ to $x = \infty$) is

$$\Delta n(x) \ = \ \Delta n_0 \cdot \exp - \frac{x}{L}$$

  - The length L is given by

$$L = \left(D_h \cdot \tau\right)^{1/2}$$

  - L is simply the diffusion length of the minority carriers (= holes in the example) as defined in the "simple" (but in this case accurate) introduction of life times and diffusion length.
  - This solution is already shown in the drawing above which also shows the direct geometrical interpretation of L.
- The important point to realize is that the *steady state* tied to this solution can only be maintained if the hole current at $x = 0$ has a constant, time independent value resulting from Fick's 1st law since we have no electrical fields that could drive a current.
  - This gives us

$$j_h(x = 0) = -\,e\cdot D\,\frac{\partial\,\Delta n_h(x)}{\partial\,x}\bigg|_{x = 0}$$

- By simple differentiation of our density equation from above we obtain

$$\frac{\partial\,\Delta n_h(x)}{\partial\,x}\bigg|_{x = 0} = -\,\frac{\Delta n_0}{L}$$

- Insertion into the current equation yields the final result

$$j_h\,(x = 0) = \frac{e\cdot D_h}{L_h}\cdot\Delta n_h(x = 0)$$

- The physical meaning is that the *hole part of the current* will decrease from this value as ***x*** increases, while *the total current* stays constant – the remainder is taken up by the electron current.

General Band-Bending and Debye Length

- The *Debye length* and the *dielectric relaxation* time are important quantities for *majority* carriers (corresponding to the *diffusion length* and the *minority carrier life time* for *minority* carriers). Let's see why this is so in this paragraph.
  - Both quantities are rather general and come up whenever density gradients cause currents that are counteracted by the developing electrical field.
  - An alternative simple treatment of the Debye length can be found in a basic module.
- Let us start with the Poisson equation for an arbitrary *one-dimensional* semiconductor with a varying electrostatic potential ***V*(x)** caused by charges with a density **ρ(*x*)** distributed somehow in the material. We then have

$$-\,\epsilon\cdot\epsilon_0\cdot\frac{d^2\,V(x)}{dx^2} = \epsilon\cdot\epsilon_0\cdot\frac{d\,E(x)}{dx} = \rho(x)$$

  - ***E***(*x*) is the electrical field strength; always minus the derivative of the potential ***V***.
  - The charge **ρ(*x*)** at any one point can only result from our usual charged entities, which are electrons, holes, and ionized doping atoms. **ρ(*x*)** is always the *net* sum of this charges, i.e.

$$\rho(x) = e\cdot\left(n_h(x) + N_D^+(x) - [n_e\,(x) + N_A^-(\,x)]\right)$$

  - The electrostatic potential ***V*** needed for the Poisson equation is now a function of ***x*** and shifts the conduction and valence band by the potential energy ***qV*** relative to some reference point for which one has ***V* = 0**. Since the band structure refers to the energy of electrons, we have that ***q* = –e** and thus may write

$$E_C\,(x) = E_C(\,V = 0) - e\cdot V(x)$$

$$E_V(x) = E_V(V = 0) - e\cdot V(x)$$

  - Thereby, the Poisson equation becomes

$$-\,\epsilon\cdot\epsilon_0\cdot\frac{d^2V\,(x\,)}{dx^2} = \frac{\epsilon\cdot\epsilon_0}{e}\cdot\frac{d^2\,E_C\,(x)}{dx^2} = e\left(n^{\,h}(x) + N_D^+(x) - [n^e(\,x) + N_A^-(\,x)]\right)$$

- If we now insert the proper equations for the four densities, we obtain a formidable differential equation that is of prime importance for semiconductor physics and devices, but not easy to solve.
- However, even if we could solve the differential equation (which we most certainly cannot), it would not be of much help, because we also a need a "gut feeling" of what is going on.

The best way to visualize the basic situation is to imagine a homogeneously doped semiconductor with a fixed charge density at its surface and no net currents (think of a *fictional insulating layer with infinitesimal thickness that contains some charge on its outer surface*).

- Carriers of the semiconducor thus can *not* neutralize the charge, and the surface charge will cause an electrical field which will penetrate into the semiconductor to a certain depth.
- This is the most general case for disturbing the carrier density in a surface-near region and thus to induce some *band-bending* .

There are *two* distinct major situations:

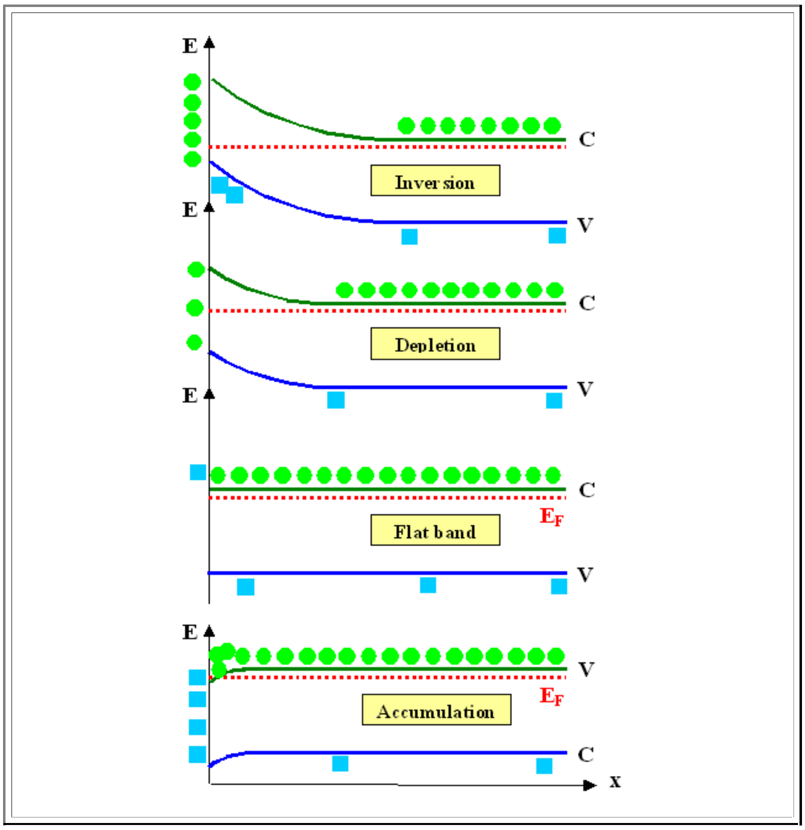
1. The surface charge has the *same polarity* as the majority carriers in the semiconductor, thus pushing them into the interior of the material.

- This exposes the ionized dopant atoms with opposite charge and a large **space charge layer** (**SCR**) will built up. This is also called the **depletion** case.
- The **SCR** is large because the dopant density is low and the dopant atoms *cannot move to the interface*. Many dopant atoms have to be "exposed" to be able to compensate the surface charge; the field can penetrate for a considerable distance.
- However: *In contrast* to what we learned about **SCRs** in **p–n** junctions, even for large fields (corresponding to large reverse voltages at a junction), the Fermi energy is $\boldsymbol{E}_\mathbf{F}$ *still constant* (currents are not possible). The bands are still bent, however, this means that $\boldsymbol{E}_\mathbf{C} - \boldsymbol{E}_\mathbf{F}$ incrases in the direction toards the surface.
- If the majority carrier density then is becoming very small in surface-near regions (it scales with **exp [– (**$\boldsymbol{E}_\mathbf{C} - \boldsymbol{E}_\mathbf{F}$ **)/(k*T*)]** after all), the minority carrier density increases due to the mass action law until minority carriers become the majority – we have the case of **inversion** .

2. The surface charge has the *opposite polarity* as the majority carriers in the semiconductor, thus accumulating them at the surface-near region of the material.

- Then majority carriers can move to the surface near region and compensate the external charge. The field cannot penerrate deeply into the material.
- This case is called **accumulation**.

The situation is best visualized by simple band diagrams, we chose the case for **n**-type materials. The surface charge is symbolized by the green spheres or blue squares on the left.



- Between depletion and accumulation must be the **flat-band** case as another prominent special case. This is not necessarily tied to a surface charge of zero (as shown in the drawing where a blue square symbolizes some positive surface charge), but for the external charge that compensates the charge due to intrinsic surface states.

We have some idea about the width of the space charge region that comes with the *depletion case*. But how wide is the region of appreciable band bending in the *case of accumulation*?

- Qualitatively, we know that it can be small - at least in comparison to a **SCR** - because the charges in the semiconductor compensating the surface charges are mobile and can, in principle, pile up at the interface

For the quantitative answer for all cases, we have to solve the Poisson equation from above. However, because we cannot do that in full generality, we look at some special cases.

First we restrict ourselves to the usual case of *one* kind of doping – **n**-type for the following example – and temperatures where the donors are fully ionized, which means that the Fermi energy is well below the donor level or $E_D - E_F >> kT$.

- We then have only *two* charged entities:

$$N_D^{\,+} = N_D$$

$$n^e = N_{\text{eff}}^e \cdot \exp -\frac{E_C - E_F}{kT}$$

- This means in what follows we only consider the *majority carriers*.

The Poisson equation then reduces to

$$\frac{\epsilon \cdot \epsilon_0}{e} \cdot \frac{d^2E_C(x)}{dx^2} = e\left(N_D - N_{\text{eff}}^e \cdot \exp -\frac{E_C(x) - E_F}{kT}\right)$$

- *And this, while special but still fairly general, is still not easy to solve* .

We will have to specialize even more. But before we do this, we will rewrite the equation somewhat.

- For what follows, it is convenient to express the band bending of the conduction band in terms of its deviation from the field-free situation, i.e. from $E_C^0 = E_C(x = \infty)$. We thus write

$$E_C(x) = E_C^0 + \Delta E_C(x)$$

- The exponential term of the Poisson equation can now be rewritten, we obtain

$$N_{\text{eff}}^e \cdot \exp -\frac{E_C(x) - E_F}{kT} = N_{\text{eff}}^e \cdot \exp -\frac{E_C^0 - E_F}{kT} \cdot \exp -\frac{\Delta E_C(x)}{kT}$$

- The first part of the right hand side gives just the electron density in a field-free part of the semiconductor, which – in our approximations – is identical to the density N_D of donor atoms. This leaves us with a usable form of the Poisson equation for the *case of accumulation* :

$$\frac{d^2E_C}{dx^2} = \frac{d^2(\Delta E_C)}{dx^2} = \frac{e^2 \cdot N_D}{\epsilon \cdot \epsilon_0} \cdot \left(1 - \exp\left(-\frac{\Delta E_C}{kT}\right)\right)$$

ΔE_C characterizes the amount of band bending. We can now proceed to simplify and solve the differential equation by considering different cases for the sign and magnitude of ΔE_C .

- Unfortunately, this is one of the more tedious (and boring) exercises in fiddling around with the Poisson equation. The results, however, are of prime importance – they contain the very basics of all semiconductor devices.

We will do *one* approximative solution *here* for the most simple case of **quasi-neutrality** which will give us the all-important *Debye length*.

- The other cases can be found in advanced modules:
  - Depletion
  - Inversion
  - Accumulation
  - Putting everything together

Quasi-neutrality is the mathematically most simple case; it treats only *small deviations* from equilibrium and thus from charge neutrality.

- The condition for quasi-neutrality is simple: We assume $|\Delta E_C| << kT$.
- We then can approximate the exponential function by its *Taylor series* and stop after the second term. This yields

$$\frac{d^2(\Delta E_C)}{dx^2} = \frac{e^2 \cdot N_D}{\epsilon \cdot \epsilon_0} \cdot \frac{\Delta E_C}{kT}$$

- That is easy now, the solution is

$$\Delta E_C(x) = \Delta E_C\,(x = 0) \cdot \exp - \frac{x}{L_{Db}}$$

- The solution defines L_{Db} = **Debye length** for ***n**-type semiconductors* **=** Debye length for electrons, we have

$$L_{Db} = \sqrt{\frac{\epsilon \cdot \epsilon_0 \cdot kT}{e^2 \cdot N_D}}$$

- Obviously the Debye length L_{Db} for *holes in **p**-type semiconductors* is given by

$$L_{Db} = \sqrt{\frac{\epsilon \cdot \epsilon_0 \cdot kT}{e^2 \cdot N_A}}$$

For added value, our solution also gives the field strength of the electrical field extending from the surface charges into the depth of the sample.

- Setting it to zero at the top of the valence band in the p-type material (as it is conventionally done), the electrostatic potential is related to the conduction band edge by $E_C\,(x) = E_g - e \cdot V(x)$. As discussed already above, the minus sign stems from the negative charge of an electron.
- Since the field strength $E(x)$ is minus the derivative of the electrostatic potential, we now have

$$E(x) = -\frac{dV(x)}{dx} = 1/e \cdot \frac{dE_C(x)}{dx} = -\frac{1}{e \cdot L_{Db}} \cdot \Delta E_C(x)$$

- Note that in the case of accumulation at the surface of an n-type semiconductor, $\Delta\ E_C(x)$ is negative, so the electric field comes out positive – in full agreement with the surface (at $x = 0$) being positively charged in this case.

The Debye length gives the typical length within which a *small* deviation from equilibrium in the *total charge density* – which for doped semiconductors is always dominated by the *majority carriers* – is relaxed or screened; in other words, it is no longer felt.

- L_{Db} is a *direct material parameter* – its definition contains nothing but prime material parameters (including the doping).
- For medium to high doping densities, it becomes rather small. The dependence of the Debye length on material parameters is shown in an illustration.
- The Debye length is also a prime material quantity in materials other than semiconductors - especially in ionic conductors and electrolytes (for which it was originally introduced). It also applies to metals, but there it is so small that it rarely matters.

The Debye length comes up in all kinds of equations. Some examples are given in the advanced modules dealing with the other cases of field-induced band bending

- *The Debye length is to majority carriers what the diffusion length is to minorities*. And just as the *diffusion length* is linked to the minority carrier lifetime τ , the *Debye length* correlates to a specific time, too, called the dielectric relaxation time τ_d .
- This will be the subject of the next paragraph.

Dielectric Relaxation Time

Let's start from the same situation that lead to the Debye length: A doped semiconductor, all dopants ionized, and some small disturbance in the charge equilibrium expressed as some small $\Delta\rho(x, t)$ somewhere, starting at some time t_0; i.e. we still assume quasi-neutrality.

- The Poisson equation now is extremely simple, we write it directly for the electrical field strength and have

$$\boxed{\frac{\mathrm{d}E(x,t)}{\mathrm{d}x} = \frac{\Delta\rho(x,t)}{\epsilon\cdot\epsilon_0}}$$

We now want to find out about *how long it takes* to establish a steady state, so we need some expression for **d(Δρ)/d*t***. The Poisson equation won't help because it does not explicitly contain the time dependence.

- But simply using the continuity equation for the relevant charge density $\Delta\rho$ provides a **d(Δρ)/d*t*** term. Since we are treating quasi-neutrality, we neglect all terms with *gradients* in the carrier density (this will turn out to be fully justified).
- Since the only relevant current is the drift curent $j(x) = \sigma E(x) = \rho\cdot\mu\cdot E(x)$, this leaves us with the following *continuity equation*

$$\boxed{\frac{\partial(\Delta\rho)}{\partial t} = -\rho\cdot\mu\cdot\frac{\partial E(x)}{\partial x}}$$

Inserting **d*E*/d*x*** from the Poisson equation gives

$$\boxed{\frac{\partial(\Delta\rho)}{\partial t} = -\frac{\rho\cdot\mu}{\epsilon\cdot\epsilon_0}\cdot\Delta\rho}$$

- ρ is the total carrier density, we can write it as $\rho = \rho_0 + \Delta\rho \approx \rho_0$ since we have quasi neutrality; **μ**, as always, is the mobility of the carrier in question.

This is a differential equation for $\Delta\rho(x, t)$ with the simple solution

$$\boxed{\begin{gathered}\Delta\rho(x,t) = \Delta\rho(x,0)\cdot\exp -\frac{t}{\tau_d}\\[2ex] \tau_d = \frac{\epsilon\cdot\epsilon_0}{\mu\cdot\rho_0}\end{gathered}}$$

- With τ_d = **dielectric relaxation time** = another basic material constant for the same reason as the Debye length.
- The *dielectric relaxation time* tells us exactly what we wanted to know: How long does it take for the majority carriers to respond to a disturbance in the charge density.

While this definition of some special time is of some interest, but not overwhelmingly so, the situation gets more exciting when we consider relations between our basic material constants obtained so far:

- Since $\mu\cdot\rho = \sigma$, the conductivity of the material (for the carriers in question), we have the simple and fundamental relation

$$\tau_d = \frac{\epsilon \cdot \epsilon_0}{\sigma}$$

- Now let's see if there is a correlation to the Debye length:
  - We use the Einstein relation $\boldsymbol{D = \mu(kT/e)}$, the Debye length definition ($\boldsymbol{L_{Db} = \{(\epsilon \cdot \epsilon_0 \cdot kT)/(e \cdot \rho)\}^{1/2}}$, pluck it into the definition of the dielectric relaxation time (again replacing $\boldsymbol{e \cdot N_D}$ by ρ) and obtain

$$\tau_d = \frac{L_{Db}^2}{D}$$

$$L_{Db} = \sqrt{D \cdot \tau_d}$$

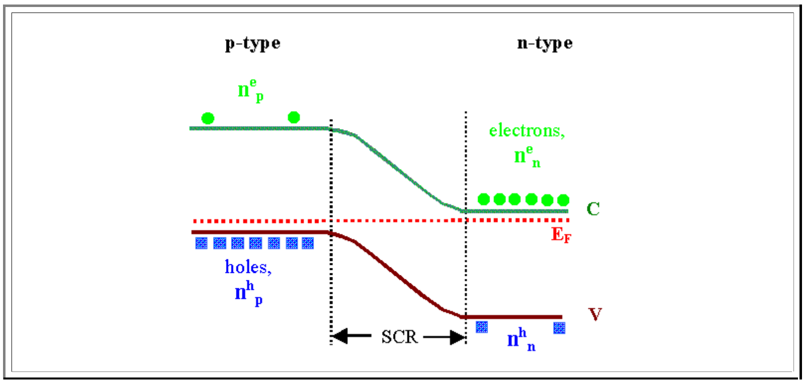
  - This is exactly the same relation for the majority carriers between a *characteristic time constant* and a *length* as in the case of the minority carriers where we had the minority lifetime τ and the correlated diffusion length $\boldsymbol{L}$.
  - The physical meaning is the same, too. In both cases the times and lengths give the numbers for how fast a deviation from the carrier equilibrium will be equalized and over which distances small deviations are felt.
- This merits a few more thoughts.
  - If the carrier density is high, τ_d is in the order of *picoseconds* and $\boldsymbol{L_{Db}}$ extends over *nanometers*. Any deviation from equilibrium is thus almost instantaneously wiped out, or, if that is not possible, contained within a very small scale.
  - And this is the regular situation for *majority* carriers. The few minority carriers always present in the semiconductor, too, can be safely neglected.
- For *minority* carriers, however, the situation is *entirely different*.
  - Their density is very small; τ_d and $\boldsymbol{L_D}$ consequently are no longer small.
  - Moreover, whatever disturbance occurs in the density of *minorities*, there are plenty of majorities that can react very quickly (with their τ_d) to the electrical field always tied to a $\Delta\rho_{min}$.
- The majority carriers are always attracted to the minorities and thus will quickly surround any excess minority charge with a "cloud" of majority carriers (which is called *screening*), essentially compensating the electrical field of the excess minorities to zero.
  - They will, of course, eventually remove the excess charge by recombination, but that takes *far longer* than the time needed to do the screening.
  - Since the electrical field is now zero, the excess charge cannot disappear or spread out by field currents – only spreading by diffusion in the density gradient (which is automatically introduced, too) is possible.
- But this is exactly the process that we have neglected in this discussion (we had all density gradients in the continuity equation set to zero!).
  - *Dielectric relaxation* (i.e. the disappearance of charge surpluses driven by electrical fields) is thus not applicable to minority carriers. Charge equilibration there is driven by diffusion - which is a much slower process!
  - This then justifies the simple approach we took before, where we only considered the diffusion of minorities and did not take into account the majority carriers.

2.3.5 Junction Reconsidered

- In this section we will give the **p–n** junction a new look using a somewhat more advanced point of view.
    - A full treatment of **p–n** junctions in any kind of *three-dimensional* semiconductor, taking into account arbitrary doping *profiles* , *finite* size, and effects of the *interfaces* and *surfaces*, is one of the more difficult things to do in semiconductor theory; we will not attempt it here.
    - In all standard treatments of junctions, we always look at special (unrealistic) cases and use (lots of) approximations. This, admittedly, can become somewhat confusing.
    - It is thus very advisable to become really well acquainted with the simple treatments given in section 2.2.4; this will clear the mind for the essentials.

Junction Without Contributions from the Space Charge Region

- Here we look at an advanced treatment of a **p–n** junction, but we still will have to make some simplifications.
    - We consider a "*narrow* " "*abrupt*" one-dimensional junction in an infinitely long crystal, which is formed by a sudden change of doping, i.e. there are *no gradients* of the doping densities N_D and N_A to the left or right of the junction.
    - "Narrow" then means that the width of the space charge layer is much smaller than the diffusion length of minority carriers, but still much larger than the mean free path length needed for the thermalization of carriers.
    - In consequence, we do not consider recombination in the **SCR**, and we can assume thermal equilibrium in the bands, i.e. we can use the quasi-Fermi energies.
- First, we look at the junction in equilibrium, i.e. there is no net current and the Fermi energy is the same everywhere (we have the same situation as shown before; but for diversity's sake, the **p**- and **n**-side reversed).



- In what follows, there will be a lot of shuffling formulas around – and somehow, like by magic, the ***I–V*** characteristics of a **p–n** junction will emerge. So let's be clear about what we want to do in the major steps (highlighted by the cyan background).
- *The first basic goal* is to find an expression for the *carrier density at the edge of the space charge region*.
    - We know in a qualitative way from the consideration of pure diffusion currents that the minority carrier density around the edge of the space charge region is somewhat larger than in the bulk under equilibrium conditions – there is a Δn_{min} given by $\Delta n_{min} = n_{min}(x) - n_{min}(\text{bulk})$, and Δn_{min} will increase for forward bias, i.e., for non-equilibrium conditions.
    - We also know that Δn_{min} induces a diffusion current and that we therefore need a "real" current to maintain a constant Δn_{min}. Finding $\Delta n_{\ min}$ thus will *automatically give us the necessary currents* belonging to the non-equilibrium as defined by the voltage.
- We first look at the various energies involved:
    - The energy difference between the left side (= **p**-side, *raised index "**p**"*) and the right side (= **n**-side, *raised index "**n**"*) of the junction is given by

$$\boxed{E_C{}^{p} - E_C{}^{n} \ = \ E_V{}^{p} - E_V{}^{n} \ = \ e \cdot \Delta V}$$

    - Here, ΔV is the difference of the (yet-to-be-determined) electrostatic potential between the **n-** and **p**-side, taken far away from the junction; the details (especially about its sign) will be given below.
- The band energy levels $E_C{}^{n,p}(x)$ and the potential $V(x)$ are functions of x, which makes all densities functions of x, too. *We will, however, write all these quantities without the "(**x**)" from now on.*
    - As long as we discuss equilibrium, the Fermi energy is constant and the carrier densities are given by their usual expression. We consider them separately for the left- and right hand side of the junction, i.e., for the **n**- and **p**-part. (Note that here the usual minus sign in the exponent was used to change the order of the terms in the differences.)

$$n_e^p = N_{eff}^e \cdot \exp \frac{E_F - E_C{}^p}{kT}$$

density of electrons on the p-side

$$n_h^p = N_{eff}^h \cdot \exp \frac{E_V{}^p - E_F}{kT}$$

density of holes on the p-side

$$n_e^n = N\,_{eff}^e \cdot \exp \frac{E_F - E_C{}^n}{kT}$$

density of electrons on the n-side

$$n_h^n = N_{eff}^h \cdot \exp \frac{E_V{}^n - E_F}{kT}$$

density of holes on the n-side

- We also have the mass action law, here applicable everywhere since we are in full equilibrium:

$$n_e{}^p \cdot n_h{}^p = n_h{}^n \cdot n_e{}^n = n_{min} \cdot n_{maj} = n_i{}^2 = N_{eff}{}^e \cdot N_{eff}{}^h \cdot \exp - \frac{E_g}{kT}$$

What we *need to know* to get on is the ***x**-dependence of the energies* or of the potential ***V*(*x*)** – this simply means we need the quantitative band diagram that so far we always just drew "by feeling". (This is one of the **essential** points why we reconsider the p–n junction here; the other one will be the usage of the quasi-Fermi energies.)

- For this we need to solve the **Poisson equation** and this demands to specify the total charge **ρ(*x*)** so that we can write down the charge as a function of ***x***. *This is easy in principle*:

The total (space) charge **ρ(*x*)** at any point along the junction is the sum of all charges: Electrons (**n_e(*x*)**), holes (**n_h(*x*)**), ionized donors (**$N_D{}^+$(*x*)**), and ionized acceptors (**$N_A{}^-$(*x*)**). We have as before

$$\rho(x) = e \cdot \left(n_h(x) - n_e(x) + N\,_D{}^+(x) - N_A{}^-(x) \right)$$

- Inserting **ρ(*x*)** into the Poisson equation gives

$$- \frac{\epsilon\,\epsilon_0}{e} \cdot \frac{d^2V(x\,)}{dx^2} = n_h(x) - n_e(x) + N_D{}^+(x) - N_A{}^-(x)$$

- Solving this equation with the proper boundary conditions will yield ***V*(*x*)** and everything else – *but not so easily* because the situation is complicated: Since **n_h(*x*)** and **n_e(*x*)** depend on ***V*(*x*)** via the Fermi distribution, this is also an *implicit* equation for ***V*(*x*)**.
- It is, however, not too difficult to find good approximations for "normal", i.e. highly idealized junctions; this is shown in an advanced module accessible through the link.

For our final goal, which is to describe the **current–voltage characteristic** of a **p–n** junction, we use the same *approximations* and conventions, namely:

1. The *zero point of the electrostatic potential* is identical to the valence band edge in the **p**-side of the junction, i.e. **$eV^p = E_V{}^p = 0$** as shown in the complete illustration to the situation shown in the picture above. This is a simple *convention* without any physical meaning.

2. *All dopants are ionized, their density is constant up to the junction, and there is only one kind on each side of the junction*, i.e.

$$n_h{}^p(\text{bulk}) \;=\; N_A \;=\; N_A{}^-$$

$$n\,{}_e{}^n(\text{bulk}) \;=\; N_D \;=\; N_D{}^+$$

- This is a *crucial assumption*. Note that while $\boldsymbol{n_{h,e}{}^{p,n}}$ **(bulk)** are constant, this is not required for $\boldsymbol{n_{h,e}{}^{p,n}(x)}$ around the junction.

3. We also assume that away from the junction, the **Si** extends into infinity (or at least to a distance much larger than several diffusion lengths) to both sides of the junction – in total we use the "abrupt" "large" junction approach

- This gives us for the carrier densities in equilibrium anywhere in the junction:

$$n_h(x) \;=\; N_A \cdot \exp - \frac{e \cdot V(x)}{kT}$$

$$n_e(x) \;=\; N_D \cdot \exp - \frac{e \cdot [V^n \;-\; V(x)]}{kT}$$

- Here, $\boldsymbol{V^n}$ is the constant value of the potential deep in the **n**-type region. Note that, having chosen the zero point for $\boldsymbol{V(x)}$ at the **p**-side of the junction where there is the negative pole of the electric field, it holds inside the **SCR** that $\boldsymbol{0 \leq V(x) \leq V^n}$.
- These equations mean that the carrier density is whatever you have in the undisturbed **p-** or **n-**part (i.e., the dopant density) times the Boltzmann factor of the energy shifts relative to this situation.

$\boldsymbol{V^n}$ is the difference of the built-in potential for equilibrium conditions, it is thus determined by the difference in the Fermi energies of the n- and the p-side before contact (relative to the band edges) – our simple view of a junction is totally correct on this point.

- With and without an external voltage $\boldsymbol{U_{ext}}$ we have

$$V^n(U_{ext}{=}0) \;=\; \frac{1}{e} \cdot \left(E_F{}^n \;-\; E_F{}^p\right)$$

$$V^n(U_{ext}) \;=\; \frac{1}{e} \cdot \left(E_F{}^n \;-\; E_F{}^p \;+\; e \cdot U_{ext}\right)$$

- Here, the sign of $\boldsymbol{U_{ext}}$ is such that a positive external voltage *increases* the built-in potential difference. Note that this is just an interim choice; later on we will replace it by the usual standard.
- In the general case, the maximum potential at the **n**-side, $\boldsymbol{V^n}$**(bulk)**, becomes

$$V^n(\text{bulk}) \;=\; V\,{}^n(U_{ext}{=}0) + U_{ext} \;=\; \frac{1}{e} \cdot \;\Delta E_F + U_{ext} \;=\; V^n + U_{ext} \;=\; \Delta V$$

- Looking at the proper solution of the Poisson equation for our case, we realize that the space charge region was defined as the part of the **Si** where the potential was not yet constant. This means that $\boldsymbol{V^n}$ **(bulk)** $\boldsymbol{= V^n|_{\text{SCR edge}}}$ on the **n**-side, and $\boldsymbol{V^p\, |_{\text{SCR edge}} = 0}$. This is an essential point, even so it is matter-of-course.

We now can move towards our primary goal and find an expression for the carrier density at the edge of the **SCR** by considering the ratio of a carrier species on both sides of the junction. From the equations above, we obtain *for the edge of the **SCR*** :

$$\left.\frac{n_e{}^p}{n_e{}^n}\right|_{\substack{\text{SCR}\\\text{edge}}} \;=\; \left.\frac{n_h{}^n}{n_h{}^p}\right|_{\substack{\text{SCR}\\\text{edge}}} \;=\; \exp - \frac{e \cdot \Delta V}{kT} \;=\; \exp - \frac{e \cdot (V^n \;+\; U_{ext})}{kT}$$

- The *minority* carrier densities (*always at the edge of the **SCR** without indicating it anymore*) can now be written as

$$n_e^p(U_{ext}) \;=\; n_e^n(U_{ext}) \cdot \exp - \frac{e \cdot (V^n + U_{ext})}{kT} \qquad \text{electrons on the p-side}$$

$$n_h^n(U_{ext}) \;=\; n_h^p(U_{ext}) \cdot \exp - \frac{e \cdot (V^n + U_{ext})}{kT} \qquad \text{holes on the n-side}$$

- These equations are nothing but the Boltzmann distribution giving the number of particles **(n_{min})** that make it to the energy **$e(V^n + U_{ext})$** out of a total number **n_{maj}** – in thermal equilibrium. We used essentially the same equation before, but now we know the kind of approximations that were necessary and that means we also know what we would have to do for "better" solutions of the problem.

Since this is important, let's review the approximations we made:

- Besides the "abrupt" "large" junction, we used the approximations from the simple solution to the Poisson equation which implies that the potential stays constant right up to the edge of the **SCR** and then changes monotonously.
- This means that for equilibrium we must obtain the same equations by computing the minority carrier density from the mass action law, i.e.

$$n_e^p(U_{ext}{=}0) \;=\; \frac{n_i^2}{n_h^p(U_{ext}{=}0)}$$

- We will see if this is true in a little exercise:

Exercise 2.3.5-1

Show the equivalence of the two equations
for the minority carrier density!

Now comes a crucial point: We are looking at *stationary non-equilibrium*. We first review the starting point again:

- At equilibrium (**$U_{ext} = 0$**), the *majority* carrier densities **$n_e^{\,n}\,|_{\text{SCR edge}}$** and **$n_h^{\,p}\,|_{\text{SCR edge}}$** are given by

$$n_h^p \;=\; N_{eff}^p \cdot \exp - \frac{E_F}{kT}$$

$$n_e^n \;=\; N_{eff}^e \cdot \exp + \frac{E_F - E_C^n}{kT}$$

- Do you remember them? These are two of our first equations from above, but given here for the choice of **$E_V^{\,p} = 0$**.

The essential point for the majority carrier density at the edge of the space charge region for *non-equilibrium* is that it *remains practically unchanged* (approximately at its bulk value) if we now apply a voltage **U_{ext}**, i.e.

$$n_{e,h}^{\,n,p}(U_{ext})\Big|_{\substack{\text{SCR}\\\text{edge}}} \;=\; n_{e,h}^{\,n,p}(\text{equ})\Big|_{\substack{\text{SCR}\\\text{edge}}} \;=\; n_{e,h}^{\,n,p}(\text{bulk})$$

- The trick here is that we consider the majority carrier density *at the **SCR** edge* – and the position of the latter may vary with the applied voltage!
- Nevertheless, beyond that point we have the bulk behaviour of the majorities – because that's how we have defined the **SCR** edge: The bulk potential stays constant right up to the edge, and this is only possible for a constant density of majority carriers.

The minority carrier densities **$n_e^p\,|_{\text{SCR edge}}$** and **$n_h^n\,|_{\text{SCR edge}}$**, however, depend *very much on the applied voltage* as expressed in the formulae above.

- Thus, we have to adjust the minority carrier density independent of the majority density, which means we have to use the **quasi-Fermi energies**.
- In other words: While the *quasi-Fermi energy* $E_F{}^{maj}$ for majority carriers remains at the equilibrium value E_F near the **SCR**, the *quasi-Fermi energy* for the minority carriers, $E_F{}^{min}$, branches off early; the details will be shown below.

We now ask about the *difference of the minority carrier density relative to equilibrium*, i.e. we look at

$$\Delta n_{e,h}{}^{p,n}\Big|_{\substack{\text{SCR}\\ \text{edge}}} = n_{e,h}{}^{p,n}(U_{ext}) - n_{e,h}{}^{p,n}(U_{ext}=0)$$

- It comes out as

$$\begin{aligned}\Delta n_{e,h}{}^{p,n} &= n_{e,h}{}^{n,p}\cdot\left(\exp-\frac{e\cdot(V^n+U_{ext})}{kT} - \exp-\frac{eV^n}{kT}\right)\\ &= n_{e,h}{}^{n,p}\cdot\exp-\frac{eV^n}{kT}\cdot\left(\exp-\frac{eU_{ext}}{kT} - 1\right)\end{aligned}$$

Inserting the general expressions for the minority carrier density from above for the case $U_{ext} = 0$ yields the final formula for our first goal:

$$\Delta n_{e,h}{}^{p,n}\Big|_{\substack{\text{SCR}\\ \text{edge}}} = n_{e,h}{}^{p,n}(\text{equ})\cdot\left(\exp-\frac{e\,U_{ext}}{kT} - 1\right)$$

- In other words: The density of minority carriers at the edge of the **SCR** will be changed by an external voltage.

In steady state conditions (which does *not* imply equilibrium, just that nothing changes) this density must remain constant as a function of time.

- Since deep in the material the minority carrier density is unchanged and has its equilibrium value, we now must have a current, driven by the density gradient alone, and *this current must be maintained by the voltage/current source* if we want steady state.
- Physically speaking, the excess density of minority carriers will diffuse around and disappear after some diffusion lengths – deep in the material they are not noticeable any more.

This is exactly the situation treated under "useful relations" for pure diffusion currents.

- We can take the formula derived there with $\Delta n_{p,n}{}^{e,h}(x=0)$ given by the equation from above and obtain immediately for the **current–voltage relationship** of a **p–n** junction (just considering the absolute magnitudes):

$$|\,j_e(U_{ext})\,| = \frac{e\cdot D_e}{L_e}\cdot\Delta\,n_e\Big|_{\substack{\text{SCR}\\ \text{edge}}}\qquad\text{or}$$

$$|\,j_e(U_{ext})\,| = \frac{e\cdot D_e}{L_e}\cdot n_e{}^p(\text{equ})\cdot\left(\exp-\frac{eU_{ext}}{kT} - 1\right)$$

$$|\,j_h(U_{ext})\,| = \frac{e\cdot D_h}{L_h}\cdot n_h{}^n(\text{equ})\cdot\left(\exp-\frac{eU_{ext}}{kT} - 1\right)$$

We now see that the external voltage, as we have introduced it, raises the potential barrier and therefore decreases the minority carrier density – and, thus, also the current flow.

- This means that, in order to *enhance* the current flow over the p–n junction, we have to apply the external voltage in a way that it *lowers* the barrier.

- Therefore, the forward voltage is $U_D := -U_{ext}$, and since it is the forward voltage, it is also the one which is taken as positive; the subscript "D" refers to the p–n junction functioning as a **diode**.

For the *final result* we add the electron and hole currents, drop suffixes and functional arguments now unnecessary, and obtain the **diode equation** (giving the total current density, including the reverse current, counted in the standard way):

$$\boxed{j_D(U_D) = \left(\frac{e \cdot n_e{}^p \cdot D_e}{L_e} + \frac{e \cdot n_h{}^n \cdot D_h}{L_h} \right) \cdot \left(\exp \frac{eU_D}{kT} - 1 \right)}$$

This is the same equation as before if we take into account *that the pre-exponential factor can be written in many ways*. To see that, we use the following identies:

- For the diffusion length we have

$$L_{e,h} = \left(D_{e,h} \cdot \tau_{e,h} \right)^{1/2}$$

$$D_{e,h} = \frac{L_{e,h}{}^2}{\tau_{e,h}}$$

$$\tau_{e,h} = \frac{L_{e,h}{}^2}{D_{e,h}}$$

- From the mass action law, which is still valid for the bulk, and the general approximation for the majority carrier density (that is already contained in our equations) we get

$$n_{e,h}{}^{p,n} = \frac{n_i{}^2}{n_{h,e}{}^{p,n}}$$

$$n\ {}_e{}^p = \frac{n_i{}^2}{N_A}$$

$$n\ {}_h{}^n = \frac{n_i{}^2}{N_D}$$

Shuffling everything around with these identities gives us – among many other equivalent formulations – . . .

$$\boxed{j_D(U\ {}_D) = \left(\frac{e \cdot L_e \cdot n_i{}^2}{\tau_e \cdot N_A} + \frac{e \cdot L_h \cdot n_i{}^2}{\tau_h \cdot N_D} \right) \cdot \left(\exp \frac{e\ U_D}{kT} - 1 \right)}$$

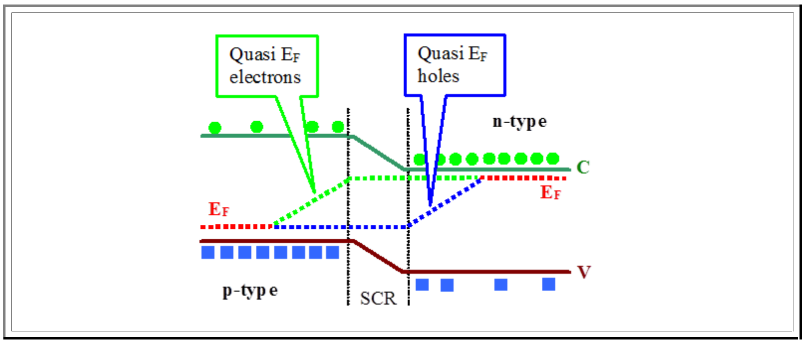
- . . . and that is exactly the equation we got before! However, we did not have to "cut corners" this time and we did not have to assume that some proportionality constant equals **1**!

More important, however: The interpretation of what happens may now be different. *Different* in the sense of looking at one and the same situation from a *different point of view*, not different in the sense that it is something else. The two points of view are complementary and not mutually exclusive; neither one is wrong!

- In the simple picture we looked at the minority carriers that had to be *generated* to account for the *loss of carriers accounting for the reverse current* and running down the energy slope.
- Here we looked at the *surplus of minorities accounting for the forward current* and which has to be moved away from the junction.
- Think about why this is the same thing! (Hint: Start from $U_D = 0$.)

What is left is just to consider the *quasi-Fermi energies* relevant for the forward direction; not only was the relevant drawing promised already above, it will also show explicitly what is meant by "surplus of minorities, having to be moved away from the junction" – because it will show us where those minorities end up.

- To cut a long story short, here it is:



- That the quasi-Fermi energies of the majorities remain constant throughout the **SCR** corresponds to the expressions giving the ratio of each carrier type on both sides of the junction.

Note that in the dawing, deliberately there are more minority carriers close to the **SCR** edges than deeper in the bulk. Yes, that's where the surplus minorities go. But that's not the end of the story:

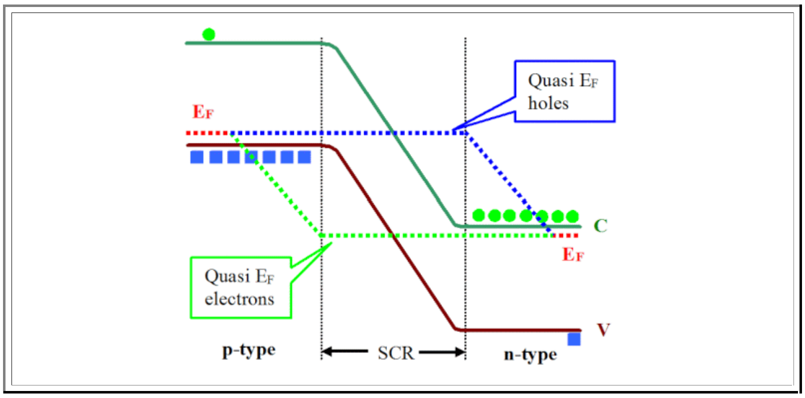
- That the quasi-Fermi energies of the minorities outside the **SCR** linearly merge towards the majorities' ones corresponds to the exponential decay of the surplus minority density away from the **SCR**, with the decay constant given by the diffusion length – as already discussed for the case of pure diffusion currents.
- Think for yourself about why all this is the case! And think about the possible consequences of the surplus minorities' presence in the case of a direct semiconductor.

Contributions from the Space Charge Region

We now should include the **generation currents from the space charge region**, as we did (in a somewhat fishy way) in our simple consideration of a junction.

- This, however, is not so easy to do in a correct (albeit still very approximate) fashion.

For the *reverse part of the generation current* from the **SCR**, we can obtain an equation directly from the Shockley-Read-Hall theory. All we have to do is to consider the **quasi-Fermi energies** of a *junction in reverse bias*. This is schematically shown in the following picture:



- The quasi-Fermi energies must behave in the way shown (the details do not matter), because otherwise the density of charge carriers (especially minority carriers!) in the junction would be too high.
- Note that in the dawing there are no minority carriers close to the **SCR** edges (deliberately!); only in those regions away from the **SCR**, where there is a single Fermi energy (shown in red), minority carriers are depicted. There, the standard full-equilibrium mass action law holds.

The decisive point is that we may consider any given thin slice of the **SCR** to be in *local equilibrium*, and that *the quasi-Fermi energy of the electrons is lower than that of the holes* throughout the **SCR**.

- The latter is a direct consequence of the applied reverse bias, increasing and steepening the potential barrier in the **SCR**, in combination with the diffusion length of the minorities being larger than the width of the **SCR** (remember the *narrow junction approximation* from above).
- This ordering of the quasi-Fermi energies is the exact opposite of the situation that we have considered so far in the recombination business, where we looked at an *increased density of carriers*, e.g. produced by irradiation with light. Then recombination outweighs generation and U_{DL}, the difference between recombination and generation, was positive.

- Hence, in the case we are considering here, U_{DL} *is negative* , i.e. there is more generation than recombination. And this means that the space charge region is busily producing carriers, always in pairs because of neutrality, which will run down the energy barrier producing *an additional reverse current*.
  - Pair production means that a deep level first emits a hole to the valence band, and then an electron to the conduction band.
  - Let's look at this using the formula for U_{DL}:

$$U_{DL} = \frac{v \cdot \sigma \cdot N_{DL} \cdot (n_e \cdot n_h - n_i^2)}{n_e + n_h + 2n_i \cdot \cosh \dfrac{E_{DL} - E_{MB}}{kT}}$$

  - For making estimates easier, we assume a mid-band level (i.e., $\cosh[(E_{DL} - E_{MB})/(kT)] = 1$) and n_e, $n_h << n_i$. This leaves us with

$$U_{DL} = -\frac{v \cdot \sigma \cdot N_{DL} \cdot n_i}{2}$$

  - For these assumptions we have seen that, treating holes and electrons on equal footing, $1/(v \cdot \sigma \cdot N_{DL}) = \tau$.
  - However, because we now have *more generation* than recombination, τ is now called the **generation life time** τ_G for this case. (More to that topic in the link.)
  - This leaves us with a *net generation of one kind of carrier* of

$$|U_{DL}| = G = \frac{n_i}{2\tau_G}$$

- The current density from the net generation of carriers in the **SCR** is then given by the product of the net generation rate with the width ***d*** of the **SCR**; adding up the holes and the electrons yields

$$j_R(\text{SCR}) = \frac{e \cdot n_i \cdot d}{\tau_G}$$

  - This is exactly the same formula (give or take a factor of **2**) as in our "quick and dirty" estimate from before. The physical reasoning wasn't so different either, if you think about it.
- How about the contribution of the **SCR** to the *forward current*?
  - The proper treatment is much more complicated and physically different from our simple explanation. The physical reasoning is as follows:
  - We have seen that we need to sustain a certain density of surplus minority carriers, $\Delta n_{e,h}^{p,n}$, at the edges of the **SCR** to maintain local equilibrium. The surplus carriers needed were injected from the other side of the junction and crossed the junction *without losses* – at least in our present approximation.
  - In reality, however, *some injected holes from the **p**-side will recombine with the injected electrons from the **n**-side*. Recombination in the **SCR** thus reduces the current needed to maintain $\Delta n_{e,h}^{p,n}$ and an additional current has to be produced which exactly compensates the losses.
- The necessary calculations are shown in an advanced module, suffice it to state here that the final result for the forward current from the **SCR** is (in a rather crude approximation)

$$j_F(\text{SCR}) = \frac{e \cdot n_i \cdot d}{2\tau_G} \cdot \exp \frac{e\,U_D}{2kT}$$

- Again, besides the factor **2** (and the new kind of life time), the same formula as before. But this time it was a kind of lucky coincidence, not really very well justified.
- *Or was it*? Think about it!

3. Silicon: General Properties and Technologies

3. Silicon: General Properties and Technologies

3.1 General Properties

3.1.1 Conductivity, Lifetime and Lattice Defects

General Remarks

This Module is unfinished

- Silicon is by far the most common semiconductor used for all kinds of products; it accounts for more than **90%** (my estimate) of semiconductor products (measured in **kg** or **$**; whatever).
  - It is also by far the most perfect semiconductor in terms of crystalline perfection, the cheapest in its field of applications if you consider value for money, and perhaps the best understood.
  - There are, however, plenty of questions concerning the properties of **Si** that are *not* well understood and thus there is plenty of research opportunity and enough room for new uses of the good old **Si**.
- Most of the basic properties of **Si** were already covered in the illustrations to the preceding chapter. Here only a short summary will be given.
- The **intrinsic conductivity** is governed by the band gap which determines the carrier density. The resistivity, in turn, is given by *carrier density times mobility*. It should thus be possible to give precise values.
  - However, upon closer inspection it turns out that not only carrier mobilities are temperature dependent, but the band gap, and the effective masses, too. That makes precise calculations difficult. We have, e.g.
    E_g(300 K)=1,1242 eV
    E_g(0 K)=1,1700 eV.
  - Measurements, while always possible, must allow for non-perfection **-** there is no such thing as an perfect intrinsic semiconductor. With lots of precautions, it is possible to grow crystals with a resistivity (at room temperature) of about **1 000 Ωcm -** still far lower than the intrinsic value.
- This should motivate a little exercise:

Exercise 3.1.1

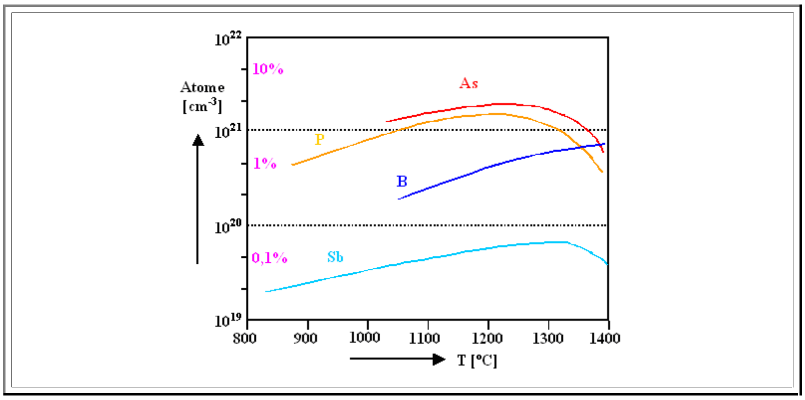
What is perfection?

- All in all, the precise temperature dependence of truly intrinsic **Si** is not so interesting anyway **-** because we never encounter it.
  - What is interesting are some basic numerical values:

Bandgap (indirect)	**$E_{\,g}$(300 K)=1,1242 eV**
Effective density of states (conduction band)	$3{,}22 \cdot 10^{19}\ \text{cm}^{-3}$
Effective density of states (valence band)	$1{,}83 \cdot 10^{19}\ \text{cm}^{-3}$
Intrinsic carrier density n_i (300K)	$1{,}3 \cdot 10^{10}\ \text{cm}^{-3}$
Intrinsic mobility of electrons (300K)	ca. 1400 cm^2/Vs
Intrinsic mobility of holes (300K)	ca. 500 cm^2/Vs
Lifetime (max).	1 ms (für ≈ 100 Ωcm)
Density	$2.33\ \text{g/cm}^{-3} \approx 5 \cdot 10^{22}\ \text{atoms/cm}^{-3}$
Surface density	$\left.\begin{matrix}11.8\\9.6\\6.8\end{matrix}\right) \times 10^{14}\ \dfrac{\text{atoms}}{\text{cm}^2}\left(\begin{matrix}\text{on }\{111\}\\\text{on }\{110\}\\\text{on }\{100\}\end{matrix}\right.$

Conductivity and Initial Doping

- As stated above, it is not easy to produce **Si** with a conductivity close to the intrinsic limit.
  - **Si** wafers commercially available therefore have resistivities $\rho \leq$ **100 Ωcm** (at room temperature); if you want something better than that you have to negotiate with the supplier.
  - An example for what is on the market can be found in the link.
- At the other end, large conductivities, obtained by heavy doping, are limited by the maximum **solubility** of the dopants. The term "solubility" refers to the (temperature dependent) equilibrium concentration of impurity atoms that can be incorporated into a crystal as single atoms.
  - For concentrations higher than the **solubility limit**, the surplus impurity atoms would tend to precipitate - and precipitates of **As**, **B**, or **P**, while possibly introducing defect levels in the band gap, are *not* active as dopants.
  - The highest achievable meaningful doping levels are thus given by the highest solubility of a doping element (always at a high temperature) - provided it can be kept in *solid solution*. This means that precipitation must be suppressed because at room temperature the solubility is usually low.
  - The picture below shows solubility data for the common **Si** dopants. In all cases the solubility is rather large (with a maximum between **1200 °C** and **1400 °C**), and may surpass the **1%** level.



  - If the maximum concentration *could* be kept in solid solution while the crystal is grown, the resistivity *could* be lower than what has been listed above. However, this is neither feasible nor sensible. Making devices always involves some heating, and if dopants are present in a high supersaturation, precipitation may start during processing, leading to all kinds of unwanted effects besides the unavoidable change in conductivity.
- There is, however, no particular need for high doping levels *and* a high precision of doping. Making devices always means that *local* doping is optimized -e.g. while making a source/drain area - and this requires medium to low doping levels of the substrate which are easily achieved.
  - High doping is only important if the wafers are used as substrate for epitaxial layers, and then the exact doping level is not crucially important.
- The real importance of the solubility curves thus is in specifying how much doping can be achieved on top of the initial doping by e.g. diffusion or ion-implantation. This will of course depend on the temperature!
  - Ion implantation done at room temperature thus always needs an **annealing** step to "**activate**" the dopants, i.e. to dissolve them in the **Si** up to the solubility limit at the chosen temperature. At the same time, this high temperature process also (hopefully) "anneals" the lattice defects produced by ion implantation.
  - There are limits to **defect annealing**, too: High doping levels change the average lattice constant - especially the small **B** ion reduces this parameter. Heavily **B**-doped layers thus want to contract, and since this is not possible on top of a solid substrate, mechanical stress is introduced which may lead to the formation of so-called "misfit dislocations" - the link shows an example.

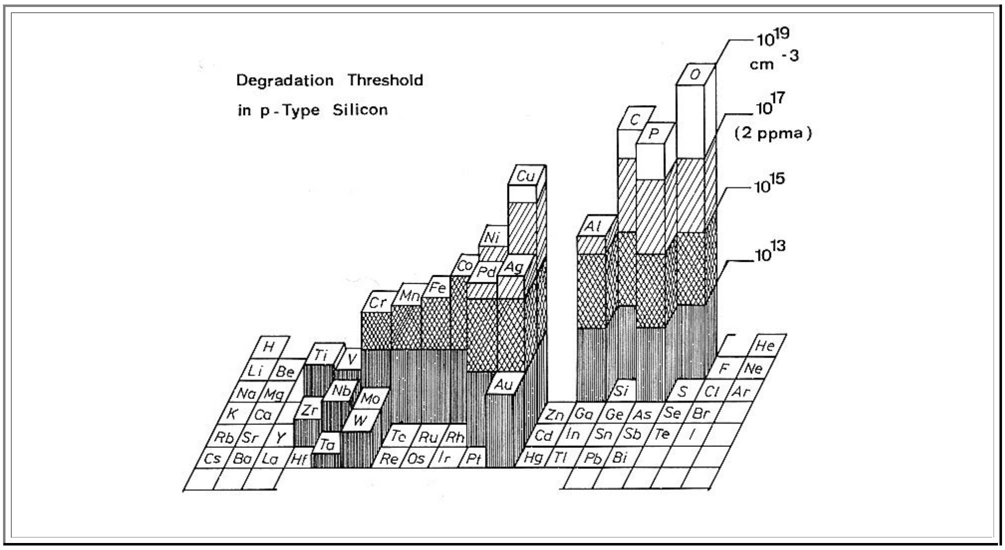
Lifetime and Diffusion Length

- Silicon is an indirect semiconductor, we expect relatively large lifetimes and diffusion lengths.
  - This is indeed the case, lifetimes as large as **1 ms** can be observed in extreme cases (This is an exceedingly long time for an electron!)
  - The corresponding diffusion length approach **mm**; again a very large distance for an atomic particle.
  - Some numerical values linking lifetime and diffusion length for Silicon can be found via the link
- Lifetime and diffusion length are direct measures of the cleanliness of **Si** with respect to "deep level" impurities which act as life time killers as directly evident from the Shockley-Read-Hall theory.

- There are many ways to measure the life time. A particularly unconventional approach giving the most precise data for the diffusion length is the "ELYMAT" technique, which is discussed in an advanced module.
- The ELYMAT and other tools like it have been used extensively in the nineties to "clean up" **Si** production and processing lines because even minute traces of contamination show up in a reduced diffusion length.

While it is generally important to maintain large diffusion lengths as a measure of cleanliness (and for some electronic properties) in integrated circuits, it is particularly important for **solar cells**.

- Any carrier generated by light that recombines in the bulk of the solar cell is lost for the external current that the solar cell is supposed to produce. Diffusion lengths thus must be larger than the absorption depth for solar light, otherwise the light generated carriers will not be able to diffuse to the the surface junction and produce current. What this means in practical (rather relaxed) terms is shown in the graph:



- For **CMOS** grade **Si** the requirements are much more stringent: Allowed levels of "life time killing" metals are much lower and specified at **< 5 · 10^{10} cm^{-3}**; i.e. at **<1ppt** (introducing now for the first time the ***ppqt*** - " **parts per quatrillion**" range).
- This is not so much because large lifetimes are so important, but because the metals killing the lifetime time tend to precipitate - and even extremely low densities (say **1 precipitate /cm^2**) of extremely small precipitates (say **10 nm**), if contained in a critical part of a device - e.g. at the **Si** - gate oxide interface - will kill a transistor and such the whole device. An example is given in the link.

Lattice Defects

Lattice defects in microelectronic **Si** are easy to deal with: *They are simply not allowed*!

- While it is indeed possible to avoid "larger" defects like grain boundaries, dislocations and sizeable precipitates in **Si** crystals (cf. the relevant chapters of the "Electronic Materials" and "Defects in Crystals" Hyperscripts), there is no way to avoid point defects or small agglomerates of point defects, also known as "***BMD***"s (**bulk microdefects**) or "***COP***"s (**crystal originated particles or pits**), or ***LLS*** (sometimes also abbreviated **LPDs**): **Localized Light Scattering Defect**.
- Why? Simply because point defects (vacancies and self interstitials) will be present in thermal equilibrium during crystal growth - and they cannot disappear at internal sinks like grain boundaries and dislocations because there aren't any; and the surface as an external sink is simply too far away for all but the surface-near point defects. The equilibrium point defects thus will be either "frozen-in" during cooling or form agglomerates which constitute the **BMDs**.
- Usually, these microdefects are few and small - it is not easy to detect them and often they are below the detection limits of the most advanced methods. Historically, however, they are periodically rediscovered because devices are becoming steadily smaller and more sensitive and their (always negative) influence on device properties is felt at some point.

As a curiosity, it shall be noted in passing that point defect equilibria in **Si** are much more complicated than in other elemental crystals and not very well understood up to this day.

- In particular, **Si** seems to be the only elemental crystal so far where self-interstitials are present in thermal equilibrium in concentrations that are comparable to vacancies (otherwise their concentration is always much lower). This implies that both vacancies and self-interstitials are involved in the formation of **BMDs** and that the diffusion of substitutional impurities (including all dopants) might be more complicated that usual.
- Some more details can be found in the "Defects in Crystals" Hyperscript.

3.1.2 Diffusion

General Remarks

If you are not very familiar with diffusion in general, it would be wise to consult some other Hyperscripts:

- Basic diffusion in "Introduction to Materials Science I" *(at present in German)*
- Point defects and diffusion in "Defects in Crystals"

The diffusion of dopants is of course one of the major topics in all process and device considerations. For any modern **Si** technology you must be able to have exactly the right concentration of the right dopant at the right place - with tolerances as small as **1%** in critical cases.

- And it is not good enough to assure the proper doping *right after* the doping process - what counts is only the dopant distribution in the *finished* device.
- Annoyingly, every time a high temperature process is executed after one of the doping steps, *all* dopants already put in place will diffuse again, and this must be taken into consideration.
- Even more annoying, the diffusion of the dopants may depend on the process - it may, e.g., be different if other dopants are present.
- A well-known example is the so-called emitter-dip or **emitter-push effect** which makes it difficult to achieve very thin base regions in bipolar transistors. The effect is due to a changed diffusion coefficient of **B** in the presence of **P**.

The only way to master diffusion in making devices is an extensive simulation of the concentration profiles as a function of all parameters involved - always in conjunction with feed-back from measurements. This requires a mathematical framework that can be based on three qualitatively different approaches:

- Use equations that describe typical solutions to diffusion problems and determine a sufficient number of free parameters experimentally. Observed but poorly understood phenomena may simply be included by adding higher order terms with properly adjusted parameters. This will always work for problems within a certain range of the experimental parameters for which the fit has been made - but not necessarily for other regions in parameter space.
- Solve macroscopic diffusion equations matched to the problem; i.e. equations of the type expressed in Ficks **1st** and **2nd** law. The input are the diffusion coefficients together with the relevant boundary conditions. This works fine if you know the the dependence of the diffusion coefficients on everything else (which you usually don't).
- Base the math on the proper atomic mechanisms. If all mechanisms and interactions are fully known, they will contain all informations and the results will be correct by necessity. Unfortunately, all mechanisms and interactions are not fully known - neither in **Si**, nor in all the other semiconductors.

So none of theses approaches works satisfactorily by itself - what is needed is a combination.

- In the eighties, e.g., it proved necessary to include diffusion mechanisms mediated by **Si** self-interstitials; a diffusion mechanisms not observed in most other materials.
- This would be not necessary for "simple" diffusion as expressed in Ficks laws with a constant diffusion coefficients - regular vacancy or interstitial mechanisms are not distinguishable at this level.
- Special effects, however, may occur and it is far easier to include these effects if the additional mathematical terms reflect the atomic mechanisms - the alternative is to add correction terms with adjustable parameters.

In any case, diffusion in **Si** (and the other semiconductors) is complicated and an issue of much research and debate. It has become extremely important to include all possible "classical" effects usually neglected because very high precision is needed for very short diffusion times (or penetration depth), but the atomic mechanisms of diffusion in **Si** are still not entirely clear.

- In what follows a few basic facts and data will be given; in due time some advanced modules with more specific items may follow.

Basic equations are the two phenomenological laws known as "**Ficks laws**" which connect the (vector) flux ***j*** of diffusion particles to the driving force and describe the local change in particle density, ρ ***(x,y,z,t)*** and the **Einstein-Smoluchowski relations** which connect Ficks laws with the atomic mechanisms of diffusion. Ficks first and second law are

First law:

$$\boxed{\boldsymbol{j} = - \boldsymbol{D} \cdot \nabla \boldsymbol{c}}$$

- With ***c*** = concentration of the diffusing particles, ***D*** = diffusion constant and ∇ = Napla operator. We have

$$\nabla c \;=\; \text{vector} = \left(\frac{\partial c}{\partial x}\,,\,\frac{\partial c}{\partial y}\,,\,\frac{\partial c}{\partial y} \right)$$

- *Second law*:

$$\frac{\partial c}{\partial t} \;=\; D \cdot \Delta c$$

  - With Δ = Delta operator (= ∇^2), and $\Delta \boldsymbol{c}$ given by

$$\Delta\, c \;=\; \frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} + \frac{\partial^2 c}{\partial z^2}$$

- An atomic view of diffusion considering the elementary jumps of diffusing atoms (or vacancies) over a distance ***a*** (closely related to the lattice constant) yields not only a justification of Ficks laws, but the relations

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

$$\nu \;=\; \nu_0 \cdot \exp - \frac{E^{\mathrm{M}}}{\mathrm{k}T}$$

  - With ***g*** = geometry factor describing the symmetry of the situation, i.e. essentially the symmetry of the lattice, and ν = jump frequency of the diffusion particle, E^{M} = activation enthalpies of migration.

- If the diffusion mechanisms involves intrinsic point defects as vacancies (**V**) or self-interstitials (**i**), their concentration is given by

$$n^{\mathrm{V,i}} \;=\; \exp - \frac{E^{\mathrm{F}}}{\mathrm{k}T}$$

  - With E^{F} = formation enthalpy of the point defect under consideration.

- The problem may get complicated if more than one atomic mechanism is involved. A relevant example for **Si** is the so-called **"kick-out" mechanism** for extrinsic point defects (= impurities):
  - A foreign atom (most prominent is **Au**) diffuses rather fast as interstitial impurity, but on occasion "kicks out" a lattice atom and then becomes substitutional and diffuses very slowly. However, the substitutional **Au** atom may also be kicked out by **Si** interstitials and then diffuses fast again. An animation of this process can be seen in the link.
  - The "kick-out" process is *not* adequately described by the simple version of the Fick equation given above.

- Since even the simple Fick equations are notoriously difficult to solve even for simple cases, not to mention complications by more involved atomic mechanisms, only the two most simple standard solutions shall be briefly discussed.

Diffusion from an Unlimited Surface Source

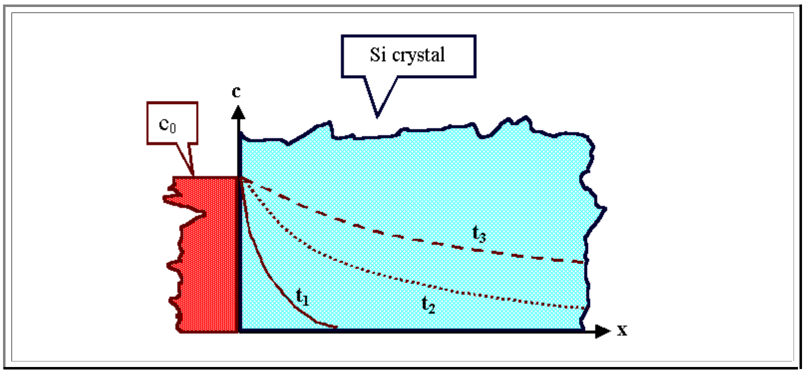
- Consider the following situation:
  - On the surface of a **Si** crystal the concentration c_0 of some dopant species is kept constant - e.g. by immersing the **Si** in a suitable gas with constant pressure or by depositing a thick layer of the substance on the surface.
  - The dopant will then diffuse into the **Si** and since the source of dopant atoms is the surface, there will be a drop in concentration of the dopant from the value c_0 at the surface to zero deep in the crystal.
  - Independent of the dopant concentrations outside the **Si**, the maximum concentration in the **Si** next to surface cannot be larger than the solubility of the dopant atom an the temperature considered; we take c_0 than as solubility limit.
- The general one-dimensional solution of the differential equations of Ficks laws for this boundary condition of an *inexhaustible source* then is given by

$$c(x,t) \;=\; c_0 \cdot \operatorname{erfc} \frac{x}{L}$$

  - With $L = 2(D \cdot t)^{1/2}$ = diffusion length, and **erfc(*x*)** = complementary errorfunction = **1 – erf(*x*)** and **erf (*z*)** = **errorfunction** given by

$$\operatorname{erf}(z) \;=\; \frac{2}{\pi^{1/2}} \cdot \int_0^z \exp - a^2 \cdot da$$

  - The errorfunction can not be written in closed form; its values, however are tabulated. A typical solution of the diffusion problem may look like this:



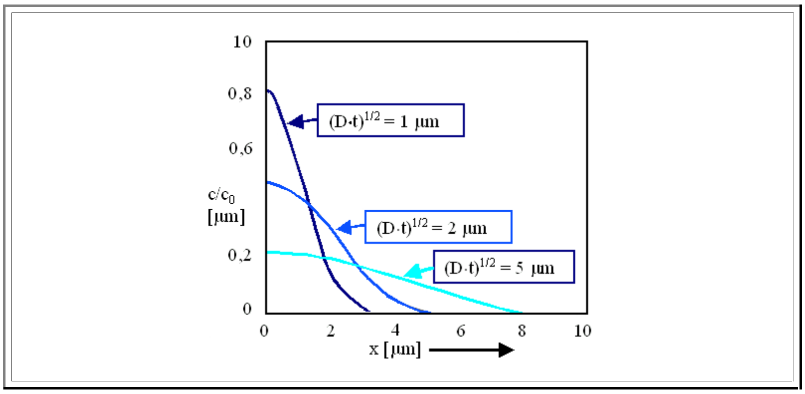
- The interesting quantity is the diffusion length ***L*** which is a direct measure of how far the diffusion particles have penetrated into the **Si**. At a distance ***L*** from the surface, the concentration of the dopant is about **1/2** c_0 or, to be exact **0,4795 ·** c_0.
  - This *diffusion length for dopants* or any other kind of atoms is not to be confused with the diffusion length of minority carriers as introduced before. Of course, the physics is exactly the same - the diffusion length for electron and holes as introduced before could just as well be obtained from solving the Fick equations for these particles.
  - Note in passing that while all definitions of diffusion lengths contain the $(D \cdot t)^{1/2}$ term, the factor **2** (or on occasion $2^{1/2}$) may or may not be there, depending on the exact solution - but this is of little consequence for qualitative discussions.
- The total quantity of dopant atoms now in the **Si** expressed as a concentration c_{total} can be obtained by integrating the "**diffusion profile**", i.e. the curve of the concentration versus depth. This is analytically possible if the integration runs from **0** to ∞ - a very good approximation for slowly diffusing atoms and thick wafers. The result is

$$c_{total} \;=\; \frac{L \cdot c_0}{\pi^{1/2}} \;=\; 0{,}56 \cdot L \cdot c_0$$

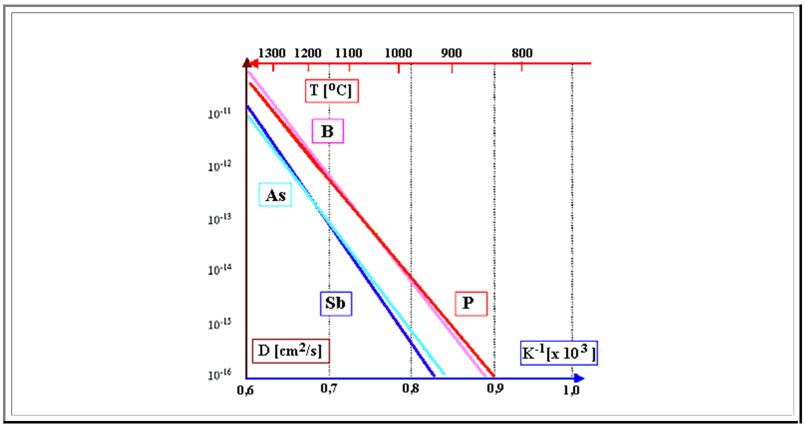
- The other standard solution for diffusion problems deals with the case of a *finite source*; i.e. only a limited amount of diffusion particles is available.
  - This is the standard case for, e.g. **ion implantation**, where a precisely measured number of dopant atoms is implanted into a surface near area.
  - For simplifying the math, we may assume that these dopants are all contained in one atomic layer - a delta function type distribution at the surface.
  - This is of course not true for a real ion implantation, where there is some depth distribution of the concentration below the implanted surface, but as long as the diffusion length obtained in this case is much larger than the distribution width after implantation, this is a good approximation.
- It is more convenient to resort from *volume concentrations* **c**of atoms to *areal densities* ***C*** because that is what an ion implantation measures: the total number of **P**-, **As**- or **B**-atoms shot into the wafer per **cm^2** called the **dose = atoms/ cm^2**. With C_0 = implanted dose and ***C*(*x*, *t*)** the area density in the **Si**, the following solution is obtained:

$$\boxed{C(x,t) \;=\; \left(\frac{C_0}{\pi \cdot D \cdot t}\right)^{1/2} \cdot \;\exp - \frac{x^2}{4D\,t}}$$

  - This is simply one half of a Gaussian distribution (the "–" sign in front of x^2 takes care of this) with a "half-width" of $(Dt)^{1/2}$; what it looks like is shown in the picture.



  - The curves can be characterized by a $(Dt)^{1/2}$ product, which again gives a typical *diffusion length*.
- The quantity of prime importance is always the diffusion coefficient of the diffusing particle. Only for "simple" mechanisms it is a simple function of the prime parameters of the point defect involved as implicitly stated above.
  - ***D*(*T*)** then follows a simple Arrhenius kind of behavior; examples for the common dopants are shown in the figure:



  - The lines shown are perfect straight lines over more than **8** orders of magnitude - provided there are no complications.
- The example of an ion-implanted layer as the source for diffusion, however, provides a good example for some of the complications that may be encountered in real **Si** diffusion:
  - *First* of all, if the distribution of implanted dopant atoms cannot be treated as a delta function, but must be taken into account as it is. Solutions then can only be obtained numerically - with some effort.
  - *Second*, if only a small area has been implanted through a mask, at least a two-dimensional problem must be solved - which is much more complicated.
  - *Third*, some dopant atoms will reach the surface after some random walk. The idealized solution assumes that they will go back into the bulk, i.e. the surface does not act as sink for diffusion atoms. This is, however, not always true and will lead to complications.

- *Fourth*, while all of the above still only amounts to a mathematical exercise in solving Ficks differential equations, there are physical problems, too: Ion implantation produces lots of surplus vacancies and interstitials which will become mobile during the diffusion procedure. The point defect concentration at the diffusion temperature thus is *not* identical to the equilibrium concentration (at least for some time), and the diffusion coefficient which always reflects the underlying atomic mechanism for equilibrium conditions, will be changed and become time dependent - a very messy situation!

In fact, the usual goal after ion implantation is to keep the implanted profile in place as much as possible - no diffusion would just be great. But you must get rid of the crystal lattice defects produced by the implantation and for that you must anneal at elevated temperatures for some time - and diffusion will take place!

- What is better: Long anneals at low temperatures or short anneals at high temperatures to remove the defects but keep your dopants in place. Not an easy question; the answer must depend on the kinetics of the defect annealing and the diffusion peculiarities of the atom under consideration.
- However, the second case is usually preferred, and a whole industry has developed around this point under the catch phrase "**rapid thermal annealing** or **rapid thermal processing** (***RTA*** or ***RTP***, respectively).

But there are more complications yet:

- The diffusion of an atom may be changed if there are noticeable concentrations of other foreign atoms around - and this includes the own species. **P**, as an example, diffuses faster in large concentrations and also enhances the diffusivity of **B** (a key word is: "emitter push effect").
- Some processes (notably thermal oxidation) produces non-equilibrium point defects (oxidation produces **Si** interstitials) which will be felt by atoms diffusing via these point defects - their diffusivity will be different if the **Si** is oxidized compared to an inert surface.
- Some atoms, as already mentioned above, diffuse by more complicated mechanisms, e.g. the kick-out mechanism. In a treatment with Ficks equations, this calls for two superimposed mechanisms, each with its own diffusion constant and some boundary conditions to assure particle conservation etc.
- A review from **1988** (which almost certainly will have been contested in the meantime in some points) covering just fast diffusing elements in Si and discussing some of the complications mentioned above, is provided in the link.
- Some prominent cases of deviations from simple diffusion behavior can be found in an advanced module

It should come as no surprise than that diffusion in **Si**, as far as the application to devices is concerned, is an active area of research and development, and that no process engineer will ever believe the results of a simulation for diffusion under a new set of conditions without experimental verification.

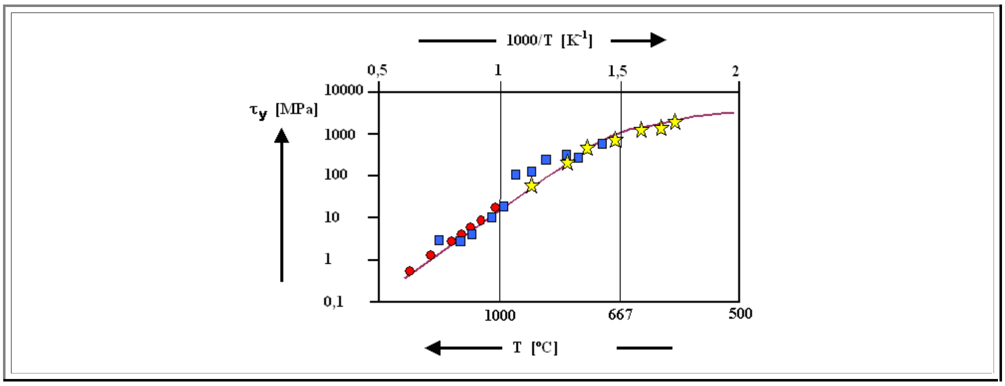
This leaves us with the question how diffusion profiles are measured - a large issue in its own right. We will deal with this only cursory, by giving catch words. You must look it up yourself if you want to know more.

- **Sheet resistance measurements**
  This is the easiest method, but it only gives the average resistance of the doped layer (always assuming that it sits on top of something pretty insulating).
- **Spreading resistance measurements**
  Here the resistance is measured as a function of depth by bevelling the sample. Cheap and easy in principle, but error prone and tricky for shallow profiles (say for diffusion length smaller than a few µm).
- **Direct doping concentration measurements by SIMS**
  Very expensive and slow - but the ultimate method.

3.1.3 Mechanical, Thermal, and Other Properties

Mechanical Properties

- Silicon at *room temperature* is brittle - and that is about all there is to know.
- Well not quite. *First* of all, fracture of **Si** is quite an interesting topic to many people (searching for "fracture+silicon" produces about **20.000** hits in the Net)
  - *Second*, there is brittle, and there is *very* brittle. What is the case for **Si**? While the microelectronics industry has learned not to break its wafers (that's the main reason why they are so thick), the crystalline **Si** (or ***Si-c***) solar cell industry cannot afford to waste **Si** and keeps its "multi-crystalline" **(10 × 10)cm^2** slices as thin as possible (about **300 µm**). As a result, breakage of the slices is becoming the major problem in industrial solar cell production. At present (**2001**) large research projects are started to find out more about fracture of (multicrystalline) **Si**.
  - Fracture toughness, to give some numbers, has been reported to be **1.19 MPa · m^{½}** for the **{100}** tensile surface and **1,05 MPa · m^{½}** for the **{111}** tensile surface - there are certainly other numbers out there, too.
  - If we compare that to, e.g., the fracture toughness of *Steel* (≈ **200 MPa · m^{½}**), *Nylon* (≈ **3 MPa · m^{½}**), or ceramics like *Silicon nitride* (**Si_3N_4**, *Silicon carbide* (**SiC**) or *Alumina* (**Al_2O_3**) which are all ≈ **(3 - 5) MPa · m^{½}**, or common *glass* (≈ **0,8 MPa · m^{½}**), we see that **Si** is about as brittle as glass and "more" brittle than some of the tougher ceramics.
- Then we have the emerging "MEMS" industry, i.e **Micro Electronic and Mechanical Systems**. Obviously, the mechanical properties, especially the elastic coefficients and the elastic moduli are of prime interest (besides fracture, which simply must be avoided).
  - Youngs modulus, is given as **131 GPa** (or **107 GPa**, or ...; it depends on the source), which again should be compared to that of *diamond* (**1 000 GPa**, the biggest there is), *"hard" steels* (≈ **200 GPa**), *ceramics* (as above; ≈ **400 GPa**), *Silver* and *gold* (≈ **80 GPa**), or *Nylon* (≈ **3 GPa**).
  - All in all, Silicons elastic properties are pretty good, but not breathtaking.
- But let's not forget that in most processes we heat up the **Si** to temperatures somewhere between **700 ^{o}C** and **1200 ^{o}C**, and at high temperatures **Si** is no longer brittle but **ductile**, i.e. it deforms plastically.
  - Plastic deformation is always characterized by the **yield stress τ_y** which describes macroscopically the minimum shear stress needed to induce *plastic deformation*, and microscopically the minimum stress needed to induce *dislocation movement*.
  - In a somewhat simple minded approach, we see that **Si** is brittle and fractures if the yield stress τ_y is larger than the stress needed for facture under the conditions used. Since the yield stress decreases with increasing temperature, we expect that plastic deformation takes over at some temperature.
  - The figure below gives a compilation of τ_y data for **Si** (from a paper by J. Rabier and J.L. Demenet; phys stat sol (b) 222, 63 (2000)).



- In short, you must expect plastic deformation if the temperature is above, say **700 ^{o}C** and you have some stress acting on your **Si**.
  - If you started with a typical, completely dislocation free wafer, the initial τ_y might be somewhat larger because you first must generate some dislocations (which then can multiply).
  - If some dislocations have been generated, however, its all over. You will experience plastic deformation and you will have introduced dislocations irreversibly into your material.
- Where do stresses come from? There are *two* major sources:

- **1.** Running a batch of **Si** wafers (or slices, or whatever) into a piece of equipment that heats up the **Si**, the outside of the wafer will always tend to be hotter than the inside during heating up, and vice verse during cooling down. Thermal expansion will be different in different parts of the specimen, and that introduces stresses directly proportional to the *temperature gradient* in the sample. There is nothing you can do except keep the temperature gradients below the critical level. Typical tricks are to move into some oven slowly, to have the equipment at some lower temperature and then go up slowly after the **Si** is inside, or to do both.
- **2.** If there are any layers on the **Si** (or inside, e.g. heavily doped regions), differences in thermal expansion coefficients will generate stresses at sometimes very large levels. Again, while you cannot avoid the stress produced by the layer you need, you can use some tricks to minimize the over-all stresses: Keep the area small (provide holes in the layer where it is not needed); put the same kind of layer on the backside **-** the combined effects cancel to some degree, and be generally aware of what you have on the backside (many layers are automatically also deposited on the backside, and there are good and not so good times in a process sequence when you can take them off).

If everything fails, you may want to try out some **Si** with a relatively high concentration of interstitial oxygen, $\mathbf{O_i}$ (say around **20 ppm**).

- These impurity atoms make dislocation movement more difficult (and thus increase τ_y) without degrading electronic properties too much (if you are lucky).

Last not least we note that the exact process of dislocation generation and movement, with particular respect to the details of the dislocation fine structure, is of major interest to basic research, since ultra-perfect **Si** is an ideal proving ground for theories concerning deformation and dislocations in covalently bonded crystals.

Thermal Properties

The most important thermal properties of Silicon are:

- **Thermal expansion coefficient** α_T,
- **Thermal conductivity k**, and, to a lesser extent
- **Specific heat** $\mathbf{c_p}$

The *thermal expansion coefficient* is crucial whenever other materials are in contact with **Si**, or if temperature gradients are encountered. In both cases it is rather easy to destroy the device by building up large mechanical stresses leading to fracture or plastic deformation. Lets see why

- **Si** is in contact with other materials *either* during processing (when, e.g., a layer of $\mathbf{SiO_2}$ or $\mathbf{Si_3N_4}$ or **Al** or ..., is deposited, often on the front *and* backside, *or*, as a finished device, when it is encapsulated in some housing, i.e. when it is packaged.

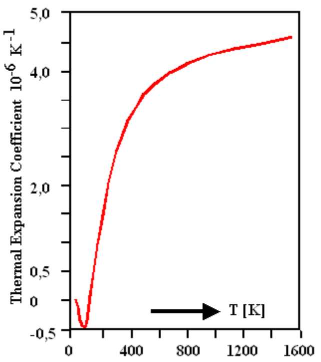
If temperature changes for a **Si** / other material *compound* **-** because you use your cellular phone during skiing and in the summer, or because processing the wafer intrinsically needs high temperature **-** mechanical stress, being roughly proportional to the difference in thermal expansion coefficients and the amount of material present, is simply *unavoidable*.

- The art of processing and packaging thus includes to keep the stress levels *always* below the critical stresses required to induce plastic deformation or fracture. You may have to optimize the thermal expansion coefficient of materials in contact with **Si** **-** if you can.
- The black polymer compound, for example, that is universally used for cheap packaging, while still very different in thermal expansion, is matched as much as possible. Since this is still not good enough, the **Si** chips are usually ground down to a thickness far below the original wafer thickness.

As we have seen above. if there is a *temperature gradient* in a piece of pure **Si** (e.g. a wafer), the hotter parts want to expand more then the cooler parts **-** again a mechanical stress is induced that is proportional to the temperature gradient *and* the thermal expansion coefficient.

- This is why you heat up the wafers slowly, giving enough time for temperature equilibration. Otherwise, plastic deformation will occur, ruining your chips and leaving a wafer that is no longer flat **-** wafer **warpage**, one of the absolutely deadly wafer diseases, has occurred

Here are data of the *linear* (as opposed to "*volume*") thermal expansion of **Si** and some comparison to other relevant materials



After Okada and Tokumaru, 1984

- At low temperatures, α_T shows a pronounced minimum; an effect not easy to understand in cubic crystals, but nevertheless observed in most of the other cubic semiconductors, too. If we take the room temperature value of about **2,5 · 10^{-6} K^{-1}**, and compare it to the values of other materials important in **Si** technology, we have

Material	Si	Ge	GaAs	SiO_2	Si_3N_4	Al	Polymers
α_T **[10^{-6} K^{-1}]**	2,5	5,8	6,86	0,5	3,2	24	ca. 50 ... 200

The *thermal conductivity* κ is important, because **Si** chips, like most semiconductor devices, are producing lots of heat in operation. It must transported out of the system and the resistance to heat flow is given by the thermal conductivity of the material.

- What do we have? Here is the **Si** value, again together with thermal conductivities of other relevant materials. Note that κ is strongly temperature and structure dependend. For **Si_3N_4**, as an example, values may scatter from **0,2 - 1,2 W·cm^{-1} · K^{-1}** .

Material	Si	Ge	GaAs	SiO_2	Si_3N_4	SiC	Diamond	Cu; Ag
κ [W · cm^{-1} · K^{-1}] at room temperature	1,5 1,4	0,6	0,46 0,54	ca. 0,014	ca. 0,2	ca. 3.5 (3 - 5)	ca. 10 - 30	4

- **Diamond**, surprisingly, is the champion - and that is why thin diamond layers are sometimes used to transport the heat generated in some device to some heat sink as efficiently as possible.

While it appears that you just must live with whatever thermal conductivity a material has - this is not entirely true. The thermal conductivity can be made substantially better, if the material is made from *one isotope only*!

- Here is a recent (may **2002**) topic form the news ticker in the semiconductor business.

Isonics Delivers Silicon-28 SOI Wafers
Online staff -- 5/2/2002
Electronic News

Isonics Corp. of Golden, Colo., today said that a major semiconductor manufacturer has taken delivery of Silicon-28 silicon-on-insulator (SOI) wafers for evaluation.
Isotopically purified silicon-28 has 60 percent more thermal conductivity than natural silicon, Isonics said. This allows for reductions in the self heating of circuits made with natural SOI wafers. Incorporating Silicon-28 into SOI wafers made using either oxygen implantation or layer transfer technologies requires no change in the manufacturing processes developed for these wafers, the company said.
"While SOI wafers are known to reduce power requirements for devices, heat has been a large concern for certain applications and is expected to become an even larger, more critical consideration as chip manufacturers continue to push for more performance," said Stephen J. Burden, Isonics VP of semiconductor materials, in a statement.
"Semiconductor manufacturers, eager to design the optimum thermal/electrical solution for their specific device, are becoming aware of the outstanding performance offered by the marriage of our high thermal conductivity silicon-28 and the film SOI wafer technology.
The wafers delivered to this customer were manufactured in cooperation with an existing thin-film SOI wafer supplier, Isonics said, instead of the company’s thick-film SOI facility.

From "Semiconductor International 5-2002"

- The *specific heat* c_p (of course for constant pressure) is not so important, but here are values anyway. In addition the **Debye temperature** Θ is shown, too

Material	Si	Ge	GaAs	SiO_2	Si_3N_4	Al
χ_π [J/g · K] at room temperature	0,7	0,31	0,35			0,9
Θ [K]	645	374	362			

Other Properties

- Here is a table found in the Net with some more non-electrical data
  - Some of the numbers deviate from the numbers given above; e.g. the thermal expansion coefficient.
  - That is just the way it is - if you look up anything, you will find different numbers. Often just a little bit different (melting points, for example), but sometimes quite different (as in the thermal expansion coefficient here).
  - Reasons for this might be:
    1. The quantities compared are actually different. In the table above, the *linear* thermal expansion coefficient is given; in the table below it might be the expansion coefficient for the volume?
    2. Watch out for units. Conversion can be tricky, especially for some quaint old British imperial units still much beloved by the Americans, too.
    3. The samples might have been different. Giving the "conductivity" or "lifetime" for **Si** without some comments, obviously does not make much sense. How about other properties?
    4. The number is simply wrong.

Si Properties	
Refractive Index	3.4179 @ 10 µm ; 3,45
Reflective Loss	46.1 % @ 10 µm
Density	2,3291 g/cm^3
Melting Point	1420 °C
Molecular Weight	28.086
Thermal Conductivity	1,63 W/(cm K); 1,4 W/(cm K)
Specific Heat	0,703 J/(g K) @ 25 °C
Thermal Expansion	4.05×10^{-6} / K @ 10...50 °C
Hardness (Knoop)	1150 (Mohs 7)
Young’s Modulus	131 GPa
Shear Modulus	79.9 GPa
Bulk Modulus	102 GPa
Rupture Modulus	340 MPa
Elastic Coefficient	C_{11} = 167 / C_{12} = 65 / C_{44} = 80 GPa
Dielectric Constant	13 @ f = 9.37 GHz

3.2 Silicon Production

3.2.1 Single Crystals and Wafers

- This is an easy subchapter because the major points concerning single crystalline Silicon production for microelectronics (and some other uses) - from making the raw (or metallurgical grade) **Si**, **poly-Si** for crystal growth, single crystals, and wafers - have been covered in the Hyperscript "*Electronic Materials*" and here we provide just a link to the starting chapter.
    - In what follows in the next subchapter, we will look briefly at methods to produce **Si** for applications other then microelectronics - which means mostly solar cells.
    - Link to starting chapter
    - Here is an overview of what can be found in " *Electronic Materials*" with direct access to all modules.
- What can you buy? Activating the link will tell.

Basics	Backbone I	Backbone II	Illustrations	Exercises	Advanced
Hyperscript: Electronic Materials					
6. Materials and Processes for Silicon Technology					
6.1 Silicon					
	r6_1_1 Silicon **r6_1_2** Silicon crystals	**r6_1_3** Other Silicon uses	**i6_1_1** CZ crystal growth **i6_1_2** Si crystal **i6_1_3** Complete wafer process **i6_1_4** Poly-Si Specs **i6_1_5** Wafer Specs **i6_1_6** Necking		**t6_1_1** Alternative poly-Si productions **t6_1_2** Crystal growth - science & art **t6_1_3** FZ crystal growth **t6_1_4** Biography Czochralski **Article** Historic review Si **Article** New developments Si crystals

3.2.2 Silicon for Solar Cells

Personal "Historical" Remarks

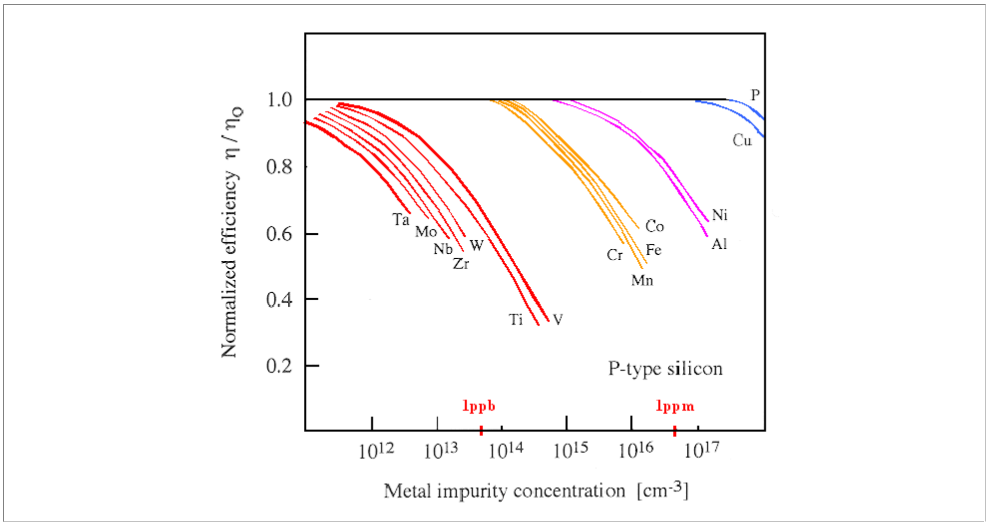
- Having worked in the development of solar **Si** from **1977** up to the present day, I remember a sizeable number of projects that tried to make progress in the following areas:
  - Making "**solar-grade**" **Si** in *cheaper* ways than with the conventional process used for microelectronics **Si**.
  - Transform solar grade **Si** *directly* (no cutting) into suitable substrates for **solar cells** (square or hexagonal plates, rectangles, or ribbons; typically about **0,5 mm - 0, 3 mm** thick) at low costs - meaning minimal losses of the starting material and simple processes.
  - Making solar cells out of the substrates at *low costs* - but with very good performance.
- The ingenuity - not to mention the money - that went into solving these questions is nothing less but amazing. There were (and still are) a large number (far more than anybody could have imagined) of approaches that have been tried - and most have been abandoned by now!
  - While recalling all these "failures" is not exactly necessary in the context of this lecture, it would be highly educational in a general context of alternative energy development. A quick scan of the readily available literature (including some **1000** pages of the proceedings of the international conference on solar energy (Vienna, **1998**) and **7** books) showed that many of the "old" approaches seem to be forgotten.
  - I will not make a concentrated effort to write the historical review that I could not find in the general literature in this Hyperscript. Only a few general remarks, illustrated by some readily available examples (to me), will be given in the backbone part. However, I might start some modules in the "advanced" section in which more historical stuff will be collected in due time.

General Requirements for Si Solar Cells

- A "solar cell primer" explaining briefly some of the more basic features of solar cells and solar energy, can be found in the link.
- When considering solar cells, all that counts in the end, is the *prize for **1 kWhr** electricity* produced with the device. While it is not particularly easy to come up with this number, it is of course directly related to the production costs of **solar cell modules** (the assembly of solar cells in a frame) and thus to production costs of a single solar cell.
  - The costs of a module - typically **(1-2) m^2** in size - can be broken down into three main components:
    35% for the solar **Si** material (ready for cell production)
    30% for the solar cell technology (making a solar cell)
    30% for module manufacture.
  - If one uses the conventional process to make single crystalline wafers as the starting material for solar cells, the costs can be broken down as follows:
    30% Starting material (**poly-Si**)
    35% Crystal growth
    30% Sawing the crystal into wafers (which also turns almost **1/2** of the expensive crystal into saw dust!)
    5% for refining (etching, polishing, cleaning).
  - In other words: While **2/3** of the costs are directly caused by **Si** and **Si** technology, there is no overwhelming single cost factor - you have to work from all angles to reduce costs.
- How large is the **market** for **Si** solar cells?
  - In **1998** solar cells based on crystalline Si producing **130 MW$_p$** (the index "**p**" refers to "peak" power) electrical power were sold, another **20 MW$_p$** or so was based on amorphous Si.
  - Considering that **1 m^2** of a solar module - very roughly - produces about **100 W$_p$** ; this translates into **1 300 000 m^2/a** of solar **Si**, or - with an average thickness of **300 µm** - into **390 m^3** of **Si**, or - with a density of **2,33 g/cm^3** - into **675 to/a** of **Si**.
  - The world production of semiconductor grade "raw" **Si** is roughly **20 000 to/a** (at very roughly **$50/kg** equalling **1 G$/a**; while the wafer business grosses something like **5 G$/a**). Solar **Si** thus accounts for less than **5%** of the eligible **Si** production - small wonder that solar **Si** is usually taken from the "left over" of the microelectronic business. *Use this link for somewhat newer information about these topics*.
  - More and newer numbers can be found in the link.
- The growth rate of solar **Si**, however, is larger than the growth rate of **Si** for microelectronics. Sooner or later, the solar **Si** community will have to make its own **Si** and this will be a major milestone in the history of **Si** production.
- Let's now look at some of the major processes on the road from sand to solar modules:

"Solar Grade" Starting Material

- As mentioned above, starting materials for solar **Si** is usually recruited from "left over" of microelectronic **Si** (**ME-Si**)
  - Those "left over" might be wafers that were rejected, the seed- and end cones of the crystals, or parts of crystals that did not meet specifications.
  - Some former **ME-Si** producers, no longer able to compete (i.e. **Si** producers form the former **UDSSR**), now sell their products as "cheap" solar **Si**.
  - "Cheap" does not mean "dirty" or otherwise degraded with respect to the life time or diffusion length. A "good" diffusion length ***L*** of the minority carriers (say at least a **100 µm**) is absolutely mandatory. What this means in terms of permissible impurities was shown before, here similar data are displayed directly for solar cell performance.



- Efforts to produce dedicated solar **Si** *directly*, essentially can exploit three ways:
  - **1.** Take cheap metallurgical grade **Si** (**MG-Si**) and find a cheaper process than the standard process to convert it to **poly-Si**. Many attempts have been made; but nothing seems to work so far.
  - **2.** Use the process for making **MG-Si**, but use sufficiently pure **SiO_2** and **C**, which will lead to relatively pure **Si** directly usable as the starting material for making solar cells. The trick is to obtain the clean ingredients cheaply (this could mean that you must make them yourself). Siemens AG spend quite some time (and money) exploring this route. The approach worked, but not necessarily cheaply, and was abandoned about **1985**.
  - **3.** Use a different reduction process for **SiO_2**. The only competitor, it appears, is the **aluminothermic reduction**

$$\boxed{3SiO_2 + 2Al \Rightarrow Al_2O_3 + 3Si}$$

  - Since **Al** is an acceptor, the resulting **Si** can be expected to be strongly **p**-doped - not acceptable for **ME-Si**, but fine for solar **Si** (which we want **p**-doped anyway).
  - While it is not directly obvious why this process should be better (mostly meaning cheaper) than the conventional process, it has been developed by Bayer AG and is ready for production - but is not used at the present time (Nov. **2000)**.
- Whichever way you produce your solar **Si**, next you must make thin sheets or slices (the name "wafer" is reserved for the **ME-Si** product), possibly by first converting your solar **Si** to something bulky material, or by making "flatware" directly. Typically you want a **10 cm x 10 cm x 0,03 cm** slice of **p**-doped **Si** for the production of the individual solar cells. This is probably the field where most new methods were developed as outlined below.

Cutting Slices from Bulk Si

- This is the main technology for making commercial **Si** solar cells. There is a bewildering multitude of processes, but all belong to two basic classes:
- *Conventional" Technologies* : Grow a (mostly single) crystal and slice it. This version accounts for (roughly) one half of solar cells sold today (Nov **2000**). Compared to **ME-Si**, money can be saved along the following lines:
  - Use less pure **poly-Si** and grow the crystal with relaxed specifications. Allow larger concentrations (and especially concentration variations) of dopant, oxygen, some impurities and tolerate some dislocations.
  - This allows for higher growth rates **-** after all it is the amount of **m^2/min** of solar cell material that defines the productivity of your factory and thus determines how many (expensive) crystal pullers you must have.
  - Some special tricks include the growth of "tri"-crystals (three single crystals joined by defined grain boundaries, or the growth of crystals with hexagonal cross section.
  - Slicing the crystals with inner-diameter diamonds saws **-** the conventional way for **ME-Si -** takes time and wastes a lot of **Si**, but produces superior flatness which, however, is not needed for solar cells. **Wire saws** with many parallel wires are used instead.
  - Polishing to a mirror finish **-** de rigeur for **ME-Si -** is abandoned in favor of a chemical etching procedure that simply removes the layer damaged by the cutting.
- ***Casting*** **-** followed by slicing; or casting directly into some kind of flat shape. The first version is now a standard technique, it accounts for (roughly) the other halve of the **Si** solar cells sold today. If you think that casting **Si** seems to be an obvious and easy way for cheap production **-** you are dead wrong: Casting **Si**, in contrast to practically all metals, is *very difficult* because:
  - **Si expands** by almost **10 %** upon solidification **-** almost everything else contracts. What happens then is known from water (**H_2O**) which is one of the few substances with the same "*melting point anomaly"* : Your water pipes and hoses, if left with freezing water in the winter, will explode without fail **-** the pressure built up during solidification will easily destroy almost any container. The only way left for casting is to make sure that you have a **directional solidification -** the liquid **Si** must always be on top of the solidified one and *never* trapped inside something.
  - **Si sticks to the walls** of the mold because it is still under internal pressure (if it would contract like most other materials, it would retract from the walls) *and* it reacts with almost anything. Getting it out of the mold (which you would like to reuse) is not easy and calls for special materials and procedures.
- Casting thin sheets in moulds therefore is practically impossible **-** what has been tried is:
  - "**Spin casting**", pioneered by the Hoxan Corp., Japan. Here a drop of molten **Si** falls on the center of a rapidly spinning wheel so that the liquid **Si** is immediately pulled out to a thin sheet which then crystallizes. Solar cells of good quality have been obtained; but it remains to be seen if the process makes it to large scale production.
  - "**Band casting**" by rapid quenching in analogy to the production method of amorphous metals. A constant stream of a molten **Si** jet impinges on a rapidly turning cooled metal wheel. The **Si** solidifies rapidly and a thin continuous band (at speeds of several **m/s**) is produced. The technique has been tried in a cooperation of Siemens and "Vacuumschmelze Hanau", but the resulting **Si** bands were of poor quality and nothing came from it.
- The method of choice then is casting big blocks of (necessarily) poly-crystalline **Si**, slice them with a wire saw, and make solar cells as as best as you can **-** considering that the material contains grain boundaries and lots of other defects which are generally detrimental to the performance of solar cells.
  - Several companies **-** most prominently "**Bayer Solar**", a company formed of the two formerly independent producers Wacker Chemitronic and Bayer AG, but also several other companies, produce poly-crystalline solar **Si** in this way which is often sold to other companies that produce the solar cells.
  - All possible problems notwithstanding **-** good solar cells with efficiencies of **12% - 14%** are routinely manufactured and this is not much worse than what one would get with single crystalline Si and standard processes. An example is shown in the link.

Conversion to "Flat" Silicon

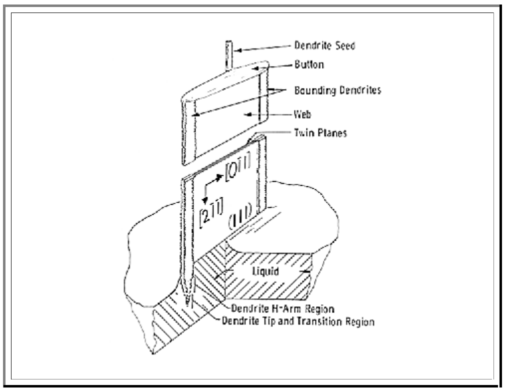
- Cutting crystals of any kind into slices is wasteful and expensive **-** large scale efforts to convert solar grade **Si** to "flat" **Si** in one fell swoop thus has been (and still is) a constant in the world of **Si** solar cells since about **1975.** Usually "flat" means some kind of long ribbon, but defined rectangles (or, as in the case of the melt spinning mentioned above), round "wafers" were also considered.
- Many different ways were explored **-** few are still with us, and only one (? I'm not quite sure about this) has made it to a production status **-** the "***EFG***" process, see below.
- A basic classification of the many ways is:
  - Technologies *not using a substrate* of some kind **-** producing "free standing" **Si** ribbons

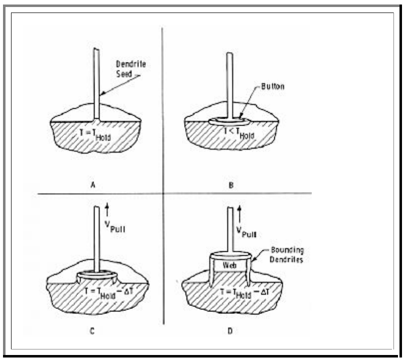
- Technologies using a *substrate* of some kind - producing **Si** layers on some other material or free-standing **Si** after separation from the substrate.

The *first category* contains tricky crystal growth methods like:

Dendritic web growth: A pretty much single crystal sheet is grown from the melt (not unlike the conventional **CZ** procedure), except that growth is restricted to a thin sheet expanded between two somewhat thicker **Si** "dendrites" as illustrated below.

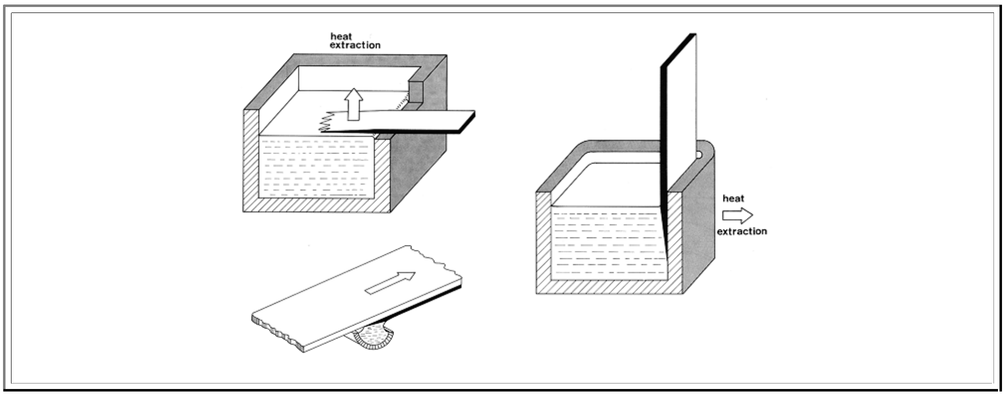
- This is something for experts in crystal growth techniques; it was tried, worked, and produced good solar cells - but it was abandoned (too expensive).
- Below a few schematic sketches of what is involved.





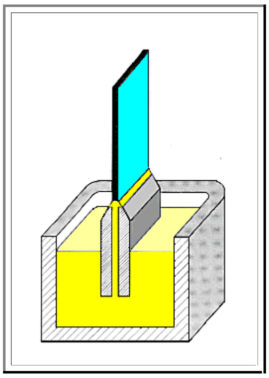
Ribbon growth directly from the melt by pulling out the crystal in an horizontal arrangement - sort of "over the rim".

- The solidification interface is steeply inclined and thus rather large which makes it easier to remove the heat of crystallization; high growth rates have been achieved.
- However, this is a tricky process difficult to run stable for a long time. And if anything goes wrong, it is very time consuming to start it again. In summary: Too expensive.
- Variants have been tried, too; below three basic principles without any further comment.



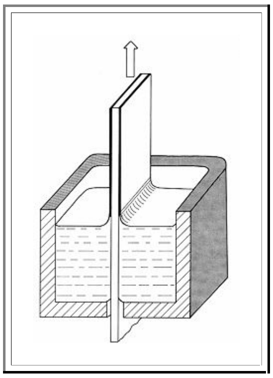
The "**EFG**" method, i.e. the "**Edge defined film-fed crystal growth technique**". This is the one (and so far only (?)) method that made it to production scale. It was started by on offspring of Mobile (oil company) called Mobile Tyco, but now is owned by **ASE**, a German based Company.

- In essence, it is a modified **CZ** crystal growth technique. The crucible filled with liquid **Si** contains a (graphite) die or nozzle and the crystal is pulled form a thin slot on top of the nozzle; well above the level of the liquid **Si** in the crucible.
- The Si used up in the crystallization is fed through capillary action to the surface of the nozzle. This looks schematically like this:

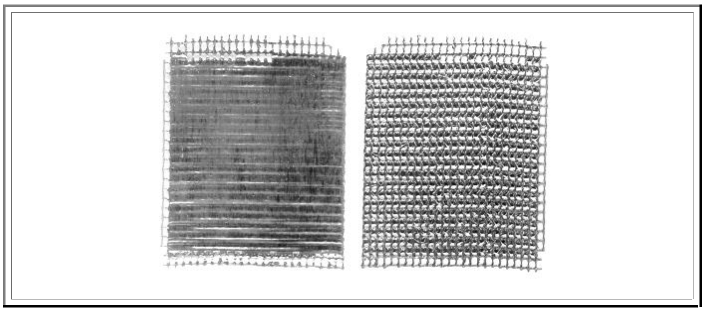


- Relatively long ribbons of good quality could be made after several years of development.
- The process was improved and became more economical by "bending" the die into a closed form, for reasons not clear to me preferably into an nonagon. The nine-sided tube is then cut along the edges and good quality material (almost, but not quite single crystalline) is obtained (if good solar **Si** is used as feedstock). Presently developments are under way to grow tubes with about **1m** in diameter!

- The *second category* contained the following methods:
- The "***S-web***" technique from Siemens AG (my first job at Siemens).
  - Here a net (mesh about **5 mm**) from high-purity carbon fibres (not cheap) is led through a solar-**Si** melt. The idea was that liquid **Si** is drawn out in the meshes of the net and stabilized by its high surface tension. It than can crystallize at leisure - the sheet forming process and the crystallization process are essentially decoupled and the removal of the heat of crystallization is no longer the limiting factor in "productivity" (**= m_2** of **Si/min**)
  - The basic principle is shown below; it can be used for a variety of substrates, not only for carbon nets, but also, e.g. for carbon "paper" (tried by a french group in the eighties) or ceramic substrates.



  - Of course, everything can be combined. **S-webs** were pulled out of dies, or by moving the net horizontally over the melt.
  - Well, the **S-web** technique did not work as originally envisioned - but it worked eventually! Long ribbons grown at high speed (up to **1m/min**) and up to **10 cm** in width could be grown; the pictures show an example. The remains of the net (it turns into **SiC**) on the backside are clearly visible on the right-hand side view.



  - Amazingly enough, the crystal quality was not as bad as it could have been - the net did not induce a lot of defects like dislocations etc. - and solar cells with efficiencies **> 10%** have been made from **S-web** material.
  - But as ever so often - there was no real advantage in prize and the technology was abandoned around **1985.** That was also true for most of the other methods touched upon above, (and for several methods not even mentioned).
- What do we learn from this? It is not so easy to make cheap solar cells from crystalline **Si**! Many billions of **$** have been spent on the effort, and untold master- and PhD theses were written. But much has been learned and the search goes on - with different materials, but also new ideas for crystalline **Si**.

Specialities

- A rapidly growing segment in the research part of solar energy are "**Thin film Si solar cells**"
  - Here a "thin" film (actually several µm to **15 µm** thick, i.e. not "thin" in the usual meaning of micro electronic technology) of **poly-Si** is deposited on a suitable substrate (e.g. some glass with a layer of ***ITO*** for the necessary electrical contact), a **pn**-junction and a contact is added to obtain a solar cell.
  - Since the cell is too thin to absorb all radiation in the infrared, the surfaces (or interfaces) should be textured and contain lots of microrprisms which reflect the radiation back into the **Si**.
  - This topic will be covered in more detail in the seminar
- Several groups (Canon, Sony and a research institute of the University of Stuttgart) use *Porous Si* in conjunction with *Waferbonding* to make solar cells.
  - *(to be filled in later)*
- Not so obvious approaches were:

- Production of suitable **Si** plates by **sintering** fine **Si** powder **-** akin to the standard process for making ceramics. This was a process pursued by Siemens in the early eighties and then carried over to a research Institute in Freiburg/Germany. While it works, it does not seem to offer clear advantages and was abandoned
- Produce **spherical solar cells** by making small spheres (in the **mm** range) of **Si** that contain a **pn**-junction on the outside (let drops of liquid **Si** fall down a tower (in vacuum of course) and add some dopant gases near the bottom). Then fuse the little spheres to flexible sheets, provide contacts (tricky!) **-** and you have a big solar cell, a whole module, in fact.
- This process was pioneered by Texas Instruments (to some extent, it appears, because it was the brain child of Jack **Kilby**, the famous co- inventor of the integrated circuit who just (**2000**) got the Nobel prize (the other inventor was Robert **Noyce** from Fairchild (which spawned Intel among many other companies)). While highly advertized in the late eighties and early nineties, it was recently sold to "Ontario Power" (where I lost its trace).

3.2.3 Crystal Lattice Defects in Si

- Strangely enough, while single crystalline **Si** as used in the micro electronics industry is the most perfect material in existence (at least on this side of Pluto), investigations of the possible lattice defects (point defects, dislocations, grain boundaries and so on) fill many volumes of scientific literature.
  - For the time being, however, we will not delve into this subject here in any great depth but refer to some links
- **Point defects** in **Si** are of prime importance for three independent reasons:
  - They are the vehicles for the all-important diffusion processes of dopants and other foreign atoms. Since there is a definite (and very unusual) contribution from self-interstitial present in thermal equilibrium, diffusion in **Si** is much more complicated than in other elemental semiconductors and still an object of study.
  - They are unavoidable defects because they are present in thermal equilibrium. This puts limits to the perfection of **Si** single crystals, and the microdefects introduced during crystal growth by agglomeration of point defects are a major object of investigation. The link leads to a recent article reviewing this part of defects in **Si**
  - Last not least: As long as we do not really understand point defects in **Si** - and we still don't - there is a need for basic research for basic research sake. After all, **Si** is not such a complicated crystal and if do not know everything there is to know about its thermal equilibrium defects, there is little hope to understand this in more complicated materials.
- "Larger" defects, i.e. **dislocations**, **grain boundaries** and **precipitates**, are also objects of intense investigations for the following reasons:
  - They are present in **poly-Si** that is used for, e.g., solar cells, and they will influence the technology and the device properties. It was (and still is) a major question if these defects are "electronically active" (meaning that they act as recombination centers), and what determines the level of those activities.
  - They may be formed during the processing of **Si** - mostly inadvertently with deadly effects, but sometimes intentionally (for the so-called "intrinsic gettering"). More about this subject in the link (and in the links going out from there).
  - They are studied in basic research to learn more about defects in covalently bonded crystals - the precise atomic structure of grain boundaries, e.g., is far from being completely understood.

3.3 General Device and Product Considerations

3.3.1 Interfaces and Contacts

This Module is unfinished

General Remarks

- Practically all **Si** devices and all other solid state devices have properties that are "*interface controlled*".
  - The "**pn**-junction" is an interface and so is the **MOS** contact. The latter, in fact, involves two interfaces: **Si**-oxide and oxide-metal. But there are many more interfaces: Metal-**Si** contacts are needed to connect the device to the outside world and there are many interfaces between the various layers in integrated circuits.
  - What kind of interface properties do we want to have? This is not a simple question: In the case of the **pn**-junction we certainly want the "**metallurgical junction**" (the area of actual contact between the **p** and **n**-type Si) to have *no structural properties* - it should be impossible to find it analytically as far as the **Si** is concerned. Of course, the metallurgical junction of a a **pn**-junction formed by diffusion is not really definable; but you also could make a **pn**-junction by epitaxy and then the metallurgical junction is wherever the surface of the substrate *was*.
  - On the other hand, it would not be too far fetched to assign all the *electronic properties* of a **pn**-junction to the interface (one certainly would do this for a metal-metal junction).
- Usually we would like a metal - **Si** contact to be *ohmic* - and not of the Schottky type. This again might be seen as an interface property. For both electronic properties mentioned we would like them to be as independent from structural properties as possible. In other words, the precise arrangement of the atoms in the interface, the presence of interface dislocations, etc., should not matter much.
- Then we have structures where we hope that the interface does not have any properties relevant to the device function.
  - Interfaces for a **MOS** contact, e.g., are not really required for the function - theoretically we could remove the dielectric and run the device in vacuum with somewhat changed parameters because of the changed dielectric constant. The question now is if there are interface properties that interfere with the device operation.
  - We certainly would tend not to worry about the properties between some insulating dielectric layers or call for "no" properties - but that is not correct, we definitely want that the stick together solidly, or, in other words, we want some bonding or adhesion.
- We thus must ask ourselves: What are the properties an interface can have and how do those properties influence or even enable device operation.
  - What are the basic properties an interface can have?
  - What is the variation range of a given property and how can it be influenced or tuned to specific needs?
  - How do specific properties interact? For example, how does the atomic structure correlate to electronic properties, e.g. states in the band gap?
- Looking at interfaces in this kind of generality, we are opening a rather big can of worms. However, we will stay short of the more structural considerations and focus on just a few examples with some importance to semiconductor devices.

Properties of Interfaces

- First, we restrict ourselves to properties that are directly linked to the interface; i.e. that stem from whatever the interface introduces *locally*. Interface states in the band gap are an interface property in this sense, but not the space charge region of a **pn**-junction, or the junction properties themselves.
  - In principle, all these properties are then given by the exact structure of the interface, i.e. the exact location of all atoms and their interaction with their neighbors, i.e. the bonding situation.
  - We may distinguish between intrinsic and extrinsic properties. *Intrinsic* properties result from an interface between two "ideal" materials that do not contain unwanted impurities - in other words we look at a "perfect" interface. Note that perfect interfaces may still contain lattice defects, e.g. so-called grain boundary dislocations, as an intrinsic part of their structure.
  - *Extrinsic* properties result from the incorporation of impurity atoms, precipitates and other structural defects not present in the intrinsic case.
- Looking at an (incomplete) list of interface properties, what comes to mind are the following topics:
  - **Interface energy**. A few numbers give an idea of the magnitudes encountered:
    - Coherent twin boundary in **Si** (smallest interface energy for **Si** grain boundaries): ≈ **60 mJ/m^2**
    - General grain boundary in Si: ≈ **600 mJ/m^2**
    - Silicon - **SiO$_2$** interface ???

- **Interface "Strength"**: How much energy (or mechanical stress) is needed to rip the two materials apart? Looking at the many interfaces of a typical integrated circuit this does not appear to be a problem - the whole thing sticks together pretty well, However, this is more a lucky accident. If you try to to replace the intermetal dielectrics (usually some variant of **SiO_2**) with some kind of **CF_4** (Teflon in other words), you will have big problems because it does not "stick". And you would dearly love to use Teflon for modern high speed devices because it is one of the few eligible materials with a very low dielectric constant. Of course, interface strength is related to the interface energy.
- **Interface structure**: There are several layers of complexity when considering the interface structure:

  - First, you may have to ask yourself if you really have a "simple" **A-B**-interface. On a more macroscopic (but intrinsic) level you may wonder if there is an intermediate layer - e.g. a thin amorphous layer of some **Si - N**i mixture between a **Si - $NiSi_2$** interface, to give an example (there isn't, by the way), or if there is an abrupt crystalline - amorphous transition at the **Si - SiO_2** interface or gradual loss of crystallinity in the **SiO_2** interface (it's rather abrupt), or if there are any preferred crystallographic orientations for interfaces between crystals (there usually are), and so on
  - For crystalline - crystalline interfaces you may ask next, if there are any intrinsic structural units, e.g. networks of dislocations (e.g. grain boundary dislocations or misfit dislocations).
  - Yet a level down, the question will be how an interface deals with intrinsic lattice defects, e.g. dislocations impinging on it, steps in the interface plane or changes in the plane of the interface (the interface may be bent, after all)
  - Finally, you have to consider the influence of extrinsic components. Point defects might segregate at the interface and even form precipitates, or small highly mobile atoms (usually **H**) may become stuck at the interface (bonding to "**dangling** " **bonds**, i.e. bonds that are nit saturated.

  .
- **Interface "chemistry"**: Interfaces are often more reactive in a general sense of the word than the bulk material. Corrosion, in particular, may not only proceed along interfaces much more rapidly than in the bulk; it may be caused by the "galvanic" properties of the interface, Somewhat more general, diffusion of atoms or molecules might be quite different in interfaces than in the bulk.
- **Interface charge:** Interfaces might be charged for a variety of reason, and this brings us closer to the more interesting properties for semiconductor devices:

  - It is almost impossible to balance charges exactly in interfaces between ionic crystals or materials with some ionic binding component. The interface then carries a net charge that is balanced either by other parts of the interface, by charged point defects, or by other interfaces. The charge is felt for a distance of about a Debye length; in all conducting media it is therefore hardly noticeable. This is an essentially *intrinsic charge*.
  - If the interface has some energy states in a bandgap, these states may or may not be occupied by electrons, depending on the Fermi energy of the system. The interface therefore may carry some variable amount of charge. Again, this is essentially an *intrinsic charge*.
  - If some *extrinsic* ions are solidly trapped in the interface - e.g. alkali metals and ***alkaline earth*** metals like **Na**, **K**, **Ca**, etc. in the **Si - SiO_2** interface - there will be some charge in the interface too. This then is an *extrinsic charge*

  Some devices, e.g. ***EPROM***s, rely on interface charge - and on the fact that it may stay there for very long times!
- **Interfaces states** - for electrons or holes. While the density of states is always affected by interfaces, they must not necessarily introduce energy levels in the band gap; i.e. only the density of states in the valence and conduction band may be affected.

Reactive and Non-Reactive Interfaces

Whenever we make an interface between materials **A** and **B** by depositing **B** somehow on **A**, the formation of the final structure with all its properties will occur between two extreme cases

- For a *non-reacting interface*, the **B**-atoms just stay pretty much wherever the are deposited.
- For a *reactive interface*, interdiffusion takes place, forming some **A-B** mixture or even defined compounds **A_x B_y**.

Non-reactive interfaces usually become reactive to some extent upon heating.

Topics to be covered

Si-SiO_2 Interface, Si - metal, silicides

3.3.2 Scaling Laws

General Consideration of Making Devices Smaller

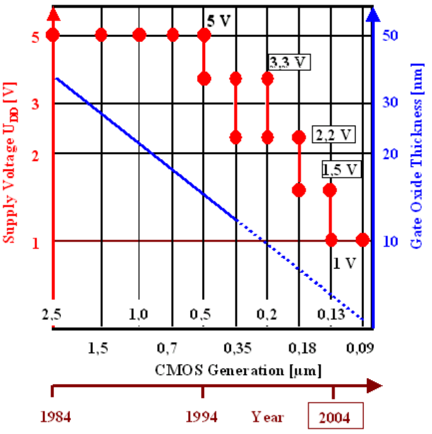
- All theories introduced so far (i.e. all of chapter 2), always assumed "infinite" or "semi-infinite" crystals. For example, the size of the crystals did not matter for the characteristics of a **pn**-junction; the dimensions of the **n**- and **p**-doped regions did not enter the equations.
  - However, if we reconsider for a moment the simple derivation of the **I-U**-characteristics of a **pn**-junction, we (hopefully) remember that the carriers responsible for the reverse current originated from a region defined by the diffusion length of the minority carriers.
  - What happens if the device is much smaller than the diffusion length ***L*** ? This will be almost always the case, considering that ***L*** is around **100 µm** and typical integrated transistor occupy hardly **1 µm^2**?
- While this question can still be answered relatively easily by conventional device physics, it nicely illustrates that we can not expect the properties of *any* device to be independent of its size as suggested by simple semiconductor physics.
  - The basic question in micro-technology thus is: What happens if you make an existing (and functioning) device smaller?
  - And making something smaller can be done in different ways: You may simply decrease the lateral extensions while leaving the depth dimension unchanged, or more realistically, you scale the lateral and depth dimensions by different factors. As an example, while you may reduce the lateral size of a source-drain region by a factor of **2**, the depth of the **pn**-junctions and the thickness of the gate oxide may scale with only a factor of **1,3**.
  - How do you find the optimum? Where are limits and how can they be overcome? In other words: What are the relevant scaling laws and when do we hit a brickwall **-** you can't make it smaller any more without insurmountable problems!
- There are some billion Dollar questions hidden in this scenario, and there are no easy answers for some of the details. There are also, however, some simple laws and rules which we will consider briefly in this subchapter.

Linear Scaling and Problems

- Lets look at some general **scaling laws**. We simply assume that we decrease all linear dimensions of an existing device by the **scaling factor *K***.
  - A first obvious conclusion is that the field strength in some insulating layer, e.g. a gate oxide, increases ***K*** -fold. We may accept that, or we might scale the voltage, too. In this case we would decrease U_{DD}, the external driving voltage, to U_{DD}/K.
  - Going through all important parameters (with some approximations if necessary), we obtain the following table

Property	**Scaling**	
All lateral and vertical dimensions	**$1/K$**	
Doping concentration	**K**	
	$U_{DD} \Rightarrow U_{DD}/K$	**U_{DD} = constant**
Packing density (No. transistor/cm^2)	**K^2**	**K^2**
Current densities	**K**	**K^3**
Field strenghts	**1**	**K**
Power loss density	**1**	**K^3**
Power loss per transistor	**$1/K^2$**	**K**
Time delay per transistor	**$1/K$**	**$1/K^2$**

- The problems you are running into are obvious. Lets look at the transition from a **1 µm** process to a **0,25 µm** process, i.e. ***K* = 4**
  - Without a fourfold reduction in the supply voltage, we would have a **64** fold increase in current densities and power loss, and a **4** fold increase in field strength. This is *not* going to work
  - Decreasing $\boldsymbol{U}_{\mathbf{DD}}$ fourfold (from **5 V** to **1,25 V**) still increases the current density fourfold, but keeps all other parameters manageable. However, our device only speeds up **4** fold, compared to **16** fold for constant. $\boldsymbol{U}_{\mathbf{DD}}$.
- So simply lets scale down $\boldsymbol{U}_{\mathbf{DD}}$ some more? Well, yes **-** but: You can't just decrease all dimensions and $\boldsymbol{U}_{\mathbf{DD}}$ just so!
  - The tjickness of gate and capacitor dielectrics might be close to absolute limits (e.g. imposed by tunneling) and simply cannot be scaled down much more.
  - Internal voltages might only be fractions of the supply voltage and reducing $\boldsymbol{U}_{\mathbf{DD}}$ may decrease signal to noise levels to unacceptable values.
  - Voltage swings for switching transistor must at least be in the order of the band gap, i.e. **1 V** for Si. Voltages thus cannot be reduced to arbitrarily small levels.
- If we look at the actual scaling of devices, much ingenuity and many additional process steps were used to avoid the simple rules of scaling. Here are a few examples:
  - Trench capacitor instead of planar capacitor. The thickness of the dielectric thus could stay relatively constant, which allowed higher supply voltages than required by scaling.
  - Electromigration resistant metallization (addition of **Cu** or other atoms to the **Al** lines, multi layers etc.) allowed larger current densities
  - "Lightly doped drain", i.e. complicated dopant profiles of the source/drain region below the gate allowed higher field strengths in the **MOS** transistors.
- The list is easily extended by **10** or more points; but you get the drift.
  - But all tricks notwithstanding: the supply voltage had to come down. The historical development is shown in the figure below.



- There are several remarkable features:
  - For almost **10** years the supply voltage was kept constant at $\boldsymbol{U}_{\mathbf{DD}}$ **= 5 V** despite a scaling of ***K* = 5**. This was possible by a dramatic increase of process complexity and materials engineering.
  - The end in scaling $\boldsymbol{U}_{\mathbf{DD}}$ is near. Now (**2001**), supply voltages are as low as **1, 5 V**, and we simply cannot go much below **1 V** with **Si**.
  - New principles are needed, because we still can make functioning transistors far below **0,1 µm**. One possible solution is to make vertical transistors. Demonstrators (Bell Labs) work with gate length of **30 nm** or less.
- What happens: You will not only live to see it, but possibly help to establish it.

Fundamental Limits

- Whatever clever tricks are used to make just one more step towards smaller devices, there are some ultimate limits. One simple example shall be given; it concerns doping:
  - Lets say we need doped areas with a typical dopant concentrations of **$\rho = 10^{16}$ cm^{-3}** or **10^{18} cm^{-3}** and maximum deviations of **± 10%**. This implies that the doped area must contain at least **100** dopant atoms because the statistical fluctuations are then **$(100)^{1/2} = 10$** giving the **10%** allowed.
  - **100** atoms at the prescribed density need a volume of (**$100/10^{16}$) cm^3** or **($100/10^{18}$) cm^3** , respectively, which equals **10^7 nm^3** or **10^5 nm^3**, respectively.
  - These volumes correspond to cubes with a linear dimension of **215 nm** or **46,4 nm**, respectively. What does this mean?
- Look at it in a different way: A typical source/drain region in a modern integrated circuit may be **0,5 µm x 0,5µm x 0,2 µm = 5 · 10^{-2} µm^3 = 5 · 10^7 nm^3**.
  - At a doping level of **10^{16} cm^3** - corresponding to the perfectly reasonable resistivity of **1.4 Ωcm** (**p**-type) or **0.5 Ωcm** (**n**-type) - we have about **500** doping atoms in there! Doping to a precision better than **$(500)^{1/2}$ = 22,4 or 4,47 %** is principially not possible assuming a statistical distribution of the doping atoms.
  - Decreasing the size to e.g. **0,1 µm x 0,1µm x 0,02 µm = 2 · 10^5 nm^3** leaves us **2** doping atoms in there - obviously absurd. Again we have to turn to novel structure - vertical transistors may do the trick once more.
- There are more "fundamental limits" - just how fundamental they are, is a matter of present day research (and a seminar topic).

3.4 Basic Silicon Devices

In this subchapter 3.4 we look at typical **Si** devices from two angles:

- *First*, the very basics of more or less *ideal* devices are given in rather short notation. This should be a reminder and not something new for you.
- *Second*, some relevant facts about *real* devices are presented without much deduction.

Some of this will be needed and come up again in chapter 8 "Speed".

3.4.1 Junction Diodes

The Ideal Junction Diode

The ideal - in the sense of most simple - **Si** junction diode has essentially one major property:

- Its ***I-U*** *characteristic* can be described to a very good approximations by the "simple" **pn-** junction theory containing the contribution of the space charge layer (otherwise you are not really describing a Si device at all). We had

$$\boxed{j \;=\; \left(\frac{e\cdot L\cdot n_i^{\,2}}{\tau\;\cdot N_A} + \frac{e\cdot L\cdot n_i^{\,2}}{\tau\;\cdot N_D}\right)\cdot\left(\exp - \frac{eU}{kT} \;-1\right) \;+\; \frac{e\cdot n_i\cdot d(U)}{\tau}\left(\exp - \frac{eU}{2kT} \;-1\right)}$$

- The major variables are the *doping levels* ***N*** (determining the **SCR** width ***d***,too), the *diffusion length* ***L***, (same thing as the life time **τ**, since they are coupled by the **Si** diffusion constant ***D***, which again is directly connected to the mobility **µ** of the carriers, which finally is mainly a function of doping), the temperature ***T***, and, of course, the junction voltage ***U***.

When does this equation break down, i.e. what distinguishes an "*ideal* " junction diode from a "*real* " one?

First, with just that equation, you could increase the *voltage* to any value you like, and the equation gives some *current* which might become very large for forward bias, and would stay small for arbitrarily large reverse bias. That is *not* realistic, of course.

- At large reverse voltages, we have a large electric field in the **SCR**, and at some point we will just have **electrical breakdown** since no material can withstand arbitrarily large field strengths. The breakdown mechanism is usually **avalanche breakdown**.
- Very large forward currents are also not realistic. The real ***I-U*** characteristics shown before indicates some reasons. One thing that goes wrong is that our equation does not treat the case of **high injection**, meaning that the concentration of minority carriers injected into the junction is larger (or at least comparable) to the equilibrium concentration in the bulk. Somewhere in the derivation of the basic diode equation we always made an assumption (hidden or openly) of *low injection*, so we cannot expect real diodes to behave ideally for *large* forward currents.

Second, we have totally neglected the *ohmic resistance* of the **Si**.

- Whatever its value R_{ser} might be, it can be seen as being switched in series to the actual diode and thus will reduce the junction voltage U_{junct} to

$$\boxed{U_{junct} \;=\; U_{ex} \;-\; I\cdot R_{ser}}$$

- In other words, we now must distinguish between the external or terminal voltage U_{ex} and the junction voltage U_{junct}, and there simply is no way to pass currents larger than U_{ex}/R_{ser}.

Third, we certainly must have some reservations about the *doping* in the derivation of the equations, too.

- The *concentration* ***N*** certainly cannot have any value. But limits here are not very important, because for the level of doping achievable in real Si diode, the equation is not too bad.
- More important are *gradients* in the dopant concentration, i.e. $\mathrm{d}N_{Acc}/\mathrm{d}\,x$ because we assumed (implicitly) that ***N*** is constant; which it rarely is in real diodes.

Fourth, we have to be a bit concerned about the *temperature*.

- The validity of the equation with respect to ***T***-variations is limited: Somewhere we assumed that all dopants are ionized and that the Fermi energy is close to the band edges which will certainly not be true at any temperature..
- In practical terms this means that we are restricted to temperatures not too far off room temperature.

Fifth and last, we have to consider the *diffusion length* ***L***.

- While we might worry a bit about the allowable range - is the equation still correct for very large or very small ***L*** - the real problem is different:
- Make an ideal diode from **Si** with ***L* = 200 µm**, for example (a regular value), and then make the **diode small** - lets say you just leave ***1 µm of Si*** to the left and right of the **SCR**. Since ***L*** was the average distance an electron or hole traveled in the **Si** before death by recombination, we have a problem now. The bulk value of ***L*** obviously can no longer summarily describe the perambulation of a minority carrier.
- Looking at it more quantitatively, we must modify the distribution of minority carriers from the edge of the **SCR** into the bulk of the **Si** for forward current flow as it was dealt with in subchapter 2.3.4 "Useful Relations" and in subchapter 2.3.5 "Junction Reconsidered". Lets look at this in an advanced module, here we only look at the results.

The Real Junction Diode.

- Lets see summarily what we must change to account for the items 1 - 5 above.
- *First* we look at intrinsic *voltage and current limitations* .
  - Avalanche *breakdown* will occur whenever the field strength in the **SCR** manages to impart enough energy to an electron or hole to generate more carriers in some scattering process. While it is clear that the he field strength in the **SCR** is mainly a function of doping, it is not so easy to derive numbers.
  - There are more breakdown mechanisms than just carrier multiplication by avalanche effects; most important, perhaps is tunneling of carriers through the potential barrier at the junction. Again, high field strengths help.
  - Important are the practical limitations in terms of usable reverse *voltages* (not field strengths per se). The range of admissible reverse voltages is large and reaches from **> 1000 V** for lightly doped **Si**, say $\mathbf{10^{14}\ cm^{-3}}$ (and some sophisticated technology) to just a few Volts on the highly doped end - take $\mathbf{10^{18}\ cm^{-3}}$.
  - Forward currents in the *high injection mode* of a real diode will be smaller than predicted by the ideal equation. In a first approximation, we simply have to reduce the slope of the characteristic by a factor of **2** - we have the same slope as in the **SCR** dominated part at very small forward currents.
  - This is what is shown in the curve for a real diode in the picture we used before.
- *Second*, how about the *ohmic resistance*?
  - It is certainly not negligible in many real diodes and is one of the major problems in solar cells.
  - It is, however, easy to address. Just do it yourself in a little exercise.

Exercise 3.4.1

Current-Voltage characteristics of a solar cell with series and shunt resistance

- *Third*, we consider *doping gradients*.
  - This is certainly the realistic case, because real diodes are mostly made by diffusing **n** or **p**-dopant into a **p** or **n**-doped substrate, respectively. At least one side of the diode thus has a doping that varies strongly with the distance from the actual junction (located at the point where $n^e = n^p$ or $N_{Don} = N_{DAcc}$. Typical profiles are given in the link
  - How do doping gradients influence the current-voltage characteristics?
- The surprising answer is: *Not much at all!* (Take that with a grain of salt)
  - The reason is that no matter how you derive the $I(U)$ characteristics, the decisive parts are only the height of energy barriers, and the recombination/generation/diffusion behavior outside the space charge region. The precise shape of the band bending, or the width of the **SCR** does not enter at all, or at best weakly (in the **SCR** term via d_{SCR}) in the basic equation from above.
- What will be influenced by doping gradients are: *First*, (minor) parameters like resistivity and mobility, and *second*, the **SCR** properties like its size, and especially its **capacitance**.
  - The first group changes the pre-exponential factor $L \cdot n_i{}^2/\tau \cdot N_{dop}$ somewhat; essentially you replace the formerly constant N_{dop} by some kind of average resulting in an effective doping N_{eff}.
  - These second set of parameters resulted from solving the Poisson equation, and we have only done this for constant dopant concentration. Redoing the calculations for *real* dopant profiles must generally be done numerically.
- However, the minor effects of doping gradients on the ***DC*** (direct current) current-voltage behavior must not induce you to think that doping gradients are unimportant!
  - The ***AC*** behavior, or, in other words, the *speed* of the junction, is very much influenced by **SCR** properties and thus by dopant gradients.

  - More to that in chapter 8 "Speed".

- *Fourth*, a quick glance at temperature effects
  - Typical ***T***-specifications for **Si** devices are **0 ^{o}C < *T* < 70 ^{o}C** for typical consumer integrated circuits or **– 55 ^{o}C < *T* < + 125 ^{o}C** for somewhat better stuff.
  - Pushing technology and materials gives maybe **+ 160 ^{o}C** for an admissible operation temperature of **Si** devices.

- While it is not only the **pn**-junction that limits the temperature region for applications of more complex devices, you simply must make sure that you have sufficient carriers (i.e. ***T*** cannot be too low), but not too many (i.e. carrier concentration must be controlled by doping and not by thermal band-band generation), limiting the upper temperature.

- *Fifth and last*, how does the size of the device influence its properties?
  - There is simple answer for simple (one-dimensional) small diodes: Replace the diffusion length ***L*** by a relevant length of the device, e.g. the distance between the edge of the **SCR** to the ohmic contact d_{Con} in all equations, and concomitantly the life time τ by the transit time t_{tran}defined via

$$\boxed{d_{\text{Con}} \;=\; \left(D \cdot t_{\text{tran}}\right)^{1/2}}$$

  - The justification is given in an advanced module.
  - In other words, we equate some relevant length d_{Con} of the device with the average distance that minority carriers travel before they disappear, and t_{tran} is the time they move around.

- This makes not only immediate sense but has far-reaching consequences, as we will see, e.g. in chapter 8**.** Some major points are listed below:
  - The size (together with the mobility) becomes the most important parameter for *speed* .
  - Since vertical dimensions are more easily made small than lateral ones, bipolar devices in a vertical stack are inherently faster than lateral **MOS** devices. In-diffusion of dopants, e.g., defining the depth of a **pn**-junction, is easily restricted to **0,1 µm**; while it takes very advanced technology to produce lateral structure sizes in this region.
  - Leakage currents decrease with decreasing device size.

- One (of several) incentives to make devices ever smaller has its roots right here.

- Well, the long and short of this is that *real* diodes are quite different from *ideal* ones - in the details! The global topics stay unchanged, lets recount them quickly:
  - - Majority and minority carrier dynamic equilibrium in the bulk, controlled by doping, carrier life time and mobility
    - Energy barrier at the junction, resulting in **SCR** and carrier concentration gradients
    - Very different behavior in reverse and forward direction
    - Forward currents mostly resulting from diffusion currents removing injected minorities
    - Reverse currents mostly resulting from field currents affecting minorities at the edge of the **SCR**

- In practice, "ideal large" diodes practically do not exist (except in the form of solar cells). Even "small" diodes with graded junctions and the like are not really used if you need a diode (but as part of more complicated devices like **MOS** transistors). Technical diodes are more sophisticated since they are optimized for specific parameters, e.g. extremely large breakdown voltages. A few examples are
  - The **PIN diode**, short for: **p**-doped **- i**ntrinsic **- n**-doped. A thin layer, as intrinsic as possible, is sandwiched between doped **Si**. Good for large forward currents and large reverse voltages. This is the standard form for diodes use for rectifying purposes.
  - **Tunnel diodes**, **varactors** , fast recovery diodes, **Gunn** diodes, ***IMPATT*** diodes, **Zener** diodes, **solar cells** - there is no shortage of names for special diodes and applications going with it. We will, however, not dwell on the subject her (in time, maybe, there might be advanced modules).

3.4.2 Bipolar Transistors

Basic Concept and Operation

- We are not particularly interested in **bipolar transistors** and therefore will treat them only cursory.
  - Essentially, we have two junctions diodes switched in series (sharing one doped piece of **Si**), i.e. a **npn** or a **pnp** configuration, with the *added condition* that the middle piece (the **base**) is *very thin*. "Very thin" means that the base width $\boldsymbol{d}_{\mathbf{base}}$ is much smaller than the diffusion length $\boldsymbol{L}$.
- The other two doped regions are called the **emitter** and the **collector**.
  - For transistor operation, we switch the emitter - base (**EB)** diode in forward direction, and the base - collector (**BC**) diode in reverse direction as shown below.
  - This will give us a large forward current and a small reverse current - which we will simply neglect at present - in the **EB** diode, exactly as described for diodes . What happens in the **BC** diode is more complicated and constitutes the principle of the transistor.
  - In other words, in a **pnp** transistor, we are injecting a lot of holes into the base from the emitter side and a lot of electrons into the emitter from the base side; and vice versa in a **npn**- transistor. Lets look at the two **EB** current components more closely transistor:
- For the *hole* forward current, we have in the simplest approximation (ideal diode, no reverse current; no **SCR** contribution):

$$\boxed{j_{\text{hole}}(U) \;=\; \frac{e \cdot L \cdot n_i^2}{\tau \cdot N_{\text{Acc}}} \cdot \exp - \frac{e \cdot U}{kT}}$$

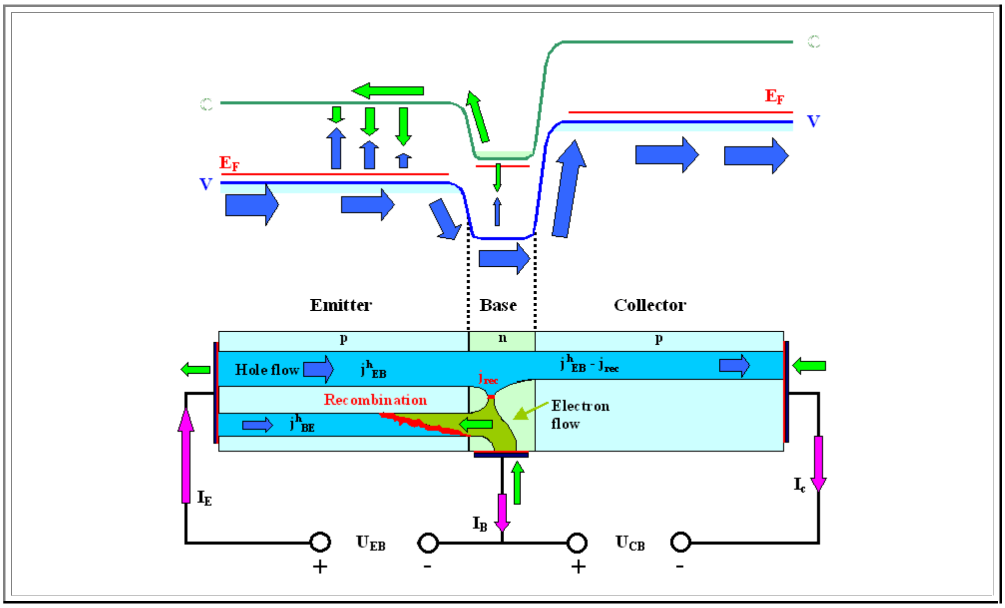
  - and the relevant quantities refer to the *hole* properties in the ***n** - doped base* and the doping level $\boldsymbol{N}_{\mathbf{Acc}}$ in the ***p** - doped emitter*. For the electron forward current we have accordingly:

$$\boxed{j_{\text{electron}}(U) \;=\; \frac{e \cdot L \cdot n_i^2}{\tau \cdot N_{\text{Don}}} \cdot \exp - \frac{e \cdot U}{kT}}$$

  - and the relevant quantities refer to the *electron* properties in the ***p** - doped emitter* and the doping level $\boldsymbol{N}_{\mathbf{Don}}$ in the ***n** - doped base*.
  - The relation between these currents, i.e. $j_{\text{hole}}/j_{\text{electron}}$, which we call the **injection ratio** κ, then is given by

$$\boxed{\kappa \;=\; \frac{\dfrac{L_h}{\tau_h \cdot N_{Ac}}}{\dfrac{L_e}{\tau_e \cdot N_{Don}}} \;=\; \frac{N_{Ac}}{N_{Don}}}$$

  - Always assuming that electrons and holes have identical lifetimes and diffusion lengths.
- The *injection ratio* κ is a prime quantity. We will encounter it again when we discuss for optoelectronic devices!
- For only one diode, that would be all. But we have a second diode right after the first one. The holes injected into the base from the emitter, will diffuse around in the base and long before the die a natural death by recombination, they will have reached the other side of the base
  - There they encounter the electrical field of the base-collector **SCR** which will sweep them rapidly towards the collector region where they become majority carriers. In other words, we have a large hole component in the reverse current of the **BC** diode (and the normal small electron component which we neglect).
  - A band diagram and the flow of carriers is shown schematically below in a band diagram and a current and carrier flow diagram.



- Lets discuss the various currents going from left to right.
  - At the *emitter contact*, we have two hole currents, $j_{EB}{}^h$ and $j_{BE}{}^h$ that are converted to electron currents that carry a negative charge away form the emitter. The technical current (mauve arrows) flows in the opposite direction by convention.
- For the *base current* two major components are important:
  - **1.** An electron current $j_B{}^e$, directly taken from the *base contact*, most of which is injected into the emitter. The electrons are minority carriers there and recombine within a distance L with holes, causing the small hole current component shown at the emitter contact.
  - **2.** An internal recombination current j_{rec} caused by the few holes injected into the base from the emitter that recombine in the base region with electrons, and which reduces $j_B{}^e$ somewhat. This gives us

$$j_{BE}{}^h \;=\; j_B{}^e \;-\; j_{rec}$$

  - Since all holes would recombine within L, we may approximate the fraction recombining in the base by

$$j_{rec} \;=\; j_{EB}{}^h \cdot \frac{d_{base}}{L}$$

- Last, the current at the *collector contact* is the *hole* current $j_{EB}{}^h - j_{rec}$ which will be converted into an electron current at the contact.
- The external terminal *currents* I_E,I_B, and I_C thus are related by the simple equation

$$I_E \;=\; I_B \;+\; I_C$$

- A bipolar transistor, as we know, is a *current amplifier*. In black box terms this means that a small current at the the *input* causes a large current at the *output* .
  - The input current is I_B , the output current I_C. This gives us a current amplification factor γ of

$$\gamma = \frac{I_C}{I_B} = \frac{I_E}{I_B} - 1$$

  - Lets neglect the small recombination current in the base for a minute. The emitter current (density) then is simply the total current through a **pn**-junction, i.e. in the terminology from the picture $j_E = j_{BE}{}^h + j_B{}^e$, while the base current is just the electron component $j_B{}^e$.
  - This gives us for I_E/I_B and finally for γ:

$$\frac{I_E}{I_B} = \frac{j_{BE}^{h} + j_B^{e}}{j_B^{e}} = \kappa + 1$$

$$\gamma = \frac{I_E}{I_B} - 1 = \kappa + 1 - 1 = \kappa = \frac{N_{Ac}}{N_{Don}}$$

- *Now this is really easy*! We will obtain a large current amplification (easily **100** or more), if we use a lightly doped base and a heavily doped emitter. And since we can use large base - collector voltages, we can get heavy power amplification, too.
  - Making better approximations is not difficult either. Allowing somewhat different properties of electrons and holes and a finite recombination current in the base, we get

$$\gamma = \frac{\dfrac{L_h}{\tau_h \cdot N_{Ac}}}{\dfrac{L_e}{\tau_e \cdot N_{Don}}} \cdot \left(1 - \frac{d_{base}}{L}\right) \approx \frac{N_{Don}}{N_{DAc}} \cdot \left(1 - \frac{d_{base}}{L}\right)$$

  - The approximation again is for identical life times and diffusion lengths.
- Obviously, you want to make the base width d_{base} small, *and* keep L large.

Real Bipolar Transistors

- Real bipolar transistors, especially the very small ones in integrated circuits, are complicated affairs; for a quick glance on how they are made and what the **pnp** or or **npn** part looks like, use the link.
- Otherwise, everything mentioned in the context of real diodes applies to bipolar transistors just as well. And there are, of course, some special topics, too.
  - But we will *not* discuss this any further, except to point out that the "small device" topic introduced for a simple p-n-junction now becomes a new quality:
  - Besides the length of the emitter and collector part which are influencing currents in the way discussed, we now have the **width of the base region d_{base}** which introduces a new quality with respect to device dimensions and device performance.
  - The numerical value of d_{base} (or better, the relation d_{base}/L), does not just change the device properties somewhat, but is the *crucial* parameter that brings the device into existence. A transistor with a base width of several **100 µm** simply is not a transistor, neither are two individual diodes soldered together.
- The immediate and unavoidable consequence is that at this point of making semiconductor devices, *we have to make things real small*.
  - Microtechnology - typical lengths around or below **1 µm** (at least in one dimension) - is mandatory. There are no big transistors in more than two dimensions.
  - Understanding *microscopic* properties of materials (demanding quantum theory, statistical thermodynamics, and so on) becomes mandatory. *Materials Science and Engineering was born*.

3.4.3 MOS Transistors

If you do not know the basic structure of a MOS transistor, you have a problem. Use the links and and make sure you understand the contents.

- Basic MOS transistor
- Making integrated MOS transistors.

4. Silicon: Special Properties and Emerging Technologies

4. Silicon: Special Properties and Emerging Technologies

4.1 Silicon on Insulator

4.1.1 General Remarks

Intended Topics:

4.1 Silicon on Insulator

- *advantages and problems, basic device structure*
- Modern developments
  Oxygen Implantation; waferbonding, smart cut technology

4.2 Etching of Silicon

- Chemical etches
  Isotropic and anisotropic dissolution, defect etches and anodic etching
- Micromechanics and microsystem technology
  Basic considerations, special process steps
- Electrochemical etching, Porous Silicon and applications
  Photonic crystals, filters, sensors, microtechnology, integrated wave guides, ...

4.3 Specialities

- Amorphes Si and applications
  Structural and electronic properties, H - passivation, solar cells and FPDs
- SiGe: Materials ascpects and devices
  HEMT, detectors (incl. Ge),

4.1.2 Modern Developments

Intended Topics

4.1 Silicon on Insulator

- General
  advantages and problems, basic device structure
- Modern developments
  Oxygen Implantation; waferbonding, smart cut technology

4.2 Etching of Silicon

- Chemical etches
  Isotropic and anisotropic dissolution, defect etches and anodic etching
- Micromechanics and microsystem technology
  Basic considerations, special process steps
- Electrochemical etching, Porous Silicon and applications
  Photonic crystals, filters, sensors, microtechnology, integrated wave guides, ...

4.3 Specialities

- Amorphes Si and applications
  Structural and electronic properties, H - passivation, solar cells and FPDs
- SiGe: Materials ascpects and devices
  HEMT, detectors (incl. Ge),

4.2 Etching of Silicon

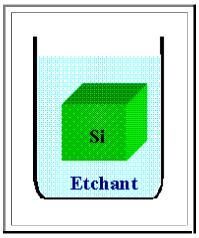
4.2.1 General Remarks

What is Etching?

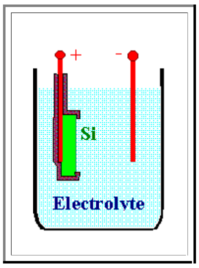
- **Etching** of silicon can mean several things:
  - The **chemical** dissolution of **Si**. In other words, **Si** will dissolve upon simple immersion in certain chemicals.
  - The **electrochemical** dissolution of **Si**. In this case dissolution takes place in certain chemicals called electrolytes - different from the ones for purly chemical etching - if, and *only if*, electrical current is passed through the **Si** - electrolyte junction.
  - The dissolution of Si in a **Plasma**. Dissolution takes place if **Si** is exposed to a suitable Plasma, i.e. a low density "vapor" of electronically excited ionized atoms or molecules and free electrons which often have considerable kinetic energies in addition. Usually plasma etching takes place at low concentrations, i.e. at low pressures. Plasma etching may use chemicals (usually gases) different from those used for chemical or electrochemical etching.
- Etching is not only a key process for any **Si** technology, it is rather poorly understood. In particular, the electrochemical and plasma etching of **Si** is full of surprises, empirical receipes, and counts among the "black arts" compared e.g., to rather well understood if complex processes like ion implantation or diffusion.
  - In typical microelectronic products, chemical etching has been abandoned in favor of plasma etching around **1985**; electrochemical etching so far has never been used.
  - However, for **microsystems** or "***MEMS*** " (= micro electronic and mechanical systems) applications, *chemical* etching is the crucial process par excellence, we will cover that briefly in an own subchapter.
  - Electrochemical etching, though known since the fifties, was seriously investigated only since about **1990**. It produces a (still growing) wealth of new phenomena and has a large potential for new products which is investigated by many research groups. The first large scale product based on **Si** electrochemistry was introduced in **1999**; it is mentioned int he contect of the "Silicon on Insulator" subchapter
  - There are special big conferences dedicated to etching of **Si**, especially to plasma etching and electrochemical etching.
  - We will neithrt treat plasma etching in this script, nor electrochemical etching as an emerging new technique.

Basic Experiments

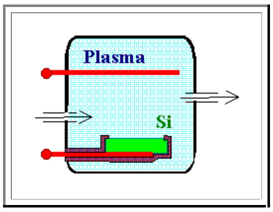
- The simple drawing below illustrates the basic etching experiment for the three etching modes mentioned.



- *Chemical etching*
- Silicon, here in the form of a cube - e.g. a single crystal with **{100}** surfaces - is immersed in an etchant and will dissolve.



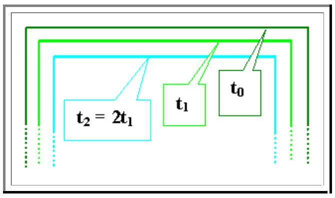
- *Electrochemical etching*
- Some parts of the **Si** sample are exposed to the electrolyte. The back side contact and other parts are protected. A voltage between the back side contact and a counterelectrode of some inert material (usually **Pt**) allows to apply a voltage which may cause a current to flow across the interface **Si** - electrolyte.
- Many chemical reactions may occur, some will dissolve the **Si**.



- *Plasma etching*
- In a vacuum vessel, some gas is introduced and excited to a Plasma by, e.g., applying a high-voltage/high-frequency power source between the Si and sone other electrode.
- **Si** may be etched, the products of the reactions and the unused gas are moved out of the system.

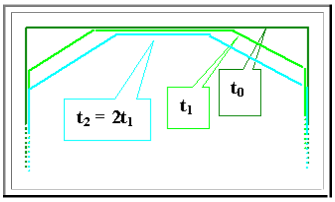
- There are many (much more complicated) ways of producing a plasma and having it interacct with the **Si** surface.

- If any etching occurs, it may happen in several, very distinctive ways. They can be discussed most easily if we imagine that the **Si** sample is a single crystal with, e.g., **{100}** surfaces as indicated the first picture.
  - The **{100}** cube may dissolve homogeneously. This looks something like this



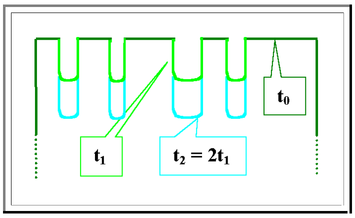
  - The cube dissolves homogeneously, i.e. the dissolution rate is constant and independent on any crystallographic or other details. The dissolution is **isotropic** .

  - The **{100}** cube may dissolve in a way that changes its shape. This happens whenever the dissolution of some crystallographic planes is faster that some other ones, or, generally speaking, if the dissolution rate depends on the crystallographic orientation.



  - Certain crystallographic planes dissolve much faster than others, the dissolution is **anisotropic**.

  - The dissolution may proceed by forming a pattern of its own e.g. by etching pores in the crystal. This self-induced patter formation process may have some crystal anisotropy in addition, e.g. the pores are always oriented in certain crystallographic directions. This not only may get extremely complicated in theory, it actually does happens in the (electro)chemical etching of **Si** and it realy is excruciably complicated!



  - Patterns emerge; in the example small pores grow into the Si. There are many other patterns obtainable; especially with electrochemical etching.
  - While the places where pores start to grow may be totally random, the growth direction of the pores may be in a special crystallographic direction - some crystal anisotropy comes in on a secondary level.

  - Other patterns may be tied to defects in the crystal, i.e. a little pore or pit develops wherever a defect, e.g. a dislocation line intersects the surface - we have defect etching. Again, this may work (very) differently on different crystallographic surfaces

- By now it becomes clear that etching is very complex and has many facets to it. The situation is becoming even more complex if we consider etching through a mask, i.e. only defined parts of the **Si** surface are exposed to the etching agent.
  - This is of course the standard procedure encountered in technology. All the observations made above apply and some new points may come in on their own.
  - However, we will not treat this case here, but in the context of the upcoming chapters and subchapters.

- Etching of **Si** always can be done in two principially unrelated ways (which often occur in a mixture, causing some of the complication):
  - **Direct dissolution**, cursorily expressed by the "chemical" formula
    Si + something = Si-compound(dissolved) + something else.
    A simple example (and reality is actually not as simple as that) could be

$$\mathbf{Si + 2HF \Rightarrow SiF_2 + H_2}$$

  - **Oxidation** in a first step, **followed by oxide dissolution** in a second step, e.g.

$$\mathbf{HF + HNO_3 \Rightarrow SiO_2 + ...}$$

$$\mathbf{SiO_2 + 2HF \Rightarrow SiF_2 + ...}$$

  - These simple examples make clear that even the basic chemistry is not so simple; we will come to that later.

- One last general remark:
  - The truncated chemical equations above use hydrofluoric acid (**HF**), and nitric acid (**HNO_3**). This is not accidential, but what you have to use in many cases. Some other chemicals may be useful, too, but there is an important general rule:

Etching Silicon always uses some of the most dangerous chemicals known to mankind!
Beware and be sure you know what you are doing!

- Hydrofluorous acid is nothing to joke about! The same is true for most of the other chemicals involved.
  - However, taking proper precautions, experiments are possible and routinely done. Huge quantities of the dangerous chemicals are used up every day in semiconductor technology and there have been very few serious accidents.
  - Do not be afraid of using chemical etches! Make sure you know the safety measures and have somebody with experience initiate you. Most accidents happen because people are careless or afraid!

4.2.2 Chemical Etching of Silicon

Isotropic Etching of Silicon

Module is unfinished

- If we neglect exotic mixtures of chemicals, the universal isotropic **Si** etchant is a mixture of **HNO_3 + HF + CH_3COOH** .
    - In other words: Mix nitric acid, hydrofluoric acid and acidic acid (**= HAc**)! The resultant mixture is known as ***CP*x**; the **x** being **1 ...4**, depending on the exact composition - and this abbreviation was chosen for good reasons!
    - The etching characteristics you obtain depend very much on the exact composition and the temperature, possibly on the doping of the **Si**, and to some extent on some small amounts of other chemicals that might be added to the mixture.
    - Some data for two typical mixtures are presented below.

Composition HF : HNO3 : HAc	Temp. °C	Etch rate µm/min	Anisotropy <100> : <111>	Masks for selective etching	Remarks
1 : 3 : 8	20	3	1 : 1	none	Etch rate strongly reduced for doping n: $<3 \cdot 10^{17}$, p: $<3 \cdot 10^{15}$
1 : 2 : 1	20	40	1 : 1	none	

- **CP** etches attack about everything - don't store them in a glass bottle, it will dissolve too!
    - Only some polymers, most notably ***PVC*** and Teflon, are **CP** prove
- They generally work by oxidizing the **Si** (thats what the is doing) and dissolving the oxide (the job of the **HF**).
    - **HAc** is mostly just for diluting the mixture
    - However, **HAc** is also....
- A big problem with the **CP** etches is that they also dissolve all possible masks - usually **SiO_2** or **Si_3N_4** layers - so they cannot be used for **selective etching**
    - At best, **Si_3N_4** may last for some time - if you hurry up selective etching becomes possible in a confined way.
- For a strong imbalance between **HNO_3** and **HF**, the etchant may change its character:
    - Most notably (and not very suprisingly), the etch rate comes way down
    - More surprisingly (but not really, if you think about it): it may now be a **defect etch**, i.e. it attacks **Si** much faster at the place of defects.
- While only **HF** dissolves **SiO_2** , all strong oxidizing agents can oxidize **Si**. It thus is possible to replace the **HNO_3** by some other oxidant. Essentially, two oxidizers are used:
- First, **H_2SO_4** can be used instead of **HNO_3**, typically in a ratio **HF : H_2SO_4 : HAc = 1 : 1: 5**
    - While inferior in over-all "quality" to the **CP** etches, it does not attack **Si_3N_4** masks very strongly and thus can be used for selective (isoptopic) etching.
    - Etch rates are around **(2...5) µm/min**, again depending on someother factors too.
- Second, **CrO_3** is used, a relatively weak oxidizer for **Si**. It only works on "soft spots", i.e. at surface areas were the bonds are weakened because of defects.
    - **Hf + CrO_3 + HAc +** many other chemicals (with no clear role) is the base of most defect etchants - a very important techique in semiconductor development

5. Fundamentals of Optoelectronics

5. Fundamentals of Optoelectronics

5.1 Materials and Radiative Recombination

5.1.1 Basic Questions and Material Issues

Basic Questions

- With "**optoelectronics**" in the context of this lecture we mean *only* electronic devices based on semiconductors where *recombination processes emit light*. We call this radiation process **spontaneous emission** of light, because it just happens statistically without any other ingredients but electrons and holes.
    - We thus do not (yet) look at the opposite process - the *absorption* of light, important in photo-diodes (or solar cells as already discussed before).
    - *Transmission* of light in waveguides (which might be integrated on a chip) is also not considered here.
- We have seen that silicon is an indirect semiconductor and recombination proceeds via deep levels. The energy released does not produce photons in an appreciable amount and **Si** is therefore not useful for optoelectronic applications.
    - This is, however, not a general truth. While recombination via *deep levels* as the required third partner does not produce light, indeed, recombination proceeding via some other third partner may. We will see that there are indirect semiconductors that emit enough light to be useful for practical applications.
    - But generally, we are looking for *direct* semiconductors where it can be expected that recombination does result in light production.
- This leads us to some general questions which can be translated to requests for material properties. Let us consider the more fundamental ones:
- The *first* question is: *What is the wavelength of the light produced?*
    - If the light is produced by direct band-to-band recombination, we have of course

$$\boxed{h \cdot \nu \; = \; E_C \; - \; E_V}$$

    - If the radiative recombination proceeds from some other energy states we simply replace $E_C - E_V$ by ΔE, the relevant energy difference.
    - With $c_{mat} = \nu \cdot \lambda$, and c_{mat} = velocity of light in the material = c_0 / n (with c_0 = velocity of light in vacuum, and n = refractive index of the material), we obtain

$$\boxed{\lambda \; = \; \frac{h \cdot c_0}{\Delta E \cdot n}}$$

- This leaves us with further material-related questions:
    - *First*, we are asking about the value of ΔE, or in most cases $E_C - E_V$, for optoelectronic materials. The answer comes from the band diagram and from finer details, still to be considered.
    - *Second*, we now must know the (possibly complex) **refractive index** of the material. This is not only important for the value of the wave length *in* the material, but especially for the transmission of light *out* of the material – where the difference in the refractive indices between two materials is a prime property of interest. The **refractive index** of a semiconductor is a new property that we did not address before.
- The *second* question coming to mind is: *How much light is produced by recombination*, or more precisely, what is the **quantum efficiency** η_{qu}, defined as the fraction of recombination events producing light?
    - For this we have to look at the various pathways open for recombination. In the preceding chapters we have discussed recombination in general and for *one* particular mechanism in detail.
    - However, there might be several mechanisms for recombination open to minority carriers, and only one might produce light. We thus must consider recombination in optoelectronic materials in more detail.
- *Third*, we may wonder about the *absolute intensity* or even more specific, *the intensity density* we can produce.
    - In other words, how many light producing recombination events per second are attainable in a *given volume* of material? What is the limit and which factors determine it?
    - Or, reformulated in technical terms: How do we produce a large non-equilibrium density of minorities (and majorities, too)? How do we inject electronically (by currents driven by external voltages) high densities of carriers in small volumes with no way out but recombination?
    - This question leads to the overwhelmingly complex issue of heterojunctions, quantum-wells, and the like.
- Now that we produced light, we must ask our *fourth* question: *How do we get it out of the semiconductor* ? Which percentage of the light produced will actually escape – some light, for sure, will be absorbed *in* the material.
    - Will it come out in all directions, with a preferred direction, or even as a ***LASER*** beam?
    - What do we have to do to optimize whatever we want?
    - How do we meet the two basic requirements for LASER activity – *inversion* and *feed-back* (whatever that means; we will come back to it)?
- We now ask the *fifth* question: *What can we do to modify the wavelength of the emitted light?*

- Can we "tune" a given semiconductor, or mix different ones? What are the criteria for success?

And *finally*, for question number *six*, we must consider the technology: *How do we make the needed materials and devices?*

- What technologies exist, what are the pros and cons?
- What can we use from **Si** technology? What kind of technologies do we need that are specific to optoelectronics ?

All these questions are interrelated; very often we can not deal with just one at a time. In other words, if you optimize one property, you will change most of the others, too.

- Selecting materials and processes for an optimized device is a complex process and still a topic of front-end research.
- While much progress has been made during the last **20** years or so, there is still no cheap and reliable *blue* semiconductor *Laser* although first prototypes based on the fairly new *technical* semiconductor **GaN** (gallium nitride) have been introduced around **1998**.
- It is a safe bet that we will see a lot more progress for the next **20** years, and that it will come from rather involved physics and materials research employing quite a bit of quantum theory, new concepts like "**photonic crystals**" and "**spintronics**" – buzz words that you, the student, may not have heard yet, but that may well become part of your professional career.

Material Issues

Pondering the questions from above, it becomes clear very quickly that we need several suitable semiconductors to cover all aspects of optoelectronics.

- It also becomes clear that we have to look at more properties than we did for **Si** – the *dielectric constant*, e.g., is a prime parameter now.
- The value of the band gap, too, is now of prime importance. (We wouldn't have cared much if it would have been somewhat different in **Si**!)
- The precise mechanisms for recombination are a prime matter of interest. In **Si**, all that counted was that you had some, but not too many deep level around midgap, giving a large recombination life time. With optoelectronics we may have to make sure that recombination proceeds exactly as needed.
- Let's start with looking at major semiconductors and their properties in detail. (If you are interested in more data, or in other semiconductors, have a look at the relevant data collection of the Ioffe Institute or at that of Derek Palmer.)

Properties	**Si**	**Ge**	**GaAs**	**InP**	**InSb**	**In$_{0.53}$Ga$_{0.47}$As**	**GaP**	**GaN**	**SiC**	**Diamond**	**Remarks**
Crystal											
Unit weight [mol]	29.09		144.63	145.79		168.545					
Density [g/cm^3]	2.33	5.32	5.32	5.49	5.77	5.49	4.14	6.1	3.166 (cubic) 3.211 (hex)	3.51	
Crystal structure	Diamond	Diamond	Zincblende (Sphalerite)	Zincblende	Zincblende	Zincblende	Zincblende	Wurtzite	Many variants: cubic, hex, rhombohedral	Diamond	
Lattice constant [nm]	0.5431	0.565	0.565	0.587	0.648	0.5867	0.545	a = 0.319 c = 0.519	a = 0.30 c many values	0.357	
Transport properties											
Band gap [eV]	1.12	0.66	1.42	1.35	0.17	0.75	2.26	3.4	2.39 - 3.26	5.47	
Type	indirect	indirect	direct	direct		direct	indirect	direct	indirect	indirect	
Effective e$^-$ mass [m*/m$_0$]	0.98						0.35	0.2	0.24 - 0.7		
Effective h$^+$ mass [m*/m$_0$] light heavy	0.16 0.49		0.082 0.45	0.12 0.56	7.3	0.051 0.50	0.14 0.79	0.3 1.4	0.9		
N_{eff} of CB [10^{18} cm^{-3}]	28 (32)	10.4	0.47	0.54	0.042	0.21	18				
N_{eff} of VB [10^{18} cm^{-3}]	10 (18)	6	7	2.9		7.4	19				
n_i [10^6 cm^{-3}]	6,600 (13,000)		2.2	5.7		63,000					
Mobility (undoped) [cm^2/Vs] μ_n μ_h	1,500 450	1,900 3,900	8,500 450	5,000 200	80,000 1,250	14,000 400	300 150	1,000 200	500 - 1,000 20 - 50	200 - 2,200 1,800 - 2,100	

Lifetime (general) [µs]	2,500		0.01	0.005		0.02					
Mechanism of luminescence	none		band-band	band-band		band-band	**exciton, if doped**	band-band			
Dielectric properties											
Dielectric constant at high frequency (static)	11.7	16	10.9 (12.9)	9.6 (12.4)	15.7 (16.8)	13.7	9.1 (11.1)	ϵ_{xx}, ϵ_{zz}: 5.35, 5.8 (9.5, 10.4)	ϵ_{xx}, ϵ_{zz}: 6.5, 6.7 (9.7, 10)	5.5	
Breakdown field strength [kV/cm]	300		350	400		100		5,000			
Specific intrinsic resistance [MΩcm]	0.2		310	11		0.000,8					
Electron affinity [eV]	4.0	4.05	4.07	4.4	4.59	4.63	4.3	4.1			
Thermal Properties											
Expansion coefficient [10^{-6} /K]	2.6	5.9	6.86	4.75	5.37	5.66	5.3	5.59 (a) 3.17 (c)		1	
Therm. conductivity [W/cmK]	1.5	0.58	0.45	0.68	0.18	0.05	1.1	1.3	5.0	22	Cu: 4.01
Specific heat [J/gK]	0.7	0.31	0.35	0.31	0.2	0.29	0.43	0.49	0.671	0.428	Cu: 0.38
Melting point [^{o}C]	1,412	937	1,238	1,062	527	970	1,457	2,500			

- Silicon is included as a reference (if there are several numbers, they are from different sources). We find expected properties, but also, perhaps, some unexpected ones.
  - *Dielectric constants are relatively large* even at the very high optical frequencies. This is not necessarily self-evident since at least for **Si** the only polarization mechanism operational is atomic polarization, i.e. the shift of electrons relative to the atom core.
  - There is at least one recombination mechanism not mentioned before: *Exciton* recombination.
- Let's continue by looking at recombination mechanisms in more detail.

5.1.2 Recombination and Luminescence

High Injection Approximations for Recombination Rates

- **Luminescence** is the word for *light emission* after some energy was deposited in the material.
  - **Photoluminescence** describes light emission stimulated by exposing the material to *light* – by necessity with a higher energy than the energy of the luminescence light. Photoluminescence is also called **fluorescence** if the emission happens less than about **1 µs** after the excitation, and **phosphorescence** if it takes long times – up to hours and days – for the emission.
  - **Cathodoluminescence** describes excitation by energy-rich *electrons*, **chemoluminescence** provides the necessary energy by *chemical reactions*.
  - Here we are interested in **electroluminescence**, in particular in **injection luminescence.**
- *Injection luminescence* occurs if *surplus* carriers are injected into a semiconductor which then recombine via a *radiative channel*.
  - This implies *non-equilibrium*, i.e. $n_e \cdot n_h > n_i^2$ and *net recombination rates* U given by the basic equation from the recombination theory for direct semiconductors:

$$\boxed{U = R - G_{therm} = r \cdot (n_e \cdot n_h - n_i^2) = r \cdot n_i^2 \cdot \left(\exp \frac{E_F^e - E_F^h}{kT} - 1 \right)}$$

  - Some, but not necessarily all of the recombination events described by U produce light, and these *radiative recombination channels* are of particular interest for optoelectronics.
- Since optoelectronic devices usually are made to produce *plenty* of light, the deviation of the carrier densities from equilibrium must be large to obtain large values of U.
  - If we write the densities, as before, as $n_{e,h} = n_{e,h}(equ) + \Delta n_{e,h}$, we now may use *the simplest possible approximation* called **high injection approximation**:

$$\boxed{\Delta n_{e,h} \gg n_{min}(equ)}$$

  - i.e. the minority carrier density is far *above* equilibrium.
  - That is different from the approximation made before, where we assumed that $\Delta n_{e,h}$ was small.
- The surplus carriers contained in $\Delta n_{e,h}$ are always *injected* into the volume under consideration (called **recombination zone** or **recombination volume**), usually by forward currents across a junction. They always must come in equal numbers, i.e. in pairs to maintain charge neutrality; otherwise large electrical fields would be generated that would restore neutrality. We thus have

$$\boxed{\Delta n_e = \Delta n_h}$$

  - The recombination volume usually is the space charge region of a junction or an other volume designed to have *low carrier densities* in equilibrium. Since the equilibrium density of both carrier types in the **SCR** is automatically very low, we may easily reach the high injection case. For a bulk piece of a (doped) semiconductor this is much more difficult – you would have to illuminate with extremely high intensity to increase the minority carrier density by some factor.
- The surplus density of carriers decays with a characteristic lifetime τ which is given by the individual life times of all recombination channels open to the carriers. Since $R \gg G_{therm}$ for the high injection case, we have in analogy to the approximation made for (small) deviations from equilibrium:

$$\boxed{U = R - G_{therm} \approx R = \frac{n}{\tau}}$$

  - We call this approximation (where we neglect G) "**high-injection** " approximation or the *high injection case* because the high density of surplus carriers is usually provided by injecting them over a forwardly biased junction into the region of interest.
  - Note that while the rate equations are formally the same for high or low injection (or everything in between), τ is not a constant but may depend on the degree of injection (as we will see).

- Now we have to look at all the possibilities for recombination – called **recombination channels** – that are open for carriers as possible ways back to equilibrium. Recombination channels generating light we will call **radiative channels**.
- The **band-band recombination channel** (with which we started above, using the full equations) can now be extremely simplified:

$$\boxed{R_{\text{b-b}} = \mathbf{v} \cdot \sigma \cdot n^2}$$

  - or, considering that **v** · σ may no longer be totally correct as the proportionality factor,

$$\boxed{R_{\text{b-b}} = B_{\text{b-b}} \cdot n^2}$$

  - and the index "***b-b*** " denotes band-band recombination. The proportionality constant ***B*** is occasionally called a **recombination coefficient**.
- If we use the same approximations for the recombination channel via deep levels, we obtain a rather simple relation, too, for the recombination rate R_{dl}

$$\boxed{R_{\text{dl}} = B_{\text{dl}} \cdot n}$$

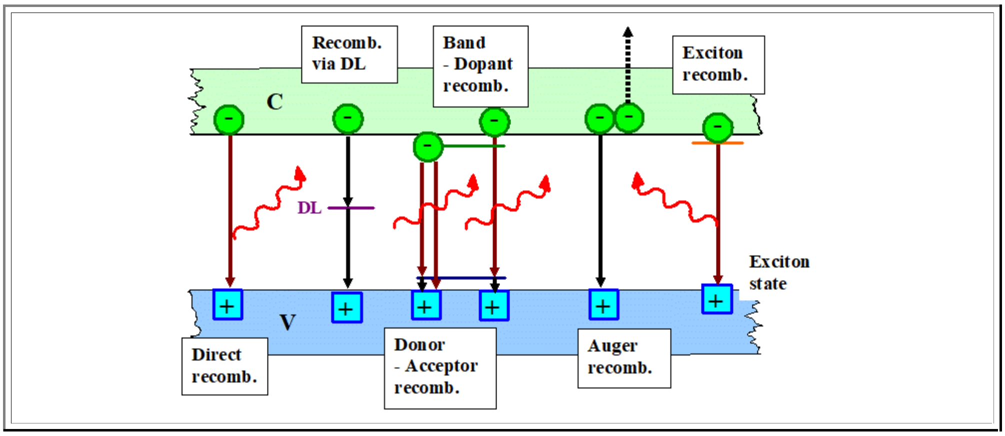
  - With B_{dl} = recombination coefficient for this case.
- Before we look at further recombination channels, we will give some thought to the *equilibrium case*.
  - In *thermal equilibrium*, we still have generation and recombination described by the equilibrium rates G_{therm} and R_{therm} and $U_{\text{therm}} = G_{\text{therm}} - R_{\text{therm}} = 0$.
  - Now a *tough question* comes up: If recombination occurs via band-band recombination and results in the emission of a photon, does this mean that our piece of semiconductor, just lying there, would emit photons and thus *glow in the dark*?
  - Obviously that can not be. Energy would be transported out of the semiconductor which means it would become cooler just lying there, a clear violation of the "second law". On the other hand, a single recombination event "does not know" if it belongs to equilibrium or non-equilibrium, so radiation must be produced, even in equilibrium. We seem to have a *paradox* .
- The apparent paradox becomes solved as soon as we consider that any piece of a material "glows" in the dark (or in the bright) because it emits and absorbs radiation leading to an equilibrium distribution of radiation intensity versus wave length – the famous "**black body**" **radiation** of **Max Planck**.
  - Recombination events in equilibrium do produce light – but the photons mostly will become reabsorbed and, in general, will not leave the material. The small amount that does escape to the environment must be exactly balanced by electromagnetic radiation absorbed from the environment.
  - This topic will be considered in more detail in an advanced module.

Additional Recombination Channels

- So far we considered only band-band recombination and recombination via deep levels. There are, however, more recombination channels, some of which are particular to compound semiconductors.
- But first we look at universal mechanisms occurring in all semiconductors. They are:
  - **Auger recombination**. In this case the energy of the recombination event is transferred to another electron in the conduction band, which then looses its surplus energy by "thermalization", i.e. by transferring it to the phonons of the lattice. This means that *no light is produced*.
  - Donor–acceptor recombination or recombination via "**shallow levels**". This includes transitions from a donor level to an acceptor level or to the valence band, and transitions form the conduction band to an acceptor level.
  - **Mixed forms**: From a donor level via a deep level to the valence band, etc.
- Now for *material specific* recombination channels. The most important one with direct technical uses is recombination via "**localized excitons**".
  - **Excitons** are something like hydrogen atoms (or, even closer in similarity, positronium = atom consisting of electron and positron) – except that a *hole* and not a proton is the partner of the electron. They are thus electron–hole pairs bound by electrostatic interaction. They can form in any semiconductor, are mobile and do not live very long at room temperature because their binding energy is very small. They decompose ("get ionized") into a free electron and a free hole.
  - If you wonder why they do not simply recombine, think about it. They can't possibly have the same wave vector (how would they "circle" each other then?) and thus need a third partner for the recombination process to occur.

- On occasion, however, they might become *trapped* at certain lattice defects and then recombine, *emitting light*. **GaP**, though an indirect semiconductor, can be made to emit light by enforcing this mechanism.
- We will come back to excitons later; more about them can be found in an advanced module.

- The picture below illustrates these points.



- The picture is far too simple and we will have to consider some of the processes in more detail later (especially recombination via excitons). Here we look at Auger recombination and donor–acceptor recombination.
  - Even without going into details, it is rather clear that (radiative) donor–acceptor recombination as well as band–dopant recombination (in both variants) are not all that different from direct (and radiative) band–band recombination. Especially for relatively high doping densities, when the individual energy levels from the doping atoms overlap forming a small band in the band gap, we might simply add the dopant states to the states in the conduction or valence band, respectively.
  - We then can treat donor-acceptor recombination as subsets of the band-band recombination, possibly adjusting the recombination coefficient $B_{b\text{-}b}$ somewhat.
- This leaves us with **Auger recombination**. This is an important recombination process that cannot be avoided *and that always reduces the quantum yield of radiation production*.
  - It has not been covered in the treatment of recombination before, and we will not attempt a formal treatment here. It is, however, simple to understand in the context of the high-injection approximation used for optoelectronics.
  - Since you need *three* carriers at the *same time* at the *same place* (the e^- and h^+ that recombine plus a third carrier to remove the energy), the Auger recombination rate, R_A, is proportional to the third power of the carrier density n:

$$R_A = B_A \cdot n^3$$

  - This means that for large carrier densities n (always way above equilibrium), and therefore large doping, Auger recombination sooner or later will be the dominating process, hence limiting the yield of radiative transitions.

Total Recombination and Quantum Yield

- All recombination processes will occur independently and the total recombination rate will be determined by the combination of all channels.
  - The situation is totally analogous to the flow of current through several resistors switched in parallel. The individual recombination rates R_i add up (like the currents) and for the total recombination rate we have

$$R_{total} = \Sigma_i R_i = \Sigma_i \frac{n}{\tau_i} = n \cdot \Sigma_i \frac{1}{\tau_i}$$

  - The total recombination time τ_{total} is thus defined by

$$\frac{1}{\tau_{total}} = \frac{1}{\tau_{b\text{-}b}} + \frac{1}{\tau_{dl}} + \frac{1}{\tau_A} + \frac{1}{\tau_{exciton}} + ...$$

  - Since we are only interested in radiative and non-radiative channels, we may write this as

$$\frac{1}{\tau_{\text{total}}} = \frac{1}{\tau_{\text{rad}}} + \frac{1}{\tau_{\text{non-rad}}}$$

$$R_{\text{total}} = R_{\text{rad}} + R_{\text{non-rad}} = \frac{n}{\tau_{\text{rad}}} + \frac{n}{\tau_{\text{non-rad}}}$$

- The quantum efficiency η_{qu} introduced before now can be calculated. It is given by the fraction of R_{rad} relative to R_{total}, or

$$\eta_{\text{qu}} = \frac{R_{\text{rad}}}{R_{\text{total}}} = \frac{1/\tau_{\text{rad}}}{1/\tau_{\text{rad}} + 1/\tau_{\text{non-rad}}}$$

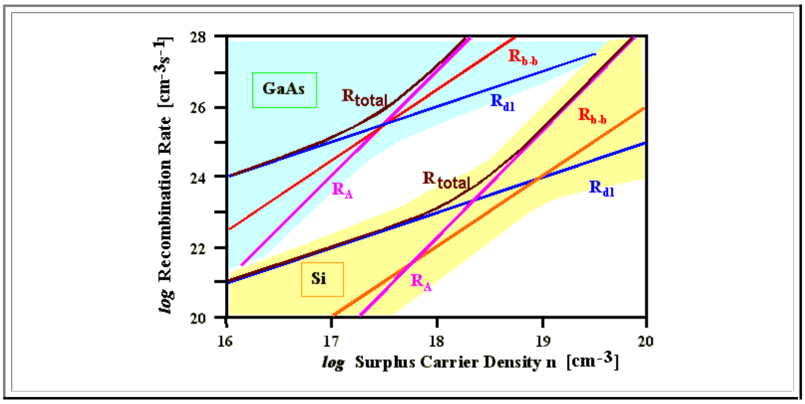
  - Obviously, the result is

$$\eta_{\text{qu}} = \frac{1}{1 + \dfrac{\tau_{\text{rad}}}{\tau_{\text{ non-rad}}}}$$

- That is easy enough, but now need some numbers for the recombination coefficients in order to get some feeling for what is going on in different semiconductors.
  - It should be clear that the B_i defined above are related to quantities like the thermal velocity, the capture cross sections, the density of deep (and shallow) levels, and so on – they depend to some extent on the particular circumstances of the semiconductor considered. e.g. doping, cleanliness, defect density, etc.
  - It should also be clear the B_i are not absolute constants for a given materials but only useful as long as the approximations used are holding. in other words, there are no universal numbers for a certain semiconductor. We only can give typical numbers.
  - With this disclaimers in mind, we use the following values (if two numbers are included, they come from different sources). Yellow denotes the indirect semiconductors and the **GaP** value is for the very unlikely direct recombination without excitons.

(T = 300K)	**Si**		**Ge**		**GaAs**		**InP**		**GaP**	
	B	τ [µs]	*B*	τ [µs]	*B*	τ [µs]	*B*	τ [µs]	*B*	τ [µs]
B_{dl} [s^{-1}]	$1 \cdot 10^5$				$1 \cdot 10^8$					
$B_{\text{b-b}}$ [cm^{3}s^{-1}]	$1 \cdot 10^{-14}$ $1.8 \cdot 10^{-15}$	5,500	$5.3 \cdot 10^{-14}$	200	$3 \cdot 10^{-10}$ $7.2 \cdot 10^{-10}$	0.015	$1.26 \cdot 10^{-9}$	0.008	$5.4 \cdot 10^5$	2,000
B_{A} [cm^{6}s^{-1}]	$2 \cdot 10^{-32}$				$1 \cdot 10^{-27}$					

- Now we can construct a *recombination rate vs. surplus carrier density diagram* as follows:



- We can see a few interesting points:
  - The recombination rate in **Si** is generally much smaller than in **GaAs** – a direct effect of the much larger lifetimes.
  - Direct recombination in **Si** is not strictly forbidden – rather, it is just unlikely. At a typical carrier density of **10^{18} cm^{-3}** we have about **10^{22}** photons generated in **Si** per **s** and **cm^3** compared to about **$3 \cdot 10^{26}$** in **GaAs**.
  - ***R_{b-b}*** in **GaAs** is similar to the recombination rates of the Auger and deep level channels at densities around **$4 \cdot 10^{17}$ cm^{-3}**, whereas in **Si**, for most densities ***R_{b-b}*** is much smaller than the other recombination rates.
  - Although for large carrier densities the Auger recombination process always dominates, it may still be useful to increase ***n***: While the quantum efficiency goes *down*, the amount of light produced still *increases* with ***n***.
  - For very large carrier densities (say **10^{19} cm^{-3}** and beyond as occasionally encountered in power circuits), even **Si** may produce some visible light.
- The **GaAs** curves now provide a first answer to our second question about the quantum efficiency.
  - For ***n*** **= 10^{16} cm^{-3}**, we have about **$4 \cdot 10^{22}$** radiative recombination events per **s** and **cm^3** out of a total of about **10^{24}** per **s** and **cm^3** , which gives a quantum efficiency of **4 %**.
  - At the high density end, around ***n*** **= 10^{19} cm^{-3}**, the situation is similar, the quantum efficiency is in the few percent range.
  - The highest quantum efficiency is around **30 %** for densities around ***n*** **= $5 \cdot 10^{17}$ cm^{-3}**.
- Of course, given the values of the recombination coefficients, we could calculate the quantum efficiency precisely, but that would not be very helpful because real devices are more sophisticated than the simple forwardly biased junction implicitly assumed in this consideration.
  - This means that we now must look more closely at the important compound semiconductors, especially on how they are doped and what typical differences to **Si** occur.
  - We will, however, first do a little exercise for injection across a straight **p-n** junction in order to get acquainted with some real numbers for carrier densities produceable by injection.

Exercise 5.1.2-1

Calculate carrier densities from the forward current of junctions.

5.1.3 Doping of Compound Semiconductors

Some Basic Considerations

- Essentially, the semiconducting properties of silicon stem from the **sp^3** hybrid bonds formed between electrically neutral atoms.
  - Two **Si** atoms donate one electron each to all four **sp^3** hybrid bonds, forming the familiar diamond type lattice.
  - Substituting a **Si** atom by a group **III** or group **V** element produces a mobile hole or electron *and* an immobile ion in the familiar way.
- All **III-V** compounds also form **sp^3** hybrid orbitals, but there is now a *big difference* to **Si** (or **Ge**, or diamond-**C**): The binding, which was *totally covalent* for the elemental semiconductors, now has an *ionic component*.
  - This is simply due to the fact that different atoms differ in their electronegativity, describing the affinity to electrons of the element. The numbers in brackets give the electronegativity quantitatively: The larger the value, the stronger the effect.

IIa, b	III	IV	V	VI
Be (1.6)	B (2.0)	C (2.6)	N (3.1)	O (3.4)
Mg (1.3)	Al (1.5)	Si (1.9)	P (2.2)	S (2.6)
Zn (1.7)	Ga (1.8)	Ge (2.0)	As (2.2)	Se (2.6)
Cd (1.5)	In (1.5)	Sn (1.7)	Sb (1.8)	Te (2.7)
Hg (1.9)	Tl (1.6)	Pb (1.6)	Bi (1.7)	Po (2.0)

  - A more electronegative element will attract the electrons from the partner more strongly, become more negatively charged, and thus increase the ionic part of the binding.
- The percentage ***p*** of the ionic binding energy varies for the various compounds.
  - The difference in electronegativity of the atoms in a compound semiconductor gives a first measure for ***p***.
  - To give some examples: For **Si** we have ***p* = 0**, for **GaAs** we find ***p* = 0.08**.
- Doping is still achieved by introducing specific atoms as *substitutional impurities*. But in contrast to elemental semiconductors, we now have more possibilities as can be seen by looking at the relevant part of the periodic table (the elements in yellow cells are never used for doping of *compound* semiconductors).
  - We can replace the group **III** elements by group **II** elements to produce *acceptors* and the group **V** elements by group **VI** elements to produce *donors* – in principle.
  - But we can also replace both the group **III** and group **V** elements by group **IV** elements which, however, may generate donor *or* acceptor levels in the band gap of some compounds – we have **amphoteric doping**.
  - We could also replace the atoms of the compound by an **isoelectronic** atom – group **III** elements by some other group **III** elements, and the same thing for the group **V** partner. In **Si**, this would mean replacing a **Si** atom by e.g. a **Ge** or **C** atom – which is not very exciting. In compounds, however, *doping with isoeletronic atoms* produces differences in the ionic part of the binding and therefore *local potential differences* with noteworthy effects as we shall see below.
  - We could even replace an atom of the lattice by a small molecule – isoelectronic or not – and achieve a doping effect.
- Well, in **Si** we could also use molecules and all group **III** or group **V** elements – but in reality only **B**, **As**, **P**, and sometimes **Sb**, **Ga** or **Al** are used.
  - In **Si**, we do not use the group **III** elements **In** and **Tl**, neither the group **V** elements **N** and **Bi**. The reasons are "technical": They may be difficult to incorporate in a crystal, their solubility may be too small, their diffusivity too high (or too low?), or their energy levels in the band gap not suitable.
- The same situation occurs with compound semiconductors.
  - There are *optimum solutions* to doping, depending on the type of semiconductor, the technology available or mandated by other criteria, and so on and so forth.
  - There are therefore no general rules for optimal doping, and here we will only discuss *amphoteric* doping and *isoelectronic* doping in somewhat more detail.
  - The energy levels of some dopants and other impurities in some **III-V** semiconductors are shown in an illustration module.

Amphoteric Doping

- A prime case of amphoteric doping is the incorporation of **Si** into **GaAs**. If the **Si** atoms replace **Ga** atoms, they act as *donors*, and as *acceptors* if they occupy **As** lattice sites. How does this work?
  - A **Si** atom has four electrons disposable for binding. If it replaces a **Ga** atom that had only three electrons, the **As** partner has to supply one electron less than before to make up for its "deficient" partner **Ga**, and the surplus electron will only be weakly bound – it will easily escape into the conduction band.
  - ***Si*** on a ***Ga*** *site* causes the release of an electron from the **As**, because the electrons are more strongly bound to **Si** than to **As** – thus, in this case **Si** acts as *donor* even so the electron is actually supplied by the **As**.
  - Contrariwise, if an ***As*** *atom* is replaced by a ***Si*** atom, the new twosome is now short one electron. It can therefore take up an electron from the valence band – **Si** now acts as an *acceptor*.
- While this seems to offer an elegant way for doping, the tough question now is: How do we *control* which lattice sites are occupied by **Si**?
  - In other words: On what kind of lattice sites will we find the **Si** atoms after it was ion-implanted, diffused into the crystal from the outside world, or incorporated directly during crystal growth procedures, or thin film growth?
- This question cannot be easily answered from first principles. Two guidelines are:
  - At *low* temperatures (**$T \leq 700$ °C**), **Si** will prefer to sit on **As** sites – it acts as an *acceptor*, with an energy level about **35 meV** above the valence band edge.
  - At *high* temperatures (**$T \geq 900$ °C**), it tends to sit at **Ga** sites and acts as a *donor*, with an energy level about **6 meV** below the conduction band edge.
  - If **Si** is incorporated into a **GaAs** melt, *large* **Si** concentrations tend to produce *donors*, *small* concentrations *acceptors*.
- More about amphoteric doping and its practical aspects will follow in the various chapters about specific compound semiconductors.

Isoelectronic "Doping" and Bound Excitons

- If we introduce isoelectronic replacements in the lattice – e.g., a group-V atom (like **N** or **Sb** substituting a **P** atom), or a molecule (like **ZnO** substituting a **GaP** pair in the **GaP** lattice), or even a larger entity (like the **Li-Li-O complex in GaP**, with one interstitial **Li**, the other **Li** substituting **Ga**, and **O** substituting **P**) – we do not "dope" in the conventional sense of the word. We rather change the ionic component of the local binding.
  - Since the introduction of isoelectronic elements is deliberate with a specific purpose in mind, we deal with it under the general heading "doping", keeping in mind that we do not change the carrier *density* in this way, but their *recombination behavior*.
  - Doping with isoelectronic elements may not do much in most compound semiconductors, but it can have pronounced effects in others by providing *new* radiative recombination channels due to an interaction of the isoelectronic dopant and electron–hole pairs called *excitons*.
  - The paradigmatic semiconductor for isoelectronic doping is **GaP**, an *indirect* semiconductor. It is used as a strong emitter of green light, however, by doping it with isoelectronic elements and using the *radiative decay of excitons bound to the doping elements*.
- How does this work? There are several crucial ingredients, all from rather involved solid state physics:
  - First, we need *excitons*.
  - Second, we need to have the *excitons bound to isoelectronic dopants*.
  - Third, we need to have a *radiative recombination* of the electron and hole constituting the exciton – *despite the nominal violation of the crystal momentum*.
- A detailed treatment of these points (explaining, among other things, why this "works" in **GaP**, but not in many other compound semiconductors) is not possible in the context of this lectures course. We will only superficially look at the basics; somewhat more information is contained in an advanced module.
- Again: What is an **exciton**? Imagine the generation of an electron–hole pair, e.g., by irradiating a semiconductor with light.
  - If the photon energy is large enough to lift an electron all the way from the valence band to the conduction band in a direct transition, you now have a *free* electron and a *free* hole which move about the crystal in a random way.
  - Now imagine that the hole and the electron stay so close to each other (i.e. less than a Debye length) in *real space* that they feel the Coulomb attraction. They are then *bound* to each other with a certain binding energy E_b, and their coordinates in r-space are (nearly) identical, but still undetermined.

  - Bound together like that, they behave like an individual particle, and this particle is called exciton (standard symbol: **X**). Note that, since it consists of two fermions with spin = ½, an exciton is a boson. Thus, in principle, arbitraily many excitons can be present.

- We already know a system where *one* negative elementary charge is bound to *one* positive one, forming a new "particle" – it is called "*hydrogen atom*". The only difference between an exciton and a hydrogen atom is that the mass of the hole is much smaller than the mass of the proton (and, of course, that our exciton can only exist in a solid).
  - From a more advanced treatment of the hydrogen atom, where the electron mass is *not* neglected with respect to the proton mass, we can immediately carry over the solution of the relevant Schrödinger equation to an exciton. Of interest are especially the allowed energy states E_X of the exciton, for which we obtain

$$E_X(k) \;=\; E_{gap} \;-\; \frac{m_{red}\cdot e^4}{8\cdot(h\cdot\epsilon_0\,\epsilon_r\cdot n)^2} \;+\; \frac{h^2\,k^2}{8\pi^2\cdot(m_e+m_h)}$$

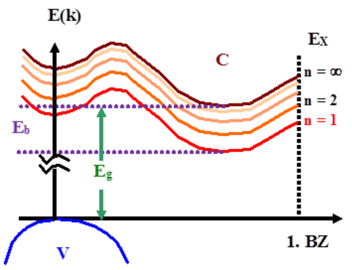
  - With n = quantum number **= 1, 2, 3, ...** and m_{red} = *reduced* mass (from $1/m_{red} = 1/m_e + 1/m_h$). Although not written explicitly, always the relevant *effective* masses (of electron and hole) have to be used for the exciton.
  - The first term simply accounts for the crystal energy, the second one is straight from the hydrogen atom (see below for details), and the third term is a correction if the two particles are not at the same place in ***k***-space (it is zero for $k_e = -k_h$ or $k_e = k_h = 0$ as it will be for most direct semiconductors), giving the kinetic energy of the exciton as a whole.

- In total we have a system of energies *smaller than the band gap energy*, with the "deepest" level defined by the following energy difference relative to the band gap – which is just the exciton binding energy mentioned above (also known as the Rydberg energy of the exciton), being directly related to the standard Rydberg energy of the hydrogen atom (**1 Ry ≈ 13.6 eV**) as follows:

$$E_b \;=\; \frac{m_{red}\cdot e^4}{8\cdot(\epsilon_0\,\epsilon_r\cdot h)^2} \;=\; \frac{m_{red}}{m_0\cdot\epsilon_r^2}\times 1\ \mathrm{Ry}$$

  - To estimate the order of magnitude relevant for the exciton binding energy, we use **0.5** m_0 for both effective masses, giving $m_{red} = 0.25\ m_0$, and a (static) dielectric constant of $\epsilon_r \approx 10$, resulting in $E_b \approx 34$ **meV**.
  - This is already a rather reasonable value. The actual exciton binding energy depends mainly on the effective mass of the respective semiconductor. As an example, a properly calculated value for **GaAs** gives $E_b = 4.4$ **meV**, which is close to the experimental one.
  - More values can be found in the advanced module.

- Due to the exciton binding energy, it takes exactly E_b less energy to create an exciton than a free electron–hole pair. Since both creation processes start by exciting the electron from the top of the valence band, one commonly finds the picture of excitonic energy levels sitting in the band gap, close to the conduction band.
  - This involves several fundamental misconceptions:
– The symmetry between electron and hole, valid for the exciton, is lost since the hole is imagined to stay at the top of the valence band;
– the excitonic energies do not represent usual electronic states of the crystal, because excitons are bosons.
  - Nevertheless, for simplicity reasons we stick to this picture. In ***k***-space for **GaP** this looks like this:



  - Since E_b for a free exciton – that moves about the crystal, transporting energy, but not charge – is just a few **meV**, similar to the typical energy difference from a donor level to the conduction band, it will not live very long at room temperature: The thermal energy then is enough to ionize the exciton, i.e. to remove the electron (or the hole; your choice – everything is rather symmetrical), and we are left with a free hole and a free electron.

  - Even so the electron and the hole are almost at the same place, recombination is *not* possible without the help of phonons, so it is rather unlikely – as stated before. Excitons in most semiconductors therefore only make their presence known at low temperatures – and in the absorption of light, because you will already find some absorption for light with an energy slightly below the band gap energy!

- Now imagine an *isoelectronic dopant in* ***GaP***, e.g. **N** instead of **P**. It distorts the potential for electrons a little bit *and strictly locally*; and in **GaP** this will lower the energy of the electrons *locally* – inside a radius similar to the lattice constant.
  - An electron thus may become bound to the isoelectronic dopant – i.e. it "revolves" around the isoelectronic atom. It may now attract a hole by Coulomb interaction and thus form a **bound exciton**. This is an easy, albeit oversimplified way, to conceive bound excitons.

- The net effect of a binding interaction of an isoelectronic dopant and an exciton is twofold:
  - The exciton energy levels move "down" from the free exciton level by an amount equal to the binding energy of the electron to the isoelectronic dopant; i.e. E_b *increases*. In some cases – naturally for **GaP** – the additional binding energy may be in the order of **10 meV**, and this pushes the exciton levels so far below the conduction band that the bound exciton is now *relatively stable* at room temperature.

- The exciton is now *localized* in space. This demands that its coordinates in $\boldsymbol{k}$-space must be somewhat undetermined thanks to the *uncertainty principle* which requires that

$$\boxed{\Delta(\hbar\, \boldsymbol{k}) \cdot \Delta \boldsymbol{r} \;>\; h}$$

  - With $\hbar \boldsymbol{k}$ = momentum. Since $\Delta \boldsymbol{r}$ is in the order of the length scale of the attractive potential, i.e., a *lattice constant*, $\Delta \boldsymbol{k}$ will be in the order of $\boldsymbol{a}/(2\pi)$, i.e., a *Brillouin zone width*.
  - In other words, the *bound exciton* can have *any* wave vector in the **1st** Brillouin zone with a certain, not too small probability.

- This has an important consequence: *Recombination for* ***bound*** *excitons is easy!* It is still an indirect recombination that needs a phonon as a third partner. But in contrast to indirect recombination between free electrons and holes, which needs a phonon with a precisely matched $\boldsymbol{k}$-vector, *any* phonon will do in this case because it always matches one of the $\boldsymbol{k}$-vectors from the spectrum accessible to the bound exciton.
  - If bound excitons exist (at room temperature), their recombination provides a very efficient channel for establishing equilibrium and thus a possibility to generate light with an energy given by the exciton energy, i.e., the bandgap minus a small exciton binding energy.
  - This is essentially the mechanism to extract light out of the *indirect* semiconductor **GaP**!

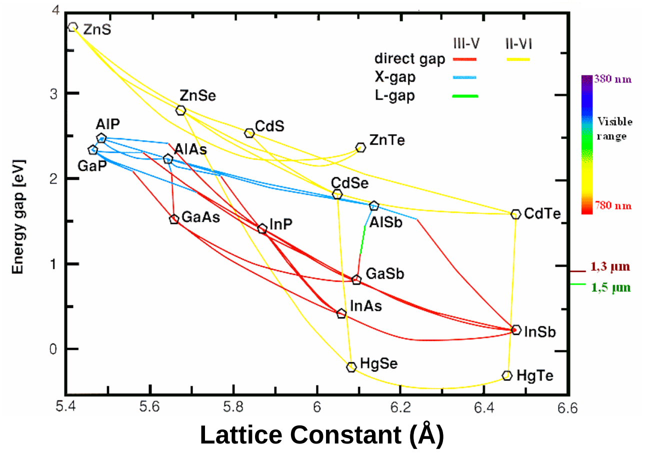
- You should now have a lot of questions:
  - Why **GaP**? How about other **III-V** compound semiconductors?
  - How about more exotic semiconductors, e.g. the **II-VI** system or organic semiconductors?
  - Anything similar for elemental semiconductors? After all, putting **Ge** into **Si** also changes the potential locally.
  - How about other defects, not necessarily isoelectronic? For example, ionized donors and acceptors also attract and possibly "bind" free electrons or holes, respectively?

- Well, this is not an advanced solid state lecture course. And even there, you may not find all the answers easily. Some answers I would have to look up, too; some, however, you can work out for yourself – at least sort of.

- Of course, there is another question looming large by now: How is doping actually *done* – at least for the more common non-**Si** semiconductors out there? What are particular problems and limits?
  - Again, answering these questions in any detail would far exceed what is possible to do during this lectures course.
  - We will, however, touch upon the subject in various and any places in the course of the following chapters.

5.1.4 Wavelength Engineering

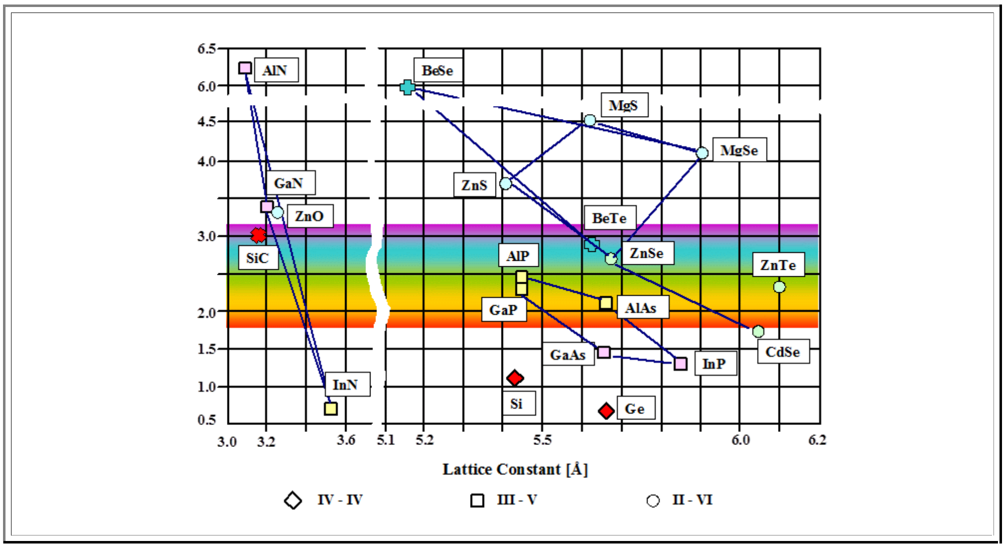
- We will now try to find some answers to our fifth question: How can we change the wavelength of the light produced by radiative recombination?
  - This question is to be understood in the sense of "changes beyond just choosing from given materials having different band gaps".
  - The recipe coming to mind is: Mix two similar (direct) compound semiconductors with different bandgaps.
  - Luckily, most **III-V** compounds are completely miscible in ternary or even quaternary crystals.
  - In other words: From the **2** compounds **GaAs** and **AlAs** we can make ternary $\mathbf{Ga_{1-x}Al_xAs}$ for $\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}$, from **GaAs** and **InP** we can produce quaternary $\mathbf{Ga_{1-x}In_xAs_{1-y}P_y}$.
- This gives a lot of options. What happens upon mixing, which changes of properties are useful, and which are not? Are there guidelines or do we have to try it out?
  - Generally, all properties of interest as given in a table in subchapter 5.1.1 will change while ***x*** and ***y*** run through the accessible range, but not necessarily linearly (or even monotonously) with the composition.
  - Here we focus on just a few of the especially important properties:
    - *Bandgap magnitude*
    - *Bandgap type* (direct or indirect)
    - *Lattice constant*
    - *Thermal expansion coefficient*
  - The two last properties will be of overriding technical importance as soon as we learn how to make heterostructures, i.e., combinations of two different semiconductors.
- There are some standard diagrams showing major properties of the most important combinations.
  - The first and most important one shows the *bandgap vs. the lattice constant* plus information about the gap type. It is shown below, with the **II-VI** compounds included for good measure:



- There is a tremendous amount of information in this diagram (note that "**X**-gap" and "**L**-gap" both denote *indirect* band gaps at the respective positions in the band diagram):
  - Most **III-V** compounds radiate at wavelengths above the visible region, i.e., in the infrared. However, adding some **Al** to **GaAs**, producing $\mathbf{Al_xGa_{1-x}As}$, will shift the wavelength into the red region of the spectrum – here are our red luminescence diodes and lasers!
  - *Very fortunate*: **GaAs** and **AlAs** have almost the same lattice constant; we can thus combine (e.g., in a stack of layers) any mixtures of these materials without encountering mechanical stress.
  - *Very unfortunate*: There are *no* **III-V** compounds in the diagram that emit *blue* light – which is a severe problem for many potential applications. While in the past, **SiC** could be used to some extent, it was only with the recent (early 1990s) advent of **GaN** that this problem was solved.
  - **SiC** and **GaN** crystals, however, are not of the "zinc-blende" type common to all the **III-Vs** in the diagram but have a *hexagonal* unit cell. *They therefore do not easily mix with the others!* To grow **GaN** layers (bulk crystals can hardly be produced!) it is therefore favorable to use a hexagonal substrate as, e.g., **SiC**, $\mathbf{Al_2O_3}$ (sapphire), or **Si(111)**, plus a special buffer layer.
  - If we want to radiate at **1.3 µm** or **1.5 µm** – infrared wavelengths of prime importance for optical communications – we should work with combinations of **GaAs**, **InAs**, and **InP**.

- Most interesting: The **II-VI** compounds are *all* direct semiconductors and span a much larger range of wavelengths than the **III-V's**. The fact that they are not much used for products tells us that there must be big problems in utilizing these compounds for mass products (prominent exception: thin-film PV modules made of **CdTe**).

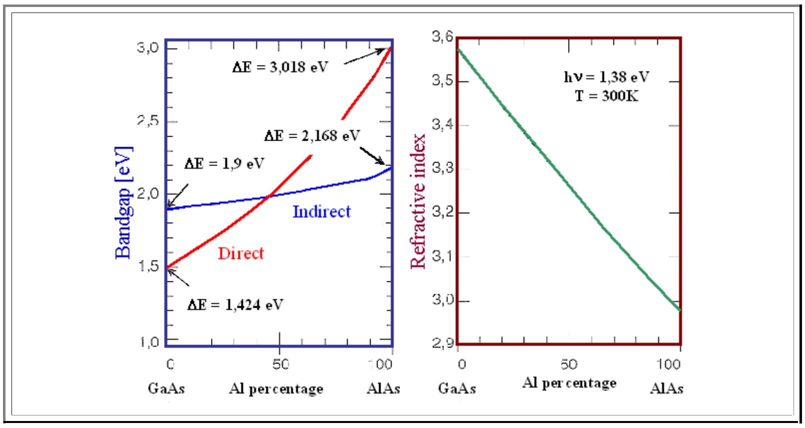
Here is another picture of the same thing including more materials:



- In the left part, all hexagonal materials are shown; for them, the value of the in-plane lattice constant (i.e., the one perpendicular to the hexagonal axis) is used.
- The group-III nitrides **AlN**, **GaN**, and **InN** span an extremely wide range of band gap energies, thereby providing a plethora of design opportunities for devices. In the first two decades of the 21st century, there were a lot of activities in this field, and the development still goes on.
- The success in this field is mainly due to two basic achievements: First, to find out how *stable p-type doping* can be reached; second, to find out how the materials can be *grown with sufficient quality* regarding high purity and low defect density, especially dislocation density (that's one of the reasons the special buffer layer mentioned above is needed for).
- The first problem was overcome by Shuji Nakamura, the second one by him and by Isamu Akasaki and Hiroshi Amano. Together, they received the **2014 Nobel prize in physics**. The Nobel lecture given by Nakamura ("Background Story of the Invention of Efficient InGaN Blue-Light-Emitting Diodes") provides valuable insight into present-day semiconductor research and device development.
- **Zinc oxide** has already found many different applications, but only very few in the field of electronics. This is mainly hampered by the same difficulty originally also faced for the nitrides: Stable p-type doping is still a problem. Nevertheless, **ZnO**-related research and development in other fields of materials science is also very interesting; see, e.g., the relevant activities in the group of Prof. Adelung here at the TF.
- In addition, among the group-III nitrides there is also **BN**, showing some similarity to **carbon**: The most stable form is **graphitic boron nitride** (i.e., covalently bonded hexagonal layers kept together by van der Waals forces), whereas the cubic modification **c-BN** (having zinc-blende structure) is metastable, just like **diamond**. Graphitic BN has a band gap of 5.2 eV, it is mainly used as a lubricant (c-BN: 6.4 eV, used as abrasive).

Let us now look a bit more closely at some other properties for the technically more relevant systems.

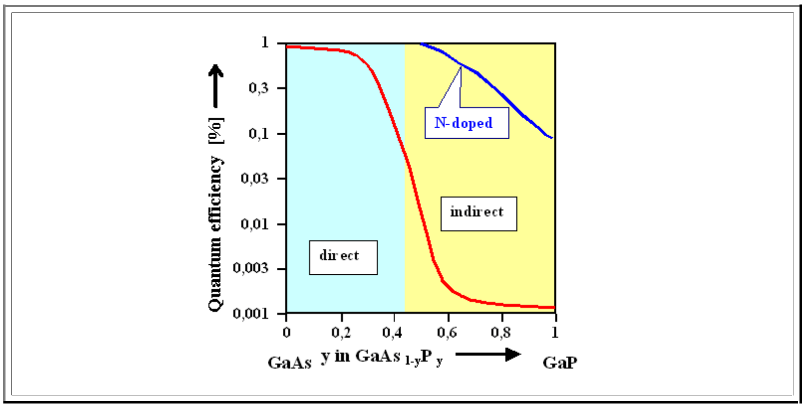
- The following diagrams show the direct and indirect bandgap and the refractive index for **$Ga_{1-x}Al_xAs$** as a function of **x**.



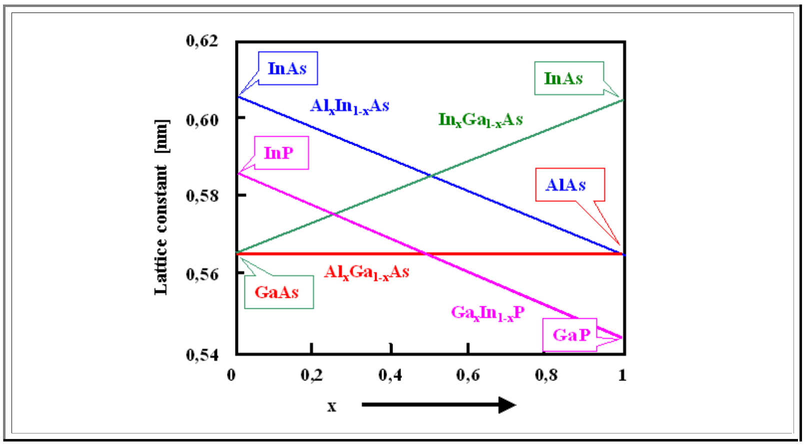
Mixing does not only affect the band gap and the lattice constant, but also the quantum efficiency of light production. The next figure shows the mixture **$GaAs_{1-y}P_y$**.

- The quantum efficiency decreases rapidly as the material approaches the indirect bandgap region.

- If an isoelectronic center – **N** in this case – is added, the **GaP** side obtains a strong radiative recombination channel via bound excitons and the quantum efficiency is two orders of magnitude larger.



- Next the lattice parameters of various mixtures as a function of **x** are shown.
  - This is easy to calculate; for complete mixing (no precipitation etc.), the lattice parameter changes linearly with the composition index **x** between the values for **x = 0** and **x = 1**.



- Finally, the technically most important systems are listed together with some key properties:
  - We see that all kinds of *ternary* and *quaternary* compounds are used, and that the *external* or total efficiency – the relation of light *out* to total power *in* – is relatively small in most cases. The external efficiency should not be confused with the quantum efficiency (relation of light *produced* to total power minus ohmic losses), since some of the light produced may never leave the device – remember the fourth question!
  - Also remember that the total efficiency of a light bulb is just a few percent. The semiconductor values don't look so bad in this context, and that for **GaAs** we can get up to **30 %** in extreme cases (and beyond!) is encouraging. From that perspective, the blue emission efficiency of the group-III nitrides is beyond expectations. Note that the somewhat exotic exciton process can account for an efficiency of **15 %**!!

Material (Dopant)	**Wavelength [nm]**	**Transition**	**External Efficiency [% of power]**	**Color**
$Al_{0.60}Ga_{0.40}N$	265	band–band	2 - 10	UV
$In_{0.16}Ga_{0.84}N$	445	band–band	50 - 70	purple
SiC (Al, N)	480	defect-related (stacking fault?)	0.01 - 0.05	blue
GaP (N)	565	exciton	0.1 - 0.7	green-yellow
$GaAs_{0.15}$ $P_{0.85}$ (N)	590	exciton	0.1 - 0.3	yellow-orange
$GaAs_{0.3}$ $P_{0.7}$ (N)	630	exciton	0.4 - 0.6	orange-red

$GaAs_{0.6}P_{0.4}$	650	band–band	0.2 - 0.5	red
$Ga_{0.6}Al_{0.4}P$ (N)	650	band–band	1 - 3	
GaP (ZnO)	690	exciton	4 - 15	
GaAs	870	band–band	0.1	infrared
GaAs (Zn)	900	band–acceptor	0.5 - 2	
GaAs (Si)	940	deep level	12 - 30	
$In_{0.73}Ga_{0.27}As_{0.58}P_{0.42}$	1310	band–band	1 - 2	
$In_{0.58}Ga_{0.42}As_{0.9}P_{0.1}$	1550	band–band		

5.2 Light and Semiconductors

5.2.1 Total Efficiency of Light Generation

Contributions to the Total Efficiency

- We will now give some thought to the second and third question raised before: *How much light is produced by recombination?* This raises the question for the value of the quantum efficiency η_{qu} mentioned before, and the total or external efficiency η_{ext} in absolute terms (third question).
  - First we will define the quantum efficiency again (and somewhat more specifically) and relate it to some other efficiencies.
  - The **quantum efficiency** η_{qu} is defined as

$$\boxed{\eta_{qu} = \frac{\text{Number of photons generated in the } \textit{recombination zone}}{\text{Number of recombining carrier pairs in the recombination zone}}}$$

  - We already know that, it could be expressed as

$$\boxed{\eta_{qu} = \frac{1}{1 + \tau_{rad} / \tau_{non\text{-}rad}}}$$

- In the high injection approximation the number of carriers is about equal to the number of carriers injected (across a junction) into the recombination zone.
  - That part of the total recombination occurring via a radiative channel determines the quantum efficiency. However, the surplus carriers in the recombination zone have one more "channel", not considered so far, for disappearing from the recombination zone: *They simply move out*!
  - In other words: parts of the injected carriers will simply flow across the recombination zone and leave it at "the other end".
- This effect can be described by the **current efficiency** η_{cu}; it is defined as

$$\boxed{\eta_{cu} = \frac{\text{Number of recombining carrier pairs in the recombination zone}}{\text{Number of carrier pairs injected into the recombination zone}}}$$

- We now define the **optical efficiency** η_{opt} as

$$\boxed{\eta_{opt} = \frac{\text{Number of photons in the exterior}}{\text{Number of photons generated in the recombination zone}}}$$

  - The optical efficiency takes care of the (sad) fact that in most devices a large part of the photons generated become reabsorbed or are otherwise lost and never leave the device.
- The total or **external efficiency** η_{ext} now simply is

$$\boxed{\eta_{ext} = \eta_{opt} \cdot \eta_{qu} \cdot \eta_{cu}}$$

  - In other words: The external efficiency is limited because not all injected carriers recombine, not all recombinations produce photons, and not all photons leave the material.
- If we want to optimize the external efficiency, we must work on all three factors - none of them is negligible.
  - We already "know" how to optimize the quantum efficiency η_{qu} by looking at the equation above. We must look for the best combination of materials producing radiation at the desired wavelength, and then dope it in such a way as to maximize the radiative channel(s) by minimizing the corresponding lifetimes. While this is not easy to do in practice, it is clear in principle.

- We do not yet know how to attack the two other problems: Maximized current efficiency and maximized optical efficiency. And theses problems are far from being solved in a final, or just semi-final way - intense world-wide research efforts center on new solutions to these problems.
  - While there are no general solutions to these problems and only some useful equations, a few general points can still be made. We will do this in the remainder of this subchapter for the current efficiency and in a separate subchapter for the optical efficiency.

Current Efficiency

- The question to ask is: Why is the current efficiency *not* close to **1** in any case?
  - After all, if we consider a simple **p–n** junction biased in forward direction in a direct semiconductor (e.g., **GaAs**), we inject electrons into the **p**-part and holes into the **n**-part, where they will become minority carriers. Some of the injected carriers will recombine in the space charge region, all others eventually in the bulk region.
  - While the quantum efficiency may be different in the different regions, because the strength of the recombination channels depends on the carrier density which is not constant across the junction, we still could assign some kind of mean quantum efficiency to the diode so that $\eta_{cu} = 1$.
  - However, we defined the efficiencies relative to a "*recombination zone* ", i.e we are not interested in radiation produced elsewhere for various reasons (to be discussed later). If we take the recombination zone to be identical to the **SCR**, only that part of the injected carriers that recombines in the **SCR** will contribute.
  - This is exactly that part of the forward current that we had to introduce to account for real ***I-V***-characteristics of **p–n** junctions – cf. the simple and advanced version in the relevant modules.
  - That part of the current that injects the carriers which *recombine in the SCR* was given by

$$\boxed{j_{rec}\,(\mathrm{SCR}) \;=\; \frac{e \cdot n_i \cdot d}{2\tau} \cdot \exp \frac{e \cdot U}{2kT}}$$

  - ***d*** was the width of the **SCR**.

- The current efficiency in this case would then be given by

$$\boxed{\begin{aligned}\eta_{cu} \;&=\; \frac{j_{rec}}{j_{diode}} \;=\; \frac{j_{rec}}{j_{non\text{-}rec} + j_{rec}}\\[2ex] &=\; \frac{1}{1 + \dfrac{j_{non\text{-}rec}}{j_{rec}}}\end{aligned}}$$

- With j **non-rec** (assuming that the electron and hole contributions and parameters are equal) given by the "simple" diode equation as

$$\boxed{j_{non\text{-}rec} \;=\; \frac{2 \cdot e \cdot L \cdot n_i^2}{\tau \cdot N_{Dop}} \cdot \left(\exp \frac{e \cdot U}{kT} - 1\right)}$$

  - we obtain for η_{cu} (neglecting the **–1** after the exponential)

$$\boxed{\eta_{cu} = \frac{1}{1 + \dfrac{4 \cdot L \cdot n_i}{d \cdot N_{Dop}} \cdot \exp \dfrac{eU}{2kT}}}$$

- η_{cu} thus decreases exponentially with the applied voltage and it would not make sense to include this effect in some averaged η_{qu}.

Why are we looking at radiation only from some confined part of the device, i.e., from the *recombination zone* , and not at the total volume, which demanded a finer look at the efficiencies? There are *practical* reasons, e.g.:

- If we consider a semiconductor *Laser* , only the radiation inside the "*resonator*" counts – everything outside of this specific recombination volume is of little interest.
- If we look at a light emitting diode – a ***LED*** – made of **GaP** doped with **N** (in addition to the normal doping) to produce the isolectronic impurities needed to bind the excitons responsible for the radiative recombination channel, it *only* radiates from the ***p****-side* because *only* electrons become primarily bound to the isoelectronic impurity and then attract a hole. In other words, only the electron part of the injected current will contribute to radiation.
- We must confine light production to areas close to the surface as shown in the next subchapter.
- Looking ahead we will learn that many optoelectronic devices are extremely complicated heterostructures which, for several reasons, need a precise definition of the *recombination volume*.
- To optimize the current efficiency then obviously means to maximize the flow of carriers into this volume, and minimize the flow out of it.

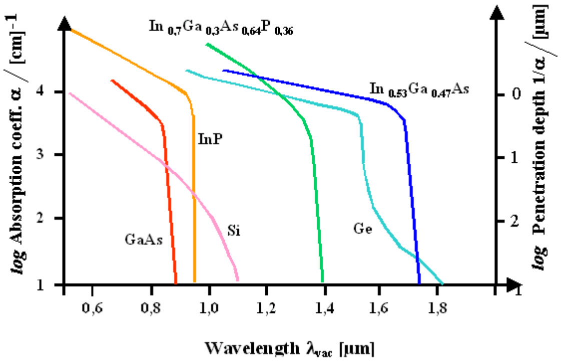
5.2.2 Absorption and Emission of Light

Absorption of Light in Semiconductors

- The optical efficiency η_{opt} is easy to understand by looking a the mechanisms that prevent photons from leaving the device. We have two basic mechanisms:
  - **1.** The photon is absorbed *before* arriving at a (possibly internal) surface of the device.
  - **2.** The photon makes it to the (internal) surface of the device, but is *reflected back* into the interior and then absorbed.
- We thus have to worry about absorption of light in semiconductors in general *and* about reflections at surfaces.
- The first topic is a science in itself. Here we only note a few of the major points:
  - In direct analogy to the various modes of radiative recombination, we have the *reverse process*, too: A photon creates an electron hole pair occupying some levels (including, e.g. exciton levels).
  - All the conservation laws must be obeyed; phonons or other third particles (in the general sense; some defects might come in handy here) may have to assist the process.
  - The dominating absorption process usually is the direct band–band process, i.e., straight up in a (reduced) band diagram from an (occupied) position in the valence band to an (unoccupied) position in the conduction band. (For indirect semiconductors this requires a larger energy than the band gap!)
- The band–band absorption process is also called the **fundamental absorption process**, it is described phenomenologically by **Beer's law**:
  - The intensity I of the light at a depth z in the semiconductor, $I(z)$, is given by

$$\boxed{I(z) \;=\; I_0 \cdot \exp(-\,\alpha \cdot z)}$$

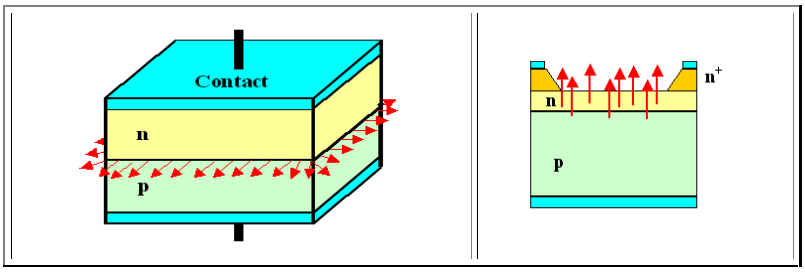
  - With I_0 = intensity at $z = 0$ and α = **absorption coefficient** of the material.
- It is clear that $\alpha = \alpha$(**h**ν) is a strong function of the energy **h**ν of the photons:
  - For **h**$\nu < E_g$**(direct)**, no electron–hole pairs can be created, the material is transparent and α is small.
  - For **h**$\nu \geq E_g$**(direct)**, absorption should be strong.
  - All mechanisms other than the fundamental absorption may add complications (e.g. "sub band gap absorption" through excitons), but usually are not very pronounced.
  - The absorption coefficients of major semiconductors indeed follow this predictions as can be seen in the following diagram:



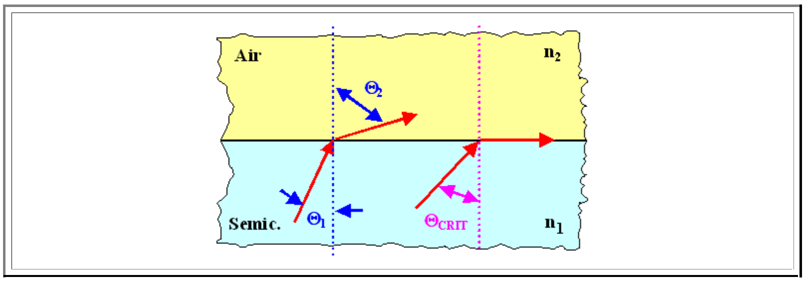
  - As expected, the absorption coefficient changes by **4 ... 5** orders of magnitude around the band edge energy, and in direct semiconductors this change is "harder" than in indirect semiconductors.
  - Note that the absorption edge of **Ge** shows the features of both an indirect and a direct transition, the latter one occurring only slightly higher in energy than the former. This is fully consistent with the band structure of germanium (cf., e.g., here).
- There are many more points to the absorption of light in semiconductors, but we will not pursue the issue further at this point.

The optical efficiency η_{opt}

- If we now look at an **LED**, we notice that light with wavelengths corresponding to the absorption edge thus will be absorbed within a few **µm** of the material – *and that automatically applies to the light emitted by radiative recombination* .
  - If we look at a naive cross section of a light emitting diode (on the left), we see that only light from the edges of the **p–n** junction has a chance to make it to the surface of the device. Obviously, this is not a good solution for a large optical efficiency.



  - If we make a junction more like in an integrated **Si** circuit (above right), the situation is somewhat improved, but it might be difficult to drive high currents in the central region of the device, far away from the contacts.
  - We might be better off in choosing an **n**-type material with a larger bandgap than the **p**-type material and see to it that light is generated in the **p**-type material. Its photon energy then would be too small for absorption in the large bandgap material and it could escape without absorption. In other words: We utilize a *heterojunction* (see below for details).
- This sequence demonstrates several important points about the realization of **LEDs**:
  - **1.** A large optical efficiency is *not* easy to achieve. Generally, much of the light produced will never leave the device.
  - **2.** The typical structures from **Si** integrated circuit technology may or may not be useful for optoelectronic applications. In general, we have to develop new approaches.
  - **3.** We always should try to produce the light close to the (possibly internal) surface of the active material. In other words, we need a defined recombination zone that is not deep in the bulk of the active material.
  - **4. Heterostructures** – meaning the combination of different semiconductors – come up quickly in optoelectronics (while virtually unknown in **Si** technology – except for high-efficiency solar cells).
- Next, let's assume that the photons make it to the surface of the device. The question now is if they are reflected back into the interior or if they can escape to the outside world.
  - This is a question that can be answered by basic optics. The relevant quantities are shown in the next picture.



  - For the light beams coming from the interior of the semiconductor to the interface (air in the picture; more generally a medium with a refractive index n_2), **Snellius' law** is valid:

$$n_1 \cdot \sin \Theta_1 \;=\; n_2 \cdot \sin \Theta_2$$

  - With n_1 **=** index of refraction of the semiconductor, n_2 **=** index of refraction of the outside (**= 1** if it is air).
- Since relevant semiconductors have rather large refractive indexes (simply given by the square root of the dielectric constant), refraction is quite severe.
  - As soon as Θ_2 reaches **90°**, light will be reflected back into the semiconductor; this happen for all angles Θ_1 larger than Θ_{crit}, the critical angle for total reflection, which is obviously given by

$$\Theta_{crit} \;=\; \arcsin \frac{n_2}{n_1}$$

  - For typical refractive indices of **3.5** (or dielectric constants ϵ_r **= 12.25**), we have Θ_{crit} **= 17°**. This is a severe limitation of η_{opt}: Assuming that radiation is produced isotropically, a cone of **17°** contains only about **2%** of the radiation!

- But the situation is even *worse* because photons within the critical angle may still become reflected – the probability is **< 1**, but not zero.
  - For the transmissivity ***T***, the fraction of light that does not get reflected, the following relation (a variant of the general Fresnel **laws** of optics) holds:

$$T = 1 - \left(\frac{n_1 \cdot \cos\Theta_1 - n_2 \cdot \cos\Theta_2}{n_1 \cdot \cos\Theta_1 + n_2 \cdot \cos\Theta_2}\right)^2 = 1 - R$$

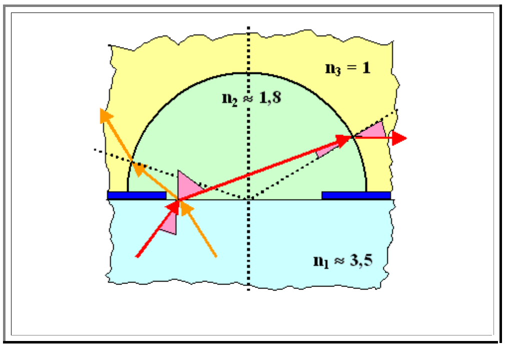
  - With ***R*** = reflectivity = {*intensity* of reflected beam} / {*intensity* of incoming beam}.

- This can be simplified to an expression for the total fraction of light leaving the semiconductor:

$$T_{total} \approx \frac{4 \cdot n_1 \cdot n_2}{(n_1 + n_2)^2}$$

  - For $n_1 \approx$ **3.5** and n_2 **= 1** (air) we have T_{total} **= 0.69**, so only about **2/3** of the radiation contained within the critical angle leaves the semiconductor.

- The total optical efficiency of **LEDs** with **isotropic** generation of radiation thus is in the **1%** region – *something we must worry about*!
  - The simplest solution is to "*grade* " the refractive index, i.e. to lower it in steps. This is most easily achieved with a "drop" of epoxy or some other polymer. How this method of **index grading** works becomes clear from the drawing:



  - Two light rays at the edge of some aperture have been traced; the relevant angles are shown as pink triangles for the red light beam. The critical angle for total reflection at both interfaces is now considerably larger. Note that the angle in the lower index medium is always larger and that this leads to a certain, not necessarily isotropic radiation characteristics of the system.
  - The polymer layer, in other words, acts as an optical system – and by giving it specific shapes we can influence the radiation characteristics to some extent.

- In total, we see that getting the light out of the device (and having it more or less focussed or otherwise influenced in its directional characteristics), is a *major part* of optoelectronic technology.
  - In fact, a totally new field of research with some bearing to these problems has recently be opened by the (first theoretical, and then experimental) "discovery" of so-called photonic crystals – activate the link for some details.

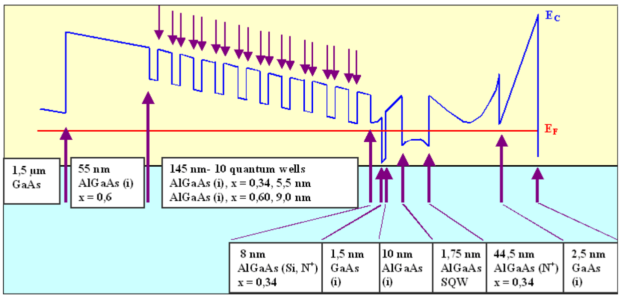
5.3 Heterojunctions

5.3.1 Ideal Heterojunctions

General Remarks

Optoelectronics, as well as practically all other devices made of compound semiconductors, always contain **heterojunctions**, i.e., junctions between two different semiconductors, for a variety of reasons.

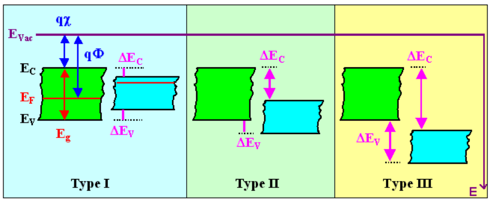
- One possible reason was already mentioned: Transparency to light generated in some active part.
- But there are many other advantages so compelling that extremely complicated heterostructures are now routinely produced despite a lot of problems that are encountered, too.
- Below, the conduction band structure of an advanced **$Al_xGa_{1-x}As$** device is shown to illustrate that point. The doping is shown in brackets; "**i**" means undoped (= intrinsic). **SQW** is short for "Single Quantum Well" – whatever that may be.
- There are many different heterojunctions, so we will not be able to delve very deep into the subject. The logic behind this structure will be made clear in a later chapter.



E_C
E_F
1,5 µm GaAs
55 nm AlGaAs (i) x = 0,6
145 nm- 10 quantum wells AlGaAs (i), x = 0,34, 5,5 nm AlGaAs (i), x = 0,60, 9,0 nm
8 nm AlGaAs (Si, N⁺) x = 0,34
1,5 nm GaAs (i)
10 nm AlGaAs (i)
1,75 nm AlGaAs SQW
44,5 nm AlGaAs (N⁺) x = 0,34
2,5 nm GaAs (i)

In this sub-chapter we will look at some major properties of heterojunctions. First of all: *How do we construct a band diagram*?

- Let's first look at the basic cases that we may encounter when considering heterojunctions. Naturally, the bandgaps are always different, but only specifying ***E*$_g$(1)** and ***E*$_g$(2)** [and of course the Fermi energy in **(1)** and **(2)**] is *not* sufficient to describe the heterosystem before the contact of the materials **(1)** and **(2)**.
- *We also must specify the exact position on the energy scale of one of the band edges*. This then gives rise to *three* distinct cases for heterojunctions as illustrated below together with the necessary definitions of the various energies needed.
- You may want to consult a special module (*in German*) dealing with some of the questions that may come up.



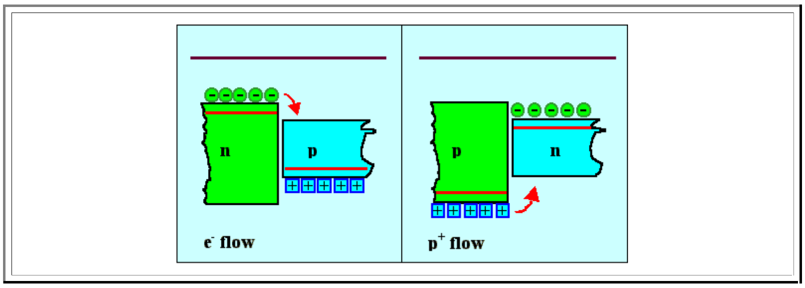
Looking at the left hand case (Type **I**), we first discuss the various energies encountered:

- In contrast to "simple" band diagrams in **Si**, the vacuum energy level is now included (and defines the zero point of the energy axis). We also have the energy of the band edges, ***E*$_C$** and ***E*$_V$**, and from their difference the bandgap energy ***E*$_g$**.
- By convention, the difference between the conduction band edge and the vacuum energy is defined as a potential called "**electron affinity** χ"; $q \cdot \chi$ is thus the work needed to move a charge from the (lower) conduction band edge to infinity.
- *Note that this is not exactly the same as the electron affinity in more chemical oriented lingo, where it is the energy gained by the reaction **X + e^- = X^-** and thus only defined by atoms that form stable negatively charged ions*.
- The difference between the Fermi energy and the vacuum energy is given by charge times the **workfunction** Θ of the material, **q** $\cdot\ \Theta$ is thus the energy needed to move a *fictitious* electron (because in semiconductors usually there is none at the Fermi energy) from the Fermi energy to infinity.
- *Note that while χ is a material constant, Θ is not – it changes with the position of the Fermi energy.*
- We may also define the differences of the band edges between the two materials, ΔE_C and ΔE_V, although they are implicitly given by the prime material parameters χ and $E_g = E_C - E_V$.

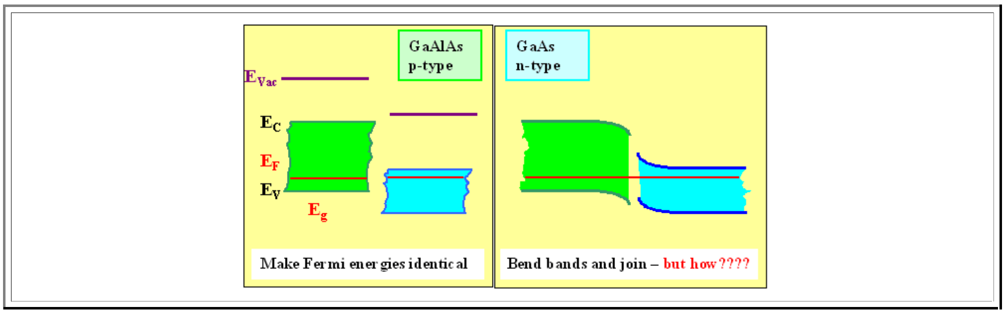
- The three types of heterojunctions possible are:
  - **Type I (straddling):** The bandgap of one semiconductor is completely contained in the bandgap of the other one; i.e. $E_C(2) > E_V(1)$ and $E_V(2) < E_V(1)$. The discontinuities of the bands are such that both types of carriers, electrons and holes, need energy (ΔE_C and ΔE_V, resp.) to change from the material with the smaller band gap to the one with the larger gap – the carriers from the other side loose this energy when they cross the junction. Type **I** heterojunctions are quite common, the important **GaAs/AlGaAs** system belongs to this kind.
  - **Type II (staggered):** The bandgaps overlap, but one $\Delta E_{C \text{ or } V}$ changes sign. The situation with respect to moving carriers from **(1)** to **(2)** or vice verse is no longer symmetrical. One kind of carrier gains energy (in the example if electrons move from right to left), the other one needs energy (the holes). The **InP/InSb** system provides an example.
  - **Type III (broken-gap):** The bandgaps do not overlap at all. The situation for carrier transfer is like type **II**, just more pronounced. The system **GaSb/InAs** belongs to this type.

Construction of Band Diagrams

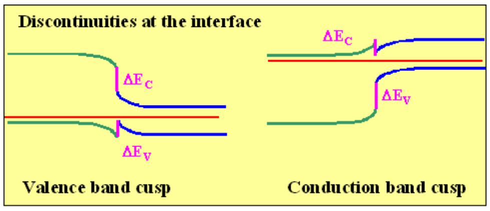
- What happens if we join the two materials? Exactly the same thing as for differently doped Si:
  - Carriers will flow across the junction, building space charges (and now possibly also interface charges) until the Fermi energy is the same everywhere in the material. Far away from the junction, everything is unchanged.
  - However, there are pronounced differences to the case of a **p–n** junction in **Si**. Let's imagine symmetrical junctions, i.e. the majority carrier density is identical in the **p-** and **n**-part. For *homojunctions* , the number of electrons flowing into the **p**-type part is then the same as the number of electrons flowing into the **n**-part. However, at least in the type **I** case, only *one kind of carrier will flow* as is obvious and shown below. The space charge regions to the left and right of the junction thus might not be symmetric.



  - Still, some kind of carrier transfer will happen and the electrostatic potential far away from the junction will rise from a constant level on one side to a different, but constant level on the other side. The difference will be equal to the difference in Fermi energy before the contact divided by the elementary charge. To the left and the right of the junction the bands are bend accordingly, and so is the vacuum energy.
- Let's see how far this recipe takes us with a simple **GaAs–AlGaAs** type **I** junction:


GaAlAs
p-type
GaAs
n-type
E_Vac
E_C
E_F
E_V
E_g
Make Fermi energies identical
Bend bands and join – but how????

  - First, we align the Fermi energies. Then, we bend the bands – in a smaller region on the more heavily doped side, but always identical for both band edges! After all, the vacuum potential at some position ***x*** is fixed and so are the band edges relative to the vacuum potential.
- This leaves us with something new: ***We cannot join the bands of the two materials!*** If we adjust the band bending on both sides so that the conduction bands match, the valence band won't match and vice verse. We must introduce a **discontinuity** right at the interface at one of the bands – or at both.
  - The next picture shows how to do that for type **I** heterojunctions for both possible doping cases.



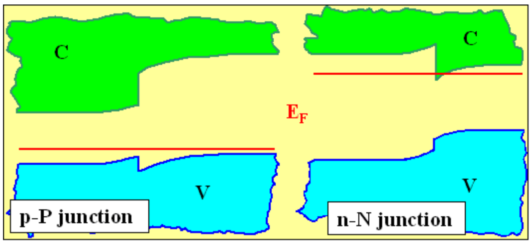
- Some kind of cusp or notch must form in the conduction or valence band, depending on the details of the system. Exactly what happens and what the cusps look like depends on many details, you must solve the Poisson equation properly for a specific case.
- Again, a discontinuity like this *must happen*, even at "ideal" interfaces. We don't know exactly what it looks like, but we can now take this potential and plug it into a one-dimensional Poisson equations and see what it means for the charge distribution.
- Unavoidably, there must be a *dipole layer* right at the interface (look at this basic module if you have problems figuring that out). The distance between the charges of this dipole layer is so small, however (atomic size) that it does not influence the carrier movement (there is an acceleration/de-acceleration process with equal magnitudes at a very small scale in classical terms or effortless **tunneling** in quantum-mechanical terms).
- This dipole layer with its sharp wiggle in the charge distribution is therefore usually not included in drawings of the heterojunction.

In reality, the matching of two lattice types with different atoms on both sides may well introduce some **interface states** in the bandgap, as discussed for the free **Si** surface.

- Depending on the Fermi energy (which is of course influenced by the interface states, too), these interface states may be charged and introduce some band bending of their own.

To make things even more complicated (for *pessimists*), or to add more possibilities for engineering with heterojunction (for *optimists*), we now can produce junctions with specific properties between materials of the *same* doping type – even for identical carrier densities.

- This type of heterojunction is sometimes called an **isotype junction** , the **p–n** type a **diode-type junction**. Some people use abbreviations, with upper case letters indicating the doping type of the material with the wider bandgap and lower case letters the other one. We then have **Pn**, **pN**, **Np**, and **nP** junctions of the *diode* type, and **Nn**, **nN**, **Pp**, and **pP** junctions of the *isotype*. If we look ahead, we can now easily denote multi-junctions like **PnP**, and so on.
- Also isotype junctions have band discontinuities at the interface, and also here the cusp is where the Fermi energy is; the next picture shows examples.



- You may already wonder what properties to expect from this kind of junction and what it is good for; we will discuss that later.

We now have a degree of freedom for all heterojunctions, which did not occur for homojunctions: *How do we distribute the discontinuity?*

- In the left hand diagrams above, e.g., we could decrease ΔE_C and increase ΔE_V by an identical amount, making the cusp more pronounced; or we do it the other way around. How can we find the real case?
- The answer is: Nobody knows how to do that in some kind of comprehensive theory. The simplest model (called the **Anderson** model) assumes that ΔE_C is equal to the difference in the electron affinities χ:

$$\Delta E_C \;=\; q \cdot [\chi(2) \;-\; \chi(1)]$$

- But that is only a rough estimate that may be quite wrong – not to mention that *bulk* electron affinities cannot be calculated with any precision, and measurements always obtain the (systematically different) *surface* electron affinities.

There is, however, one feature of the discontinuities that makes life somewhat easier: Whatever its value, it is determined by the interaction of the atoms at the interface and interatomic forces are responsible for its value.

- This simply means that its value does not change much if we change properties of the materials on a scale much larger than the atomic scale. In other words:

- The band bending necessary for adjusting the potentials on both sides of the junction, so that the Fermi energy is identical, may be seen as independent of the value of the discontinuities. If we construct a band diagram, we simply always keep the same value for the discontinuities (as determined from using the vacuum level as reference for the separate bulk materials), no matter what else we do with the bands.

Properties of Heterojunctions

How do we measure the values of the discontinuities? The answer is: Make the heterojunction and *measure* the junction properties.

- In order to do that, we need to know how the discontinuities influence measurable quantities. How, for example, does the precise nature of the discontinuities influence the current–voltage characteristic of a heterojunction? Or, if there is some radiative recombination, the quantum- or current efficiencies?
- *We are now entering deep water*. Or are we? After all, the equations for ***I–V*** characteristics of a junction (without the space charge layer part) in the simple or more complex form did not contain anything about the shape of the band bending – only the potential difference and bulk properties of materials to the left and right of the junction.
- This tells us that the basic diode characteristics (assuming that nothing happens in the space charge region) must still be valid in its general form, but with *one big difference* that transfers into a decisive property of heterojunctions:
- The hole and electron part of the total current now are *different* even for a perfectly symmetric junction!

This is most easily seen if we look at the relation between the two partial currents j_e and j_h, called the injection ratio $\kappa = j_e / j_h$. Taking the expressions from the simple diode equation given before, we obtain

$$\kappa = \frac{j_e}{j_h} = \frac{e \cdot n_e{}^p \cdot D_e / L_e}{e \cdot n_h{}^n \cdot D_h / L_h}$$

- Rewriting this in terms of the carrier mobilities and the doping densities (assuming fully ionized dopants) with the relations given before, we obtain

$$\kappa = \frac{[\mu_e \cdot n_i{}^2 / (L_e \cdot N_A)]_{\text{p-side}}}{[\mu_h \cdot n_i{}^2 / (L_h \cdot N_D)]_{\text{n-side}}}$$

- For homojunctions, the intrinsic carrier density n_i is the same on both sides, *but not for heterojunctions!* For the intrinsic carrier densities of any semiconductor we have the basic equations:

$$n^e{}_i = N^e{}_{\text{eff}} \cdot \exp\left(-\frac{E_C - E_{Fi}}{kT}\right)$$

$$n^h{}_i = N^h{}_{\text{eff}} \cdot \exp\left(-\frac{E_{Fi} - E_V}{kT}\right) = n^e{}_i$$

- With E_{Fi} = Fermi energy for the intrinsic case.

Only for the case that $N^e{}_{\text{eff}} = N^h{}_{\text{eff}}$ would the Fermi energy be in the middle of the band gap; and while we always used that approximation for **Si**, we must be more careful with compound semiconductors.

- But independent of the exact position of the Fermi energy, for total equilibrium we always have

$$n_i{}^2 = n^e{}_i \cdot n^h{}_i = N^e{}_{\text{eff}} \cdot N^h{}_{\text{eff}} \cdot \exp\left(-\frac{E_C - E_V}{kT}\right)$$

- Inserting this relation into the equation for the injection ratio κ from above, introducing the difference of the band gaps of material **1** and material **2** as $\Delta E_g = E_g(1) - E_g(2)$, and substituting "**1**" and "**2**" for "**n-side**" and "**p-side**" (because this relation is valid for *any* heterojunction of the diode type), we obtain for κ:

$$\kappa \;=\; \frac{[\,\mu_e \cdot N_{\text{eff}}{}^{e} \cdot N_{\text{eff}}{}^{h} \,/\, (L_e \cdot N_A\,)]_1}{[\,\mu_h \cdot N_{\text{eff}}{}^{e} \cdot N_{\text{eff}}{}^{h} \,/\, (L_h \cdot N_D)]_2} \cdot \exp\left(-\,\frac{\Delta E_g}{kT}\right)$$

- This is a very important equation for optoelectronics. Let's see why:
  - If ΔE_g is sufficiently large – and since it is in an exponential term, it does not have to be *very* large – it will always overwhelm the possible asymmetries in the pre-exponential term, e.g. because of different doping levels, or effective density of states between the two materials, and this means $\kappa = j_e / j_h$ is very different from **1**. Getting all signs right, we have the following situation:

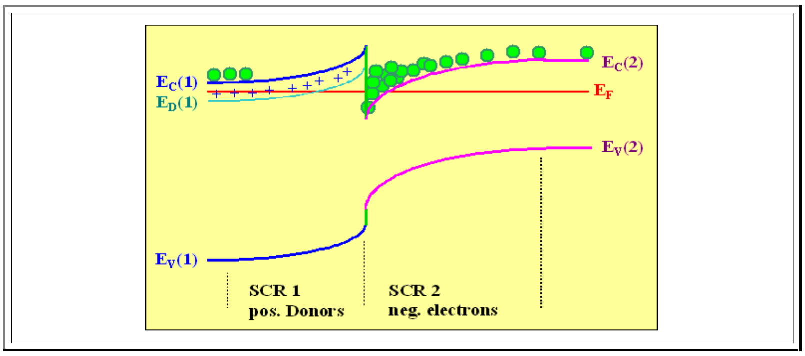
Junction type	ΔE_g	**exp** $[-\Delta\, E_g/(kT)]$	κ
Pn	>0	small	small
pN	<0	large	large

  - In other words: *In heterojunctions of the diode type, injection of the majority carriers from the material with the larger band gap (almost) always far surpasses the reverse process*.
  - To give a relevant example: For a **$GaAs/Ga_{0.7}Al_{0.3}$ As** junction with **$\Delta E_g = 0.3$ eV** and for doping densities of **10^{18} cm^{-3}** or **$2 \cdot 10^{17}$ cm^{-3}**, respectively, we have $\kappa \approx 10^6$**.**
- Why is a large value of κ so important?
  - Because if we sandwich a *small* gap semiconductor between two *large* gap semiconductors, we should be able to inject a lot of electrons from one side and a lot of holes from the other side – with no means of escape. The injected carriers *must* recombine in the small gap part, which thus is our recombination zone – we have a large current efficiency η_{cu}.
  - If you think about that a minute and try to come up with some structure, you will realize that there is a problem: You cannot have injection via *two* **p–n** junctions – one junction must be an isotype junction. But luckily, isotype junctions have similar properties: it is easy to inject majority carriers from the wide band gap side and not so easy from the other side.

5.3.2 Isotype Junctions, Modulation Doping, and Quantum Effects

Isotype Junctions

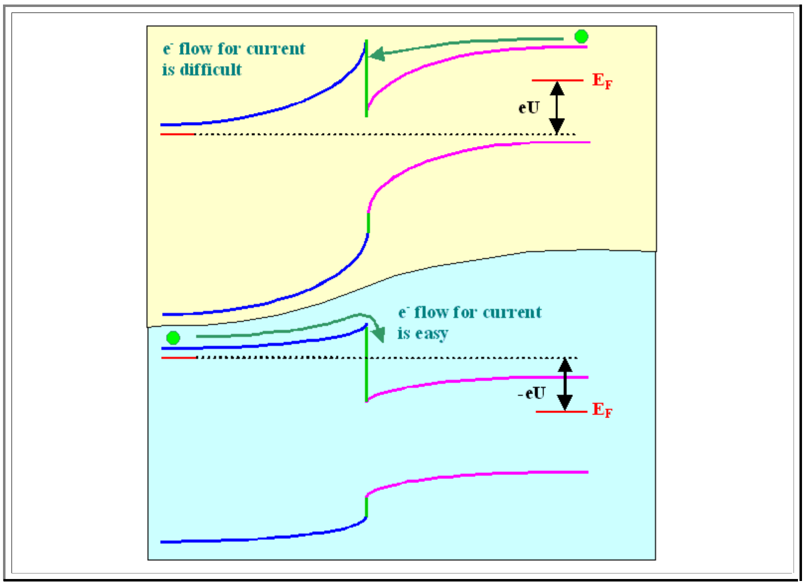
Let's look at the band diagrams of an isotype heterojunction in equilibrium; we chose a somewhat extreme case:



- Electrons were transferred form the (heavily **n**-doped) wide-gap material **1** to the (lightly **n**-doped; almost intrinsic) small gap material **2**. Space charge regions form; on the left hand side by positively charged ionized dopant atoms; on the right hand side by the increased electron density.
- In other words: In the low gap semiconductor we have carrier **accumulation** like in Si **MOS** devices.

What happens if we apply a bias ***U***?

- Well, draw the band diagrams. Move one side up by **e · *U***, and make sure that the band discontinuities do not change. What you get for the structure from above is shown below.



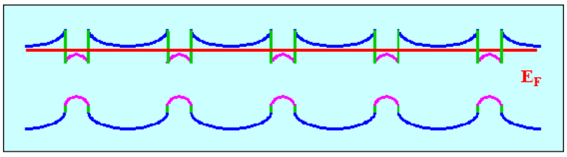
The current is always carried by majority carriers (electrons in our case). Inspecting the (exaggerated) drawings, it is clear that this is relatively easy from left to right, but *not* from right to left. Without going into the details of the characteristics, there are several novel features emerging with possible uses for devices. The first one of these effects is clear; the following ones need consideration:

- **1**. We may use an isotype heterojunction to **inject** majority carriers from the wide band gap material into the small band gap material as in the case of diode-type junction.
- **2.** We may use an isotype heterojunction to **spatially separate** the carriers generated by doping in the wide band gap material from the doped region.
- **3.** We might have peculiar **new quantum effects**.

While the first point is relatively clear, including its usability for light emitting devices (again, try to figure this out yourself), the second and third point need some explaining.

Modulation Doping

- Let's look at the isotype junction in equilibrium *again* to understand the second point.
  - What happened is essentially that the electrons missing in the space charge region on the left hand side were transferred to the potential dip on the right hand side. Of course, electrons are also running down the slope from the right, but the essential contribution is from the wide band gap material on the left (which, after all, is the cause for the dip).
  - Since we picked a highly doped material **1** and a lightly doped material **2**, we now have **a lot** of electrons (their density is essentially given by the doping in material **1**) in a crystal with **few** ionized dopant atoms.
  - And what this means is that we now have a *high density of highly mobile electrons*, because the mobility at high doping density is always severely decreased by scattering at the ionized dopants – cf. the paragraph to this topic. This effect is most pronounced at low temperatures and can lead to a mobility enhancement of an order of magnitude or even more.
  - This is not only generally useful, but can be carried to extremes. All we have to do is to make *sandwiches* as shown below.



  - With properly chosen dimensions, deep potential wells will form in the low band gap material that contain most of the electrons from the highly doped wide band gap material. This amounts to a novel way of effectively doping material **2**, called **modulation doping**.
- If the potential wells are small enough (which is usually the case), the confinement of the electrons in the wells leads to pronounced quantum effects – we therefore call these potential wells "**quantum wells**" (***QW***) and distinguish **single quantum wells** (***SQW***) and **multiple quantum wells** (***MQW***).
  - An **SQW** is obtained by sandwiching just one small gap semiconductor, a **MQW** as shown above. The introductory picture of the heterojunction subchapter showed examples of both types.
  - We may even improve on that by inserting a very thin layer (**1 . . . 10 nm**, say) of intrinsic material of a suitable band gap between the two basic materials. If properly done, this layer, while not impeding carrier flow into the potential wells for equilibration, keeps the carriers from being scattered at the interface and thus increases the mobility even more.
- But with that we have not yet exhausted the possibilities of heterojunctions – we will now turn to special quantum effects.

Quantum Confinement Effects

- Let's consider the peculiar quantum effects in modulation doped structures by looking at some typical dimensions.
  - The width of the various space charge layers must still be given by formulas not too different from the ones we had for Si. For a **GaAlAs/GaAs** system with a high doping around **10^{18} cm^{-3}** in the wide band gap **GaAlAs** side, the width of the dips with the high electron density on the **GaAs** side is about **5 ... 10 nm**, while the lateral extension is large by comparison.
  - The mean free path length of the (highly mobile) electrons is larger than the thickness of the potential dip (better called potential well for the multi-junction configuration shown above) and this means that we now have essentially a *two-dimensional* electron gas.
- What does that mean? Especially if we make the thickness of the layers extremely thin?
  - It means that we have a periodic crystal in two dimensions (***x*** and ***y***) and a one-dimensional potential well in the ***z***-direction, which is always the direction used in the pictures above.
  - The relevant Schrödinger equation is easy to write down, especially in the free electron approximation with a constant potential (**= 0**) in ***x***- and ***y***-, and a potential ***V*(*z*)** in ***z***-direction:

$$-\frac{\hbar^2}{2}\left(\frac{1}{m_x^*}\cdot\frac{\partial^2}{\partial x^2}+\frac{1}{m_y^*}\cdot\frac{\partial^2}{\partial y^2}+\frac{1}{m_z^*}\cdot\frac{\partial^2}{\partial z^2}\right)\psi(r) - e\cdot V(z)\cdot\psi(r) = E\cdot\psi(r)$$

  - This equation is solved by

$$\psi(\boldsymbol{r}) \;=\; \psi_{\text{vert}}(z) \cdot \psi_{\text{lateral}}(x,y)$$

- The two functions ψ $_{\text{lateral}}(x,y)$ and $\psi_{\text{vert}}(z)$ are decoupled, the solutions can be obtained separately. For the lateral part we simple have

$$\psi_{\text{lateral}}(x,y) \;=\; \text{Solutions of the two-dimensional free electron gas problem}$$

- The vertical part of the solution comes frome solving the remaining one-dimensional Schrödinger equation

$$\left(-\frac{\hbar^2}{2m_z^*} \cdot \frac{\partial^2}{\partial z^2} - e \cdot V(z) \right) \psi_{\text{vert}}(z) \;=\; E_{\text{vert}} \cdot \psi_{\text{vert}}(z)$$

It is rather clear that the structure of the two-dimensional problem will not be much different from that of the common three dimenional problem if we introduce a periodic potential in (**x,y**). We simply obtain Bloch waves in two dimensions instead of plane waves for ψ **lateral(x,y)**. The energy eigenvalues are unchanged, too, they were for the free electron gas (using the effective masses by now).

$$E_{\text{lateral}} \;=\; \frac{\hbar^2 k^2_{\text{lat}}}{2m^*_{\text{lat}}}$$

- The solution of the one-dimensional problem in ***z***-direction depends of course on the precise shape of ***V(z)***, but as a general feature of potential wells we must expect a *sequence of discrete energy levels*. For the most simple case of a rectangular well (with infinite height), standard calculations show that

$$E_{\text{vert}} \;=\; \frac{(\hbar\pi)^2}{2m_z^*} \cdot \frac{j^2}{d_z^2}$$

- With ***j* = 1, 2, 3, ... =** quantum number, and d_z **=** thickness of the potential well.

The total energy of an electron is now given by

$$E_{\text{total}} \;=\; E_{\text{lat}} + E_{\text{vert}}$$

$$E_{\text{total}} \;=\; \frac{(\hbar k_{\text{lat}})^2}{2m^*_{\text{lat}}} + \frac{(\hbar\pi)^2}{2m^*_z} \cdot \frac{j^2}{d_z^2}$$

- This "simply" means that the states in the conduction bands are now a *discrete series* given by the quantum number ***j*** with a density of states per level of

$$D_{\text{lat}} \;=\; \frac{\hbar \cdot m^*_{\text{lat}}}{2\pi} \;=\; \text{constant}$$

- If you like to try your hand at a little math: The formula for D_{lat} is rather easy to obtain if you follow the recipe for the three-dimensional ***DOS*** for this case.

What do we get from this? Well, a lot of special effects for enthusiastic solid state physicists, but also some (not necessarily big) advantages for devices. *However.....*

- Each **quantum well** *layer* is now something like a *one-dimensional* atom – in contrast to the three-dimensional real atoms were the wave functions of the electrons were confined in all three directions. If we move these "atoms" close together in the ***z***-direction, there must be a point where the wavefunctions in ***z***-direction (the $\psi_{\text{vert}}(z)$) start to overlap and do all the things *real atoms do at close distance*.

- The energy levels change and split, and – in analogy to a crystal formed by real atoms – a one-dimensional energy band may start to develop with an *energy range that is given by the geometry of the system*, i.e., the thickness of the layers, for a **multi quantum well structure**.

This is a momentous statement! Think about it!

It means that we can make materials with energetic properties that we can tailor *at will* (within bonds and limits, of course). We no longer must just live with bandgaps and other properties that mother nature provides, *we now can make our own systems*. At least in principle.

- Something like that we call a **metamaterial**.
- Multiple and single quantum wells are already part of recent devices as shown in the last module of this chapter.

5.3.3 Real Heterojunctions

Misfit Dislocations and Critical Thickness

So far we considered *ideal* heterojunctions. What do we mean with ideal? You can look at it in two ways

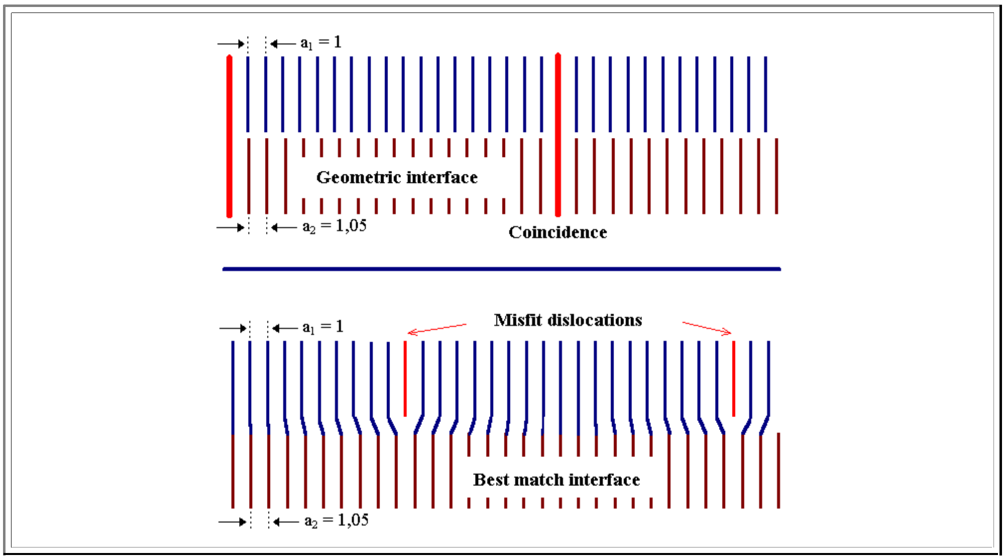
- The junction is *structurally* ideal, i.e. you just switch from one set of atoms on one side of the junction to another set on the other side. For that you need the same type of crystal lattice and identical, or at least very similar lattice constant, of course.
- The junction is *electronically* ideal, i.e. the interface does not have any interface states in the band gap (in analogy to the case of a free surface treated before) or is otherwise interfering with carrier densities and transport.

But even for these ideal conditions we have an *energy discontinuity* at the interface with a charged dipole layer if the badgap energies are different. What happens for *real* interfaces, the only ones we can actually make?

- *Real* interfaces have one thing in common: The lattice constants of the two materials joined at the interface are *never* precisely identical. And from this fact of life evolve many problems – and many ingenious technologies to avoid those problems.
- The basic problem is the same for all heterojunctions. The **lattice misfit** between the two crystal may cause the incorporation of a network of so-called **misfit dislocations** into the interface. And this misfit dislocation network is the source of practically all evil in heterojunctions – if you have it, your device will not work at all, will work only badly, or fail after some (too short) time.
- Compared to the "high physics" part of the electronic and quantum properties of heterojunctions, this looks like a mundane problem. Well, it is – but it is here where most grandiose ideas for stunning devices go down the drain. If you can not make the junction, you won't get far with your device.

If you are especially interested in this topic, or if you only have a very dim perception of lattice defects in general and dislocations in particular, you should now turn to the hyperscript "*Defects in Crystals*" either in general, or to the chapters "Dislocations" or "Phase boundaries" in particular; here we only will deal with the very basics of misfit dislocations.

- The following figure shows what misfit dislocations are and why they are formed.



If you just geometrically juxtapose two crystals, you will have a situation as shown in the upper part for a misfit of **5%**.

- Only every **20th** lattice point will precisely match between the two lattices (at so-called coincidence sites). In between, the situation does not only look highly instable, but really *is* unstable. If there is any appreciable interaction between the atoms of lattice **1** and lattice **2**, something will happen and that is almost always the case (mother nature, of course, does provide some exotic crystals with "geometric" interfaces as the exception).
- Other weird solutions are conceivable, e.g. an amorphous layer between the two crystals, some highly disordered region formed by a mixture of the two lattices – you name it. While all of this does happen on occasion, it is not the rule; certainly not for "normal" semiconductors.

Eschewing "geometric" and "weird" interfaces, there are only two reasonable options left:

- **1.** The lattices are elastically squeezed or expanded until they fit precisely. The amount of energy contained in the necessary elastic distortions is directly proportional to the volume of the deformed material; for the one-dimensional structures we are usually envisioning, the energy scales with the thickness of the layers.

- **2.** Misfit dislocations are introduced as shown above. This means that all the misfit is concentrated in a small volume around the dislocations, while in between we have a perfect fit with only a little elastic distortion. The total energy contained in the distortion around the dislocations is rather large, but does not depend much on the volume (resp. thickness) of the crystals.

- As a simple and sad consequence we then have the following basic fact of life:
  - For layer thicknesses *larger* than some system-dependent critical thickness $\boldsymbol{d_{crit}}$, the introduction of a misfit dislocation network is *always* energetically favorable.
  - Deriving a formula for the critical thickness is not without problems and some material specific idiosyncrasies, but in general we have

$$\boxed{d_{crit} \;=\; \frac{b}{8\cdot\pi\cdot f\cdot(1\;+\;\nu\,)}\cdot\ln\frac{e\cdot d_{crit}}{r_0}}$$

  - With b = **Burgers vector** of the dislocations; usually somewhat smaller than a lattice constant, f = misfit parameter = $(a_1 - a_2)/a_1$, ν = Poisson ration ≈ **0.4**, **e** = base of natural logarithms **= 2.718...**, r_0 = inner core radius of the dislocation; again in the order of lattice constant.

- Getting precise values of $\boldsymbol{d_{crit}}$ is such not easy (not to mention that the equation above has no analytical solution); but for a crude approximation that can be used for "normal" cases we simply have

$$\boxed{d_{crit} \;=\; \frac{b}{10\cdot f}}$$

  - More about that can be found in an advanced module.

- If, for example, we look at the system **GaAs/InAs** , we have lattice constants of **0.565** and **0.606** nm, so $\boldsymbol{f = 0.0726}$ (i.e. the misfit is **7.2 %**). The Burgers vectors in these crystals are usually $\boldsymbol{a/(2^{1/2})}$ ≈ **0.42nm**, which gives us a critical layer thickness of $\boldsymbol{d_{crit} = 0.58}$ **nm** – less than **2** crystal layers. [1]
  - *Shit* (really)! This looks not so good – and in fact, nobody uses the **GaAs/InAs** system for heterojunctions. But we have better couples, especially mother natures gift to optoelectronics, the **GaAs/AlAs** system and the **InAs/GaSb/AlSb** system where the misfit parameters are much smaller.
  - If we go through the numbers for **GaAs/AlAs** with **0.5653/0.5660**, we obtain $\boldsymbol{d_{\ crit} = 57}$ **nm** – a value sufficient for many applications.

- While this is nice, we must of course ask ourselves if there are ways to beat the $\boldsymbol{d_{crit}}$ equation, i.e., to produce layers with a thickness larger than the critical thickness. This is indeed the case and we will look at some of the methods to produce dislocation free heterojunctions despite the energetic limitations. More about misfit dislocations (and other problems in heterojunctions) can be found in an advanced module.

Extending the Critical Thickness

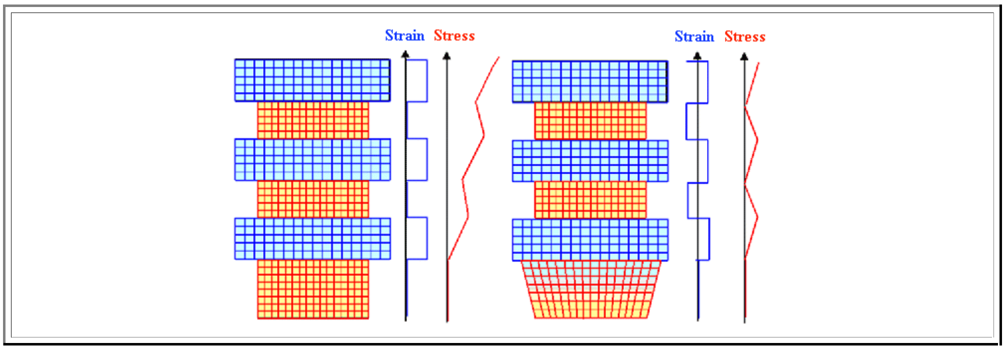
- There are some ways to beat the critical thickness to a smaller or larger extent; we will just give them a cursory glance which will not do justice to the sweat and toil as well as hard thinking that went into this problem.

- *First* , do not believe the **theory** and give up because it looks bad.
  - Even the full equation from above does not take all parameters into account. The situation may be better (or worse) than the numbers you obtain.
  - Try it experimentally, at least as the layer thickness you need is not too far (at least a factor **3** or more) above the theoretical limit. You might be lucky!
  - However, don't try for really large misfits above, say, **2 %**. Not only is the critical thickness small, but you probably will not even be able to obtain a smooth layer – islands will grow!

- *Second*, consider the **kinetics** of the layer deposition process.
  - Any formula for $\boldsymbol{d_{crit}}$ (including much more advanced treatments) is an equilibrium formula, comparing enthalpies in equilibrium.
  - However, since your layer thickness always is below $\boldsymbol{d_{crit}}$ at the beginning of the deposition process, there are no misfit dislocations in the beginning of the deposition. After the critical thickness is reached, dislocations must be nucleated and move from the surface to the interface and this is a kinetic process which you may be able to impede.

  - In other words, for optimized conditions, you may obtain dislocation free interfaces for kinetic reasons. In particular, make the nucleation of dislocations difficult by avoiding all irregularities (including temperature gradients) that may serve as nuclei.

- *Third*, minimize the elastic strain energy by using a **buffer layer**.
  - This is maybe the most important trick; especially if you want to produce many junctions for multiple quantum wells.
  - Lets look, e.g., at a **MQW** sequence consisting of the substrate material (yellow) and a material with a larger lattice constant (blue) very schematically before the "joining" of the crystals. Keep in mind that the substrate, being very thick, never "gives" – only the layers will be strained!
  - Even if the first blue layer is below the critical thickness, the stress will build up, and after a few layers you have misfit dislocations for sure (left part of the figure).
  - The break-through came with the introduction of a buffer layer in which the lattice constant is gradually changed (by gradually changing the composition) to a value halfway in between material **1** and material **2**. This is shown in the right part of the figure. The effect is that while the stress in the layers is about the same as before, it does not build up anymore with the number of layers if everything is done just right – *multiple quantum wells are possible!*



  - In reality, the buffer layer is much thicker so it cannot be strained very much as shown.

- Buffer layers of some mysterious kind, it seems, also finally helped to obtain the holy grail of heteroepitaxy: Growing **GaAs** on **Si** without misfit dislocations.
  - Not an easy task, if you consider that the misfit is about **4.1 %.** Still, **Motorola** appears to have solved the problem, if you can believe newspaper articles. The "appears" relates to what you actually read in one of Germanys finest daily; the article is contained in the link – click on it and try if you can make some sense out of it (provided you understand German). Otherwise try this link.

- *Fourth*, accept the misfit dislocations, but put them in a part of the system where they **do no harm**.
  - This approach is known under the heading "**compliant substrates**". The basic idea is simple (and illustrated in an advanced module): Make a bicrystal (usually of **Si**) by bonding two wafers together with a defined twist of up to **15°** along the axis perpendicular to the wafer. A grain boundary ("small-angle twist boundary") must form, consisting of a dense array of screw dislocations.
  - Now polish off one of the **Si** wafers until only a thin layer (**1 µm** or less) remains. This does not only sound difficult to do, it really is – but nevertheless it can be done in a large scale production. The remaining sandwich, thick **Si** / grain boundary / thin **Si**, is your compliant substrate.
  - If you now deposit a layer of anything on thin **Si** layer, any misfit (up to very large amounts) between the thin **Si** layer and the deposited layer of the other material will be accommodated by the dislocations in the grain boundary – there is no need to form new ones.
  - The important interface thus remains dislocation free and you may now be able to do things not possible so far.

[1] Note that in contrast to the elemental diamond lattice, where the smallest possible Burgers vector for a perfect dislocation is $\boldsymbol{b}_{\mathbf{elem}} = \boldsymbol{a}/(2 \cdot 2^{1/2})$, we have $\boldsymbol{b}_{\mathbf{comp}} = \boldsymbol{a}/(2^{1/2})$ because we would otherwise replace **A**-atoms by **B**-atoms in the glide plane of the dislocations (look at the Volterra "cut and paste" definition of a dislocation in the "**Defects in Crystals**" hyperscript).

5.4 Quantum Devices

5.4.1 Single and Multiple Quantum Wells

Energy Levels in a Single Quantum Well

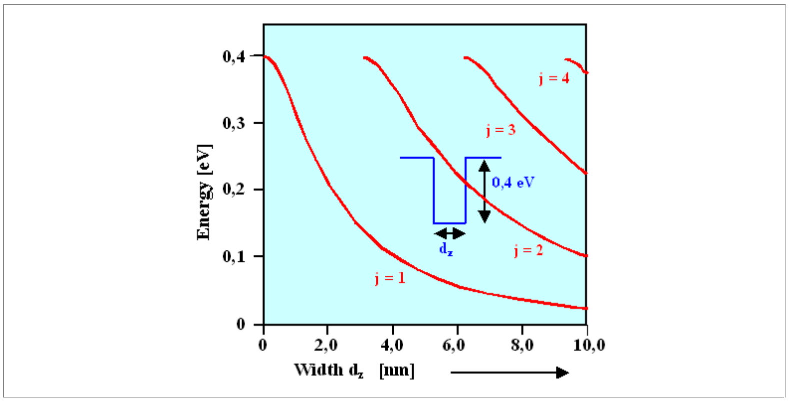
- Let's first look at an ideal **single quantum well** (**SQW**), rectangular and with an extension $\boldsymbol{d_z}$ and infinite walls (the index "**z**" serves to remind us that we always have a three-dimensional system with the one-dimensional quantum structures along the ***z***-axis).
  - We have already solved the Schrödinger equation for this problem: It is nothing else but the one-dimensional free electron gas with $\boldsymbol{d_z}$ instead of the length ***L*** of the crystal used before.
  - We thus can take over the solutions for the energy levels; but being much wiser now, we use the effective mass instead of the real mass for the electrons and obtain

$$E = \frac{(\hbar \cdot k_z)^2}{2m_z^*}$$

  - With $\boldsymbol{k_z = \pm n_z \cdot 2\,\pi/d_z}$, and **n = ± (0,1,2,3,...)**.
  - We have used *periodic boundary conditions* for this case, which is physically sensible for large crystals. The wave functions are propagating plane waves in this case. It is, however, more common and sensible to use fixed boundary conditions, especially for small dimensions. The wave functions then are standing waves. Both boundary conditions produce identical results for energies, density of states and so on, but the set of wave vectors and quantum numbers are different; we have
  - $\boldsymbol{k_z = j_z \cdot \pi/d_z}$, and $\boldsymbol{j}$ **= 1,2,3,..** (we use ***j*** as quantum number to indicate a change in the system). For the energy levels in a single quantum well we now have the somewhat modified formula

$$E = \frac{\hbar^2 \cdot \pi^2}{2m_z^*} \cdot \frac{j^2}{d_z^2}$$

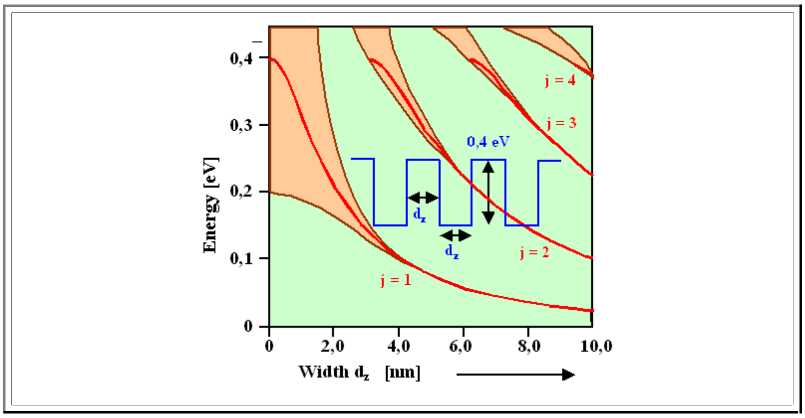
- The absolute value of the energy levels and the spacing in between increases with decreasing width of the **SQW**, i.e. with decreasing thickness $\boldsymbol{d_z}$ of the small band gap semiconductor sandwiched between the two large band gap semiconductors.
  - Large differences in energy levels might be useful for producing light with interesting wave lengths. In infinitely deep ideal **SQWs** this is not a problem, but what do we get for real **SQWs** with a depth below **1 eV**? This needs more involved calculations, the result is shown in the following figure.



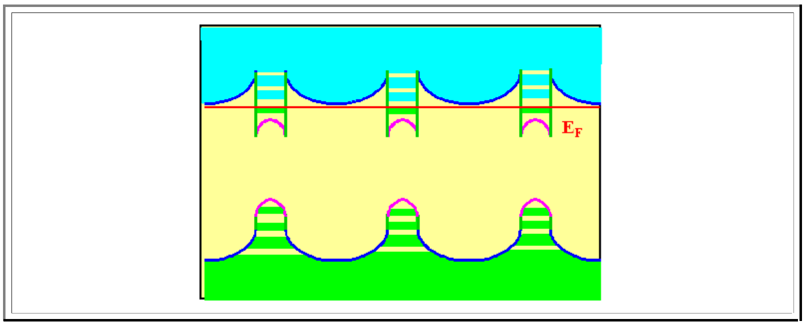
  - The **SQW** has a depth of **0.4 eV**; if it disappears for $\boldsymbol{d_z = 0}$, we simply have a constant energy of **0.4 eV** for the ground state; all excited states stop at that level.
- For layer thicknesses in the **nm** region (which is technically accessible) energy differences of **0.2 . . . 0.3 eV** are possible, which are certainly interesting, but not so much for direct technical use.
  - While **SQWs** are relatively easy to produce and provide a wealth of properties for research (and applications), we will now turn to *multiple* quantum wells obtainable by periodic stacking of different semiconductors as shown before.

Energy Bands in Multiple Quantum Wells

- Since single atoms may also be described as **SQWs** (for one electron you just have the hydrogen atom type with a Coulomb potential), we must expect that the wave function of the electrons start to overlap as soon as the single **SQWs** in the **MQW** structure are close enough.
  - The situation is completely analogous to the qualitative formation of a crystal with a periodic potential from atoms. The discrete energy levels must split into many level, organized in bands.
  - This is exactly what happened; it was **Leo Esaki** who first did the calculations (and more) for this case for which he was awarded the Nobel prize.
  - Her is the result for the same system shown above.



  - Energy bands with respectable band gaps develop indeed, and we now should redraw the band diagram from before to include this fact:



  - We have "mini-bands" in the quantum wells (for symmetry reasons also for the holes in the quantum well along the valence band); green denotes occupied levels; blue empty levels.
- What do we gain by this (besides a Nobel prize)?
  - *A structure that is used in commercially sold LASER diodes!* In other words, quite crucial parts of the information technology rely on **MQWs**.
  - We will come back to this issue in the context of semiconductor lasers.

6. Principles of the Semiconductor Laser

6. Principles of the Semiconductor Laser

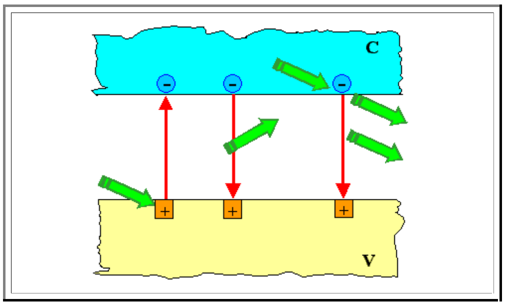
6.1 Laser Conditions

6.1.1 Interaction of Light and Electrons; Inversion

- In principle, anything that emits electromagnetic radiation can be turned into a "**LASER**", but what *is* a laser?
  - The word "*LASER*" was (and of course still is) an ***acronym***, it stands for " **L**ight **A**mplification by **S** timulated **E**mission of **R** adiation". By now, however, it is generally perceived as a standard word *in any language*, meaning something that is more than the acronym suggests (and we will therefore no longer write it with capital letters)!
  - A laser in the direct meaning of the acronym is a black box that emits (= outputs) more light of the same frequency than what you shine (= input) on it – that is the *amplifier* part. But the "*stimulated emission*" part, besides being the reason for amplification, has a second, indirect meaning, too: The light emitted is exactly in phase (or coherent to) the light in the input. Unfortunately, lasers in this broad sense do not really exist. Real lasers only amplify light with a very specific frequency – it's like electronic amplifiers for *one frequency only*.
- A laser in the *general* meaning of the acronym thus produces intense monochromatic electromagnetic radiation in the wavelength region of light (including infrared and a little ultra-violet; there are no sharp definitions) that is coherent to the (monochromatic) input. If you "input" light containing all kinds of frequencies, only one frequency becomes amplified.
  - A laser in the *specific* meaning of everyday usage of the word, however, is more special. It is a **device** that produces a coherent beam of monochromatic light *in one direction only* and, at least for semiconductor lasers, *without some input light* (but with a "battery" or power source hooked up to it). It is akin to an electronic oscillator that works by internally feeding back parts of the output of an amplifier to the input for a certain frequency.
  - Before the advent of hardware lasers in the sixties, there were already "**masers** " – just take the "**m**" for "microwave" and you know what it is.
  - And even before that, there was the basic insight or idea behind masers and lasers, and – as ever so often – it was **A. Einstein** who described the "**S**timulated **E** mission" part in **1917/1924**. More to the history of lasers can be found in an advanced module.
- Obviously, for understanding lasers, we have to consider *stimulated emission* first, and then we must look at some *feedback* mechanism.

Stimulated Emission of Radiation

- Understanding stimulated emission is relatively easy; all we have to do is to introduce one more process for the interaction between light and electrons and holes. So far we considered two basic processes, to which now a third one must be added:
  - **1. Fundamental absorption**, i.e., the interaction of a photon with an electron in the *valence band* resulting in a electron(C)–hole(V) pair.
  - **2. Spontaneous emission** of a photon by the (spontaneous and direct) recombination of an electron–hole pair.
  - **3. Stimulated emission**, as the third and new process, is simply the interaction of a photon with an electron in the *conduction band*, **forcing** recombination and thus the emission of a second photon, being an exact duplicate of the incoming one.
- All three processes are schematically shown in the band diagram below.



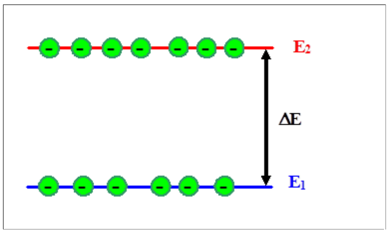
- Looking at this picture, you should wonder why one obvious further process is missing: How about an electron in the conduction band simply absorbing a photon? The electron could be moved up by the amount **hν** in the conduction band, and would come back to the band edge by tranferring its surplus energy to phonons.
  - This process does take place, but it is not very strong if we do not have many electrons in the conduction band. More importantly: It is not necessary for "lasing", but rather detrimental – we will cover it later.

- Stimulated emission, however, is *not* just the reverse of absorption. Again: Usually, photons interact with electrons in the conduction band by *transferring their energy to the electron*, which moves the electron to some higher energy level in the band (or to the next band, or, if the photons are very energetic (meaning **X**-rays), even out of the crystal) – which means that the photons are *absorbed* .
  - On the contrary, stimulated emission is a *resonant process*; it only works if the photons have exactly the right energy, corresponding to the energy that is *released* if the electron makes a transition to some allowed lower level. Then, the two photons are *exactly in phase* with each other (and propagate in the same direction). For semiconductors, this energy is pretty much the band gap energy, because all conduction band electrons are sitting at the conduction band edge (more precisely, within some small energy interval above E_C , of course), and the only available lower energy level are the free positions (i.e., occupied by holes) at the valence band edge.
  - Stimulated emission thus may be seen as a competing process to the fundamental band–band absorption process described before. But while *all* photons with an energy $h\nu > E_g$ may cause fundamental absorption, because there are many unoccupied levels above E_g, *only* photons with $h\nu = E_g$ (plus some small ΔE, possibly) may cause stimulated emission.
- Einstein showed that under "normal" conditions (meaning conditions not too far from thermal equilibrium), *fundamental absorption by far exceeds stimulated emission*. Of course, Einstein did not show that for semiconductors, but for systems with well-defined energy levels – atoms, molecules, whatever.
  - However, for the special case that a sufficiently large number of electrons occupies an excited energy state (which is called **inversion**), stimulated emission may dominate the electron–photon interaction processes. Then *two* photons of identical energy and being exactly in phase come out of the system for *one* photon going into the system.
  - The kind of *inversion* we are discussing here should not be mixed up with the *inversion* that turns **n**-type Si into **p**-type (or vice versa) that we encounterd before. Same word, but different phenomena!
  - These two photons may cause more stimulated emission – yielding **4**, **8**, **16**, ... photons, i.e. an avalanche of photons will be produced until the excited electron states are sufficiently depopulated.
  - In other words: One photon $h\nu$ impinging on a material that is in a state of *inversion* (with the right energy difference $h\nu$ between the excited state and the ground state) may, by stimulated emission, cause a lot of photons to come out of the material. Moreover, these photons are all in phase, i.e. we have now a strong and coherent beam of light – amplification of light occurred!
- We are now stuck with two basic questions:
  - **1.** What exactly do we mean with "inversion", particularly with respect to semiconductors?
  - **2.** How do we reach a state of "inversion" in semiconductors?
- Let's look at these questions separately!

Obtaining Inversion in Semiconductors

- If you shine **10** input photons on a crystal, **6** of which disappear by fundamental absorption, leaving **4** for stimulated emission, you now have **8** output photons. In the next round you have $2 \cdot (8 \cdot 0.4) = 6.4$ and pretty soon you have none.
  - Now, if you reverse the fractions, you will get **12** photons in the first round, $2 \cdot (12 \cdot 0.6) = 14.4$ the next round – you get the idea.
  - In other words, the *coherent* amplification of the input light only occurs for a *specific condition*:
- There must be *more* stimulated emission processes than fundamental absorption processes if we shine light with $E = h\nu = E_g$ on a direct semiconductor – this condition defines " *inversion* " in the sense that we are going to use it.
  - Note that the light produced by spontaneous recombination of the electron–hole pairs, generated by fundamental absorption, is not coherent to the input and does not count!
  - We only look at *direct* semiconductors, because radiative recombination is always unlikely in indirect semiconductors, and while stimulated emission is generally possible, it also needs to be assisted by phonons and thus is unlikely, too.
- We will find a rather simple relation for the dominance of stimulated emission, but it is not all that easy to derive. Here we will take a "*shortcut* ", leaving a more detailed derivation to an advanced module.

- Let's first consider some basic situations for inversion in full generality. For the most simple system, we might have two energy levels ***E*** $_1$ and ***E***$_2$ for atoms (take any atom), the lower one (***E*** $_1$) mostly occupied by electrons, the upper one (***E***$_2$) relatively empty. *Inversion* then means that the number of electrons on the upper level, ***n***$_2$, is larger or at least equal to ***n***$_1$.
- In equilibrium, however, we would simply have



Two-level system in inversion condition

$$\frac{n_2}{n_1} = \frac{D_2}{D_1} \cdot \exp\left(-\frac{\Delta E}{kT}\right)$$

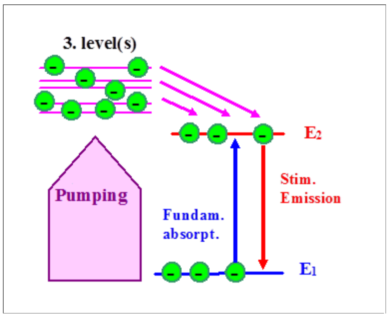
- Here, Δ***E* = *E***$_2$ **–** ***E*** $_1$, and ***D***$_{1,2}$ **=** the maximum number of electrons permitted on ***E***$_{1,2}$ (the "density of states").
- In words: In equilibrium we have far more electrons at ***E***$_1$ than at ***E***$_2$.

For inversion to occur, we therefore must be *very far from equilibrium* if Δ ***E*** is in the order of **1 eV** as needed for visible light.

- However, stimulated emission would quickly depopulate the ***E***$_2$ levels, while fundamental absorption would kick some electrons back. Nevertheless, after some (short) time we would be back to equilibrium.
- To keep stimulated emission going, we must move electrons from ***E***$_1$ to ***E***$_2$ *by some outside energy source*. Doing this with some other light source providing photons of the only usable energy Δ***E*** would not only defeat the purpose of the game (since, after all, that is the light we want to generate) – it also would never bring us back to inversion because of the depopulation of ***E***$_2$ by stimulated emission.
- *In short*: Two-level systems are no good for practical uses of stimulated emission.
- In semiconductors we could inject electrons from some other part of the device, but a semiconductor is not a two-level system, so that is not possible.

What we need is an *easy* way to move a lot of electrons to the energy level ***E***$_2$ *without* depopulating it at the same time. This can be achieved in a **three-level system** as shown below (and this was the way it was done with the first ruby laser).

- The essential trick is to have a whole system of levels – ideally a band – *above* ***E***$_2$, from which the electrons can descend efficiently to our single level ***E***$_2$ – but not easily back to ***E***$_1$ where they came from. Schematically, this looks like the figure on the right.
- The advantage is obvious. We now can use light with a whole range of energies – always larger than Δ***E*** – to "*pump*" (yes, this is the standard word used for this process) electrons up to ***E***$_2$ via the reservoir provided by the third level(s).
- The only disadvantage is that we have to take the electrons from ***E***$_1$. And no matter how hard we pump, the effectiveness of the pumping depends on the probability that a quantum of the energy we pour into the system by pumping will actually find an electron to act upon. And this will always be proportional to the number (or density) of electrons available to be kicked upwards. In the three-level system this is at most ***D***$_1$. However, if we sustain the inversion, it is at most **0.5 ·** ***D***$_1$, because by definition we then have at least one-half of the available electrons on ***E***$_2$.

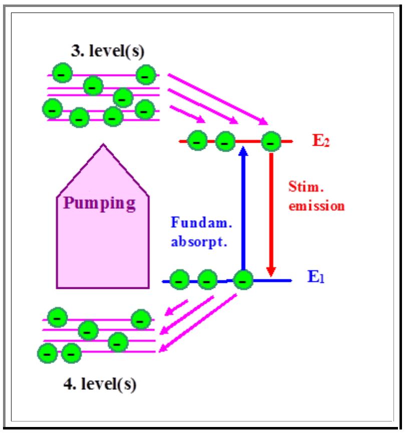


It is clear what we have to do: Provide a *fourth level* (even better: a band of levels) *below* ***E***$_1$, where you have a lot of electrons that can be kicked up to ***E*** $_2$ via the third level(s). It is clear that we are talking semiconductors now, but let's first see the basic system:

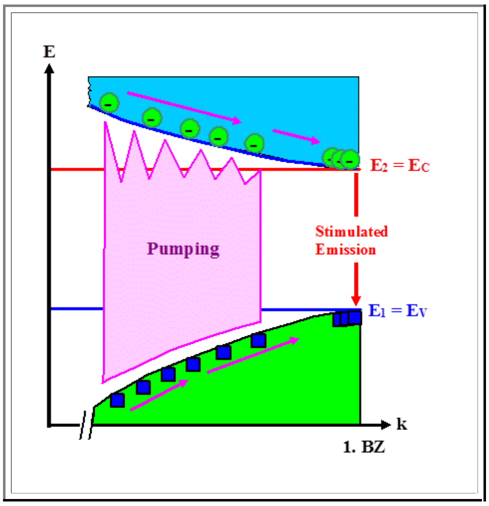
- We simply introduce a system of energy states below E_1 in the picture from above. We now have a large reservoir to pump from, and a large reservoir to pump to.

- All we have to do is to make sure that pumping is a one-way road, i.e. that there are no (or very few) transitions from the levels **3** to levels **4**.

- This is not so easy to achieve with atoms or molecules, but, as you should have perceived by now, this is exactly the situation that we have in many direct band gap semiconductors. All we have to do to see this is to redraw the **4**-level diagram at the right as a band diagram. To include additional information, we do this in *k-space*.



We have the following general situation for producing inversion in semiconductors by optical pumping:



Electrons may be pumped up from anywhere in the valence band to anywhere in the conduction band – always provided the transition goes vertically upwards in the reduced band diagram.

- The electrons in the conduction band as well as the holes in the valence band will quickly move to the extrema of the bands – corresponding to the levels E_2 and E_1 in the general four-level system.
- "*Quickly*" means within a time scale defined by the dielectric relaxation time. This time scale is so small indeed that it introduces some uncertainties in the energies via the **uncertainty relation** (which is considered in the advanced module but need not bother us here).

We have now everything needed for a "quick and dirty" derivation for the inversion condition in the sense introduced at the top.

The Inversion Condition

The condition for inversion was that there were at least as many stimulated emission processes as fundamental absorption processes. The recombination rate by **stimulated emission** we now denote R_{se}, and the electron–hole pair generation rate by **fundamental absorption** is R_{fa}. We thus demand:

$$R_{se} \geq R_{fa}$$

- In general, fundamental absorption and stimulated emission can happen in a whole range of frequencies for semiconductors. While we expect that the electrons that are being stimulated to emit a photon will occupy levels right at the conduction band edge, stimulated emission is not forbidden for electrons with a higher energy somewhere in the conduction band. While these electrons are in the (fast) process of relaxing to E_C, they still might be "hit" by a photon of the right energy at the right time and place – it is just more unlikely than at E_C.

We thus must expect both rates, R_{se} and R_{fa}, to be proportional to:

1. The **spectral intensity of the radiation** in the interesting frequency interval.

- The differential frequency interval considered extends from ν to $\nu + \Delta\nu$; the spectral intensity in this interval we name $u(\nu)\Delta\nu$ or, expressing the frequency ν in terms of energy via $E_{phot} = h\nu$, $u(E)\Delta E$.
- This value, $u(E)\Delta E$, divided by the single-photon energy $E_{phot} = h\nu$, essentially gives *the flux of photons in this frequency interval* (i.e., the number of photons arriving per second and per area, for short).

2. The **density of states** available for the processes.
    - The probability that a photon with a certain frequency ν and therefore energy $E_{phot} = h\nu$ will be absorbed by an electron at *some* position E_1 in the valence band (i.e., close but not necessarily equal to E_V), will be proportional to the density of states in the valence band, $D_V(E_1)$, *and* to the density of states exactly E_{phot} above this position in the conduction band, $D_C(E_1 + h\nu)$.
    - Contrariwise, the probability that stimulated emission takes place, triggered by a photon with energy $h\nu$, is proportional to the density of states in the conduction band *and* to the density of states $h\nu$ below in the valence band.
    - This is a crucial part of the consideration – and a *rather strange one*, too: That *both* densities of states must be taken into account – where the particle is coming from *and* where it is going to – is a quantum mechanical construct (known as Fermi's golden rule) that has no classical counterpart.
3. The **probability** that the states are actually occupied (or unoccupied).
    - The density of states just tells us how many electrons (or holes) *might* be there. The important thing is to know how many *actually are there* – and this is given by *the probability that the states are actually occupied* (necessary for absorption or stimulated emission) or *unoccupied* (necessary for the transition of electrons to this state).
    - In other words, the *Fermi–Dirac distribution* comes in. In the familiar nomenclature we write it as $f(E, E_F{}^e, T)$ or $f(E, E_F{}^h, T)$ with $E_F{}^{e,h}$ = quasi Fermi energy for electrons or holes, respectively.
    - The *crucial* point is that we take the *quasi Fermi energies*, because we are *by definition* treating strong non-equilibrium between the bands, but (approximately) equilibrium in the bands.
    - We also, for ease of writing, define a direct Fermi distribution for the holes as outlined before and distinguish the different distributions by the proper index:

$f_{e \text{ or } h}(E, E_F{}^{e,h}, T)$	=	probability that some level at energy E *is* occupied by an electron or hole
$1 - f_{e \text{ or } h}(E, E_F{}^{e,h}, T)$	=	probability that some level at energy E is *not* occupied by an electron or hole

- Remember that "not occupied by a hole" always means "occupied by an electron" – whereas the meaning of "not occupied by an electron" depends on what is referred to: Only for the valence band this means "occupied by a hole"! (Do you also remember why this is so? If not: Think about charge neutrality!)

That is all. However, the density of states are complicated functions of E, and the spectral density of the radiation we do not know – it is something that should come out of the calculations.

- But we are doing shortcuts here, and we do know that the radiation density will have a maximum around $h\nu = E_g = E_C - E_V$. So let's simply assume that the necessary integrations over $u(E) \cdot D(E)\Delta E$ will be expressible as $N_{eff} \cdot u(\nu) \cdot \Delta\nu$ with N_{eff} = effective density of states. Moreover, we assume identical N_{eff} in the valence and conduction band.
- The rates R_{se} for stimulated emission and R_{fa} for fundamental absorption then can be written as

$$R_{fa} = A_{fa} \cdot N_{eff}^2 \cdot u(\nu) \cdot \Delta\nu \cdot \left(1 - f_{h \text{ in } V}(E_1, E_F{}^h, T)\right) \cdot \left(1 - f_{e \text{ in } C}(E_1 + h\nu, E_F{}^e, T)\right)$$

$$R_{se} = A_{se} \cdot N_{eff}^2 \cdot u(\nu) \cdot \Delta\nu \cdot \left(f_{e \text{ in } C}(E_1 + h\nu, E_F{}^e, T)\right) \cdot \left(f_{h \text{ in } V}(E_1, E_F{}^h, T)\right)$$

- The A_{fa} and the A_{se} are the proportionality coefficients and we always use $f_{h \text{ in } V}$ if we consider carriers in the valence band and $f_{e \text{ in } C}$ if we consider the conduction band.

Enters Albert **Einstein**. He showed in **1917** that the following extremely simple relation *always* holds for fundamental reasons:

$$\boxed{A_{fa} = A_{se}}$$

- We will just accept that (if you don't, turn to the advanced module for a derivation) and now form the ratio R_{se} / R_{fa}. The coefficients then just drop out and we are left with

$$\frac{R_{se}}{R_{fa}} = \frac{[f_{e \text{ in } C}(E_1 + h\nu, E_F{}^e, T)] \cdot [f_{h \text{ in } V}(E_1, E_F{}^h, T)]}{[1 - f_{h \text{ in } V}(E_1, E_F{}^h, T)] \cdot [1 - f_{e \text{ in } C}(E_1 + h\nu, E_F{}^e, T)]}$$

- With some shuffling of the terms (see the exercise below) we obtain

$$\frac{R_{se}}{R_{fa}} = \frac{E_F{}^e - E_F{}^h}{h\nu}$$

- with E_1 and $E_1 + h\nu$ denoting some energy level in the valence or conduction band, respectively, implying $h\nu \geq E_g$ (since for direct semiconductors, the smallest possible difference between some energy levels in the valence band and some energy levels in the conduction band that are connected by a direct transition is E_g).

This is a rather simple, but also rather important equation. It says that we have *more* stimulated emission between $E_1 + h\nu$ and E_1 than fundamental absorption between E_1 and $E_1 + h\nu$ if the *difference in the quasi Fermi energies is larger than the difference between the considered energy levels*.

- Thus, we have as the **first laser condition**:

$$E_F{}^e - E_F{}^h \;\geq\; h\nu \;\geq\; E_g$$

- We call this "*laser condition* ", because "lasing" requires inversion, i.e. that there are at least as many electrons at the conduction band edge as we have *electrons* (not holes!) at the valence band edge.

It is clear that this involves heavy non-equilibrium conditions.

- We need to *inject a lot of electrons* into the conduction band and a *lot of holes* (= taking electrons out) into the valence band.
- And we have to keep the *injection rates* at least as large as the stimulated emission rate, i.e. we have to supply electrons (and holes) just as fast as stimulated emission takes them away if we want to keep the rate of radiation constant.

Now we know what is needed to obtain light amplification in principle. But how much amplification do we get from a piece of semiconductor kept in inversion? This will be the topic of the next module.

Exercise 6.1-1

Do the math for the **1st** laser condition

6.1.2 Light Amplification in Semiconductors

- If we produce a state of inversion, i.e. we have at least as many electrons in a high energy state as in the low energy state to which they fall be radiating recombinations, we will be able to amplify light.
  - Let's look at a real light amplifier now. The input light with an intensity ***I*** **0** enters the material, travels through the length of it, getting amplified all the time, and finally exits with a higher intensity ***I***.
- For a quantitative analysis, let's consider a semiconductor which we keep in inversion conditions extending from ***z* = 0** to ***z* = *L*** along the ***z***-axis. We now shine some light on it a ***z* = 0** and with the spectral intensity $u_\nu(z = 0)$.
  - By definition, the rate for stimulated emission events, R_{se}, will increase in ***z***-direction and so does the spectral intensity of the radiation, $u(\nu , z)$ which we now write somewhat simplified as $u_\nu (z)$.
  - We also assume that the inversion conditions are the same everywhere (i.e. they do not depend on ***z***), and now define the *net rate of stimulated emissions* in short hand:

$$R^{net}{}_{se} \;=\; R_{se} \;-\; R_{fa} \;=\; R^{net}{}_{se}(z) \;=:\; R(z)$$

  - The interesting quantity, if we want to discuss the amplification of light, is $u_\nu(z)$, corresponding to the density of photons that, per second, travel along the ***z***-axis. If we have any amplification at all, it will increase for increasing ***z*** and the rate of increase is somehow given by an amplification factor g_ν which must be a function of the "strength" of the inversion condition which in turn is tied to ***R*(*z*)**. We must expect that amplification is different at different frequencies and it is thus wise to index ***g*** with "ν ". If g_ν is constant (which we can expect for constant ***R***), the spectral intensity $u_\nu (z + \Delta z)$ at some point ***z* +** Δ***z*** will be given by $u_\nu (z)$ as follows:

$$u_\nu (z + \Delta z) \;=\; u_\nu(z) \;+\; g_\nu \cdot u_\nu(z)\Delta\, z$$

  - This means that the intensity at the end of some length Δ***z*** of material is given by the intensity available at the entrance plus the part that is generated in the length increment considered. This part is proportional to the incremental length Δ***z*** available for amplification, the factor g_ν, which is defined by this equation and which will be properly called **gain coefficient**, and the intensity available.
- Making Δ***z*** arbitrarily small, i.e. moving from Δ***z*** to **d*z***, yields a simple differential equation:

$$\frac{u_\nu(z \,+\, dz\,) \;-\; u_\nu(z)}{dz} \;=\; \frac{du_{\,\nu}\,(z)}{dz} \;=\; g_{\,\nu} \cdot u_\nu(z\,)$$

  - The solution, of course, is

$$u_\nu(z) \;=\; u_\nu(0) \cdot \exp\;(g_\nu \cdot z)$$

  - If we measure the intensity ***I*** of the light in some conventional units, we have the same relation, of course, because any measure of intensity at some frequency ν is always proportional to the number of photons. The increase in intensity then is

$$I_\nu\,(z) \;=\; I_\nu(0) \cdot \exp\;(g_\nu \cdot z)$$

- Formally, this is nothing but Beer's law of absorption if we introduce a *negative* absorption coefficient α, i.e α $= -\, g_\nu$.
- While this was fairly straightforward, two questions remain:
  - *First*, an obvious question: What determines the gain coefficient?
  - *Second,* something a bit less obvious. At this point we have made all kinds of assumptions and approximations, and it is difficult to keep track of what kind of problem we are considering compared to reality. If you think about this, it all boils down to the following question: Are there *other losses* besides fundamental absorption to the photons and to the electrons in the inversion state that we have not yet included? Because if there are, we will have a harder time to amplify light than we think we have. Our present major goal, the amplification of light, then would be harder to achieve.
- Those are rather difficult questions which we will only consider summarily in this module. There are, however, links to more advanced stuff in what follows.

Gain Coefficient and Transparency Density

- Looking back at the simple example used for defining inversion, it is clear that the gain coefficient g_ν increases if the degree of inversion, i.e. the ratio of stimulated emission to fundamental absorption events, increases.
  - This implies that g_ν increases with increasing carrier density in the conduction band which in turn demands that E_F^e, the quasi Fermi energy of the electrons in the conduction band, moves deeper into the conduction band.
  - The gain coefficient g_ν , moreover, will be largest at the frequency corresponding to the energy levels where most electrons can be found. This level moves up in energy with increasing density of the electrons; for very few electrons it is of course E_C.
- Again, from the simple example used before, we can conclude that for the onset of inversion, i.e., for identical rates of fundamental absorption and stimulated emission, nothing happens in total: Exactly the same number of photons emerges at the output as fed into the input. We have

$$\boxed{I_\nu(z) \;=\; I_\nu(0) \cdot \exp(g_\nu \cdot z) \;=\; I_\nu(0)}$$

  - which demands **exp ($g_\nu \cdot z$) = 1** or **$g_\nu = 0$**.
  - The effect now will be that the semiconductor appears *completely transparent to the light*. The necessary density of electrons (for some fixed density of holes) is called **transparency density $n^e{}_T$**.
  - If the carrier density n^e increases beyond $n^e{}_T$, the maximum value of g_ν obtained at a certain frequency (which increases sightly with n^e) will increase, too, in a pretty much linear fashion. We find more or less empirically:

$$\boxed{g_\nu{}^{\max} \;=\; a \cdot (n^e - n^e{}_T)}$$

  - The factor a may simply be considered to be a material constant called **differential gain factor**, since it depends more or less on material parameters like *density of states*, *band gap*, etc., and is hard to calculate for real materials. For example, GaAs has a value of $a \approx 2.4 \cdot 10^{-16}\ \text{cm}^2$.
- In total, we have a complex dependence of g_ν on carrier density and frequency. An advanced module will give more details.

Additional Losses

- There are two kinds of possible losses that we may have to worry about.
  - We may loose some photons somehow which then cannot stimulate electrons to emit another photon.
  - We may loose electrons in some recombination channels; these electrons then aren't available for stimulated emission.
- The first kind of loss includes fundamental absorption, but that is already included in the theory so far. Are there other optical losses in the semiconductor?
  - Well, there are. Besides fundamental absorption, we also have the kind of absorption that prevents metals from being transparent: *Photons are generally absorbed by the free carriers*, i.e., by the electrons in the conduction band. If we increase the carrier density we will increase this effect.
- The second kind of loss includes all electrons that recombine through one of the other channels available: deep levels, direct recombinations, Auger recombination, excitons, ...
  - These recombination events not only reduce the number of available electrons, but the ones disapperaring via radiative recombination produce some light of their own – with the right wave length but with *random phases*, i.e., not coherent to the light we care for. This light also becomes amplified and induces a kind of **phase noise**.
  - Luckily, all these recombination losses are negligible for real devices.
- There might be more loss mechanisms, but we will sweep 'em all under the rug and simply combine everything there is (or might be) with respect to intrinsic losses (i.e., inside the semiconductor) in an **intrinsic loss coefficient α_i**.

- The decrease in intensity due to the intrinsic losses then is

$$I_{\text{loss}}(z) \;=\; I_\nu(0)\cdot\exp\,(-\alpha_i\cdot z)$$

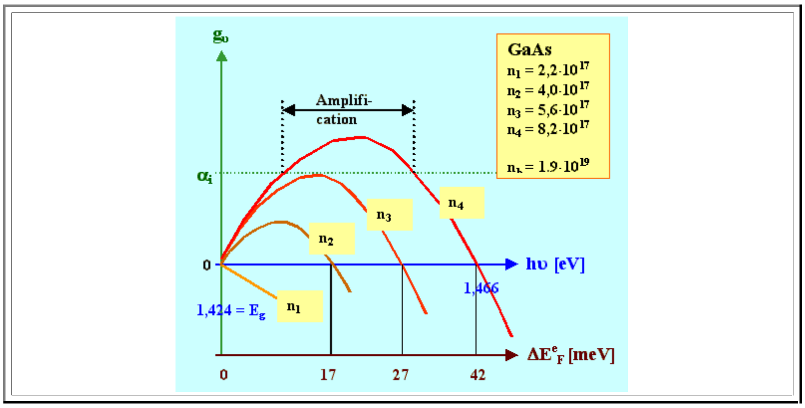
  - We can combine gain and losses then to the final equation linking the input to the ouptput:

$$I_\nu(z) \;=\; I_\nu(0)\cdot\exp[(g_\nu \;-\; \alpha_i)\cdot z]$$

  - For a constant gain coefficient along the crystal, the total output is then

$$I_\nu(L) \;=\; I_\nu(0)\cdot\exp[(g_\nu \;-\; \alpha_i)\cdot L]$$

- There is an important consequence from this equation: Amplifications demands that $g_\nu - \alpha_i > 0$ and that requires $g_\nu > \alpha_i$.
  - Just achieving inversion (corresponding to $g_\nu = 0$) thus is not good enough. There is a minimum or *threshold* value given by α_i before light amplification will occur. In other words: The density of electrons has to be larger than just the transparency density $n^e{}_T$.
- What this means in practice is shown below in a schematic way.
  - Shown are curves for g_ν for **GaAs** in a halfway realistic manner including some numbers.



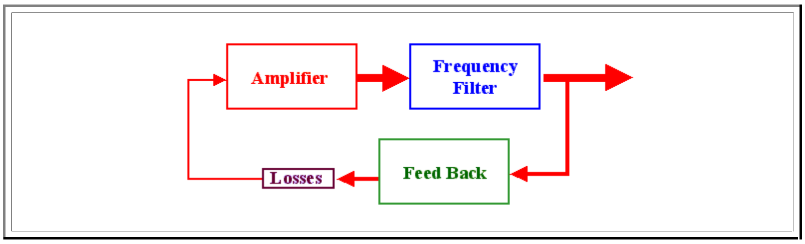
  - A constant hole density of $\mathbf{n_h = 1\cdot 10^{19}\ cm^{-3}}$, i.e., *heavily* doped **p**-type **GaAs** has been used as a reference. The electron density is raised by injection to four values marked $\mathbf{n_1}$ through $\mathbf{n_4}$.
  - The gain coefficient g_ν is given as a function of the frequency (in terms of energy). A second scale shows the necessary level of the quasi Fermi energy for the electrons as $\Delta E^e{}_F$ above the conduction band edge.
  - For the electron density $\mathbf{n_1}$ we have the onset of inversion. The gain coefficient is $g = 0$ for exactly one frequency corresponding to the band gap.
  - With increasing **n**, the gain coefficient is **> 0** for a portion of the frequency interval, peaking at about the frequency corresponding to $E_g + \tfrac{1}{2}(\Delta E^e{}_F)$ - that's where most of the electrons are!
  - Only when g_ν is above the intrinsic loss coefficient α_i, which has been drawn in in a halfway realistic manner, some amplification occurs in the part of the spectrum indicated.
- More to this in the advanced module "gain coefficient"

6.1.3 From Amplification to Oscillation: Second Laser Condition

- So far we have the principle of the amplification of light by stimulated emission. Making a laser in the conventional sense of the word still requires to produce a light beam with a "battery" – not with some input light.
  - Since we nevertheless need at least some input light, this is the same task as to produce an *oscillator* from an amplifier in electronics, and the solution of this task is achieved along identical lines: Feed back *one* frequency from the output of the amplifier to the input and make sure it is in phase (or, as we say for light, "coherent"). Thus, we have to take . . .

A General Look at Feedback and Oscillations

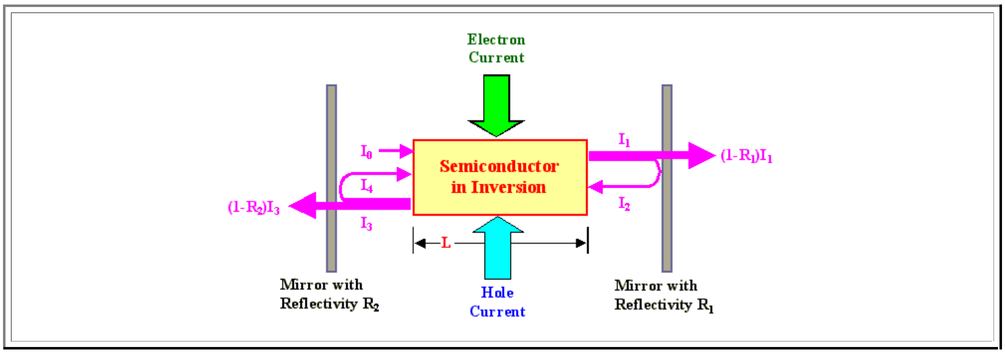
- Feeding one frequency from the output back to the input will lead to an amplification of this frequency.
  - Since then also the intensity of the feedback signal increases, this one frequency will become more amplified, and so on ... – pretty soon your system is now an oscillator for the frequency chosen. That is, you feed back a large enough part of the output to account for losses that may be occurring in the feedback loop so that still sufficient amplitude is left to drive the amplifier. The essential parts are shown in the drawing:



  - If you think about this, you will discover a problem. If there is enough signal at the input, the output will go up forever or until a fuse blows – there is no stability in the system
  - We need some kind of servo mechanism that adjusts the amplification factor to a value where only the losses are recovered by amplification, so that a stable, preferably adjustable output amplitude is obtained.
- This is clear enough for electrical signals, but how do we do this with light? Well, we do everything with mirrors:
  - **1**. The **feed back** part in general.
  - **2.** The **coherency** requirement.
  - **3.** The **selection of the frequencies**.
  - **4.** The **guidance of the light** including the "*beam shaping* " of the output.
- The **4th** point is new – after all, electrical signals go to wherever the wires go, but with light we have to make sure we get a *single* beam if we like to have one.
  - We first look at the general principle of light feedback without worrying about the three other points (which we throw in afterwards).

Feedback with Mirrors

- All we have to do is to put the piece of semiconductor that is supposed to amplify the light by stimulated emission between two *partially transparent* mirrors
  - The whole system then looks like this.


Electron Current
I_0
I_4
I_3
$(1-R_2)I_3$
Semiconductor in Inversion
I_1
I_2
$(1-R_1)I_1$
L
Mirror with Reflectivity R_2
Hole Current
Mirror with Reflectivity R_1

- Let's look at this in a quantitative way. We start the analysis by feeding some light with the intensity I_0 (from the left) to the semiconductor which we keep in a state of inversion by constantly supplying the necessary electrons and holes. It will be amplified along its way through the semiconductor (length L) if we exceed the threshold for amplification and emerge with the intensity I_1 on the right
- Parts of I_1 will be reflected; this intensity we call I_2, it is given by

$$\boxed{I_2 \;=\; I_1 \cdot R_1}$$

- R_1 is the reflection coefficient for mirror **1**. For $R_1 = 1$ all the light will be reflected, for $R_1 = 0$ all the light will be transmitted.
- The transmitted intensity is then simply $I_1 \cdot (1 - R_1)$ as indicated in the drawing.

But now we have light traveling (and getting amplified) from right to left.

- It will emerge with the intensity I_3, some of it ($I_4 = I_3 \cdot R_2$) will get partially reflected and run through the crystal, and so on and so on.

Eventually, we will reach a steady state with all intensities being constant. Let's see what that will be.

- First we write down all the relations for the intensity that we have, using the formula from before that links the output to the input:

$$\boxed{\begin{aligned} I_1 &= I_0 \cdot \exp[(g - \alpha_i) \cdot z] \\ I_2 &= I_1 \cdot R_1 \\ I_3 &= I_2 \cdot \exp[(g - \alpha_i) \cdot z] \\ I_4 &= I_3 \cdot R_2 \end{aligned}}$$

- We dropped some indices for ease of reading and obtain immediately for, e.g., I_4

$$\boxed{I_4 \;=\; I_0 \cdot R_1 \cdot R_2 \cdot \exp[(g - \alpha_i) \cdot 2L]}$$

To make life easy, we use a small trick and assume $R_1 = R_2 = R$ (a reasonable choice automatically fulfilled if we take as partially reflecting mirrors simply the surfaces of the crystal).

- Next, because light not reflected back into the crystal is lost, we express the reflection part in terms of losses by smartly defining the quantity

$$\boxed{a_R \;:=\; -\frac{1}{2L} \cdot \ln(R_1 \cdot R_2) \;=\; -\frac{1}{2L} \cdot \ln(R^2) \;=\; -\frac{1}{L} \cdot \ln R}$$

- Since $R < 1$, α_R is always positive because of the minus sign. This gives us

$$\boxed{R_1 \cdot R_2 \;=\; R^2 \;=\; \exp(-2 \cdot \alpha_R \cdot L) \quad \Rightarrow \quad \exp(-\alpha_R \cdot L) = R}$$

- The losses of the external output due to the partial reflection as it would appear to an "outside" observer thus are assigned to the crystal, too, and the factor ½ or **2**, respectively, appears because the light travels *twice* through the crystal.
- This gives the final form for I_4:

$$\boxed{I_4 \;=\; I_0 \cdot \exp\Big([g - (\alpha_i + \alpha_R)] \cdot 2L\Big)}$$

What does this equation tell us? It contains two essential pieces of information:
1. The condition for *starting the process*, and
2. the conditions for the *stationary state*.

Let's look at this in detail:

- The requirement for *starting the process*, i.e. for starting the oscillator, is that after *one* cycle (from I_0 to I_4) we must have recovered I_0. Or, in formal words, the *starting condition* is

$$\begin{gathered} I_4 \geq I_0 \\ [g - (\alpha_i + \alpha_R] \cdot 2L \geq 0 \end{gathered}$$

- This then defines a **threshold value g_{th}** for the gain factor g which is given by

$$g_{th} = \alpha_i + \alpha_R = \alpha_i - \frac{1}{L} \cdot \ln(R)$$

- If the system has a gain coefficient above this value, *one* photon will be enough to start the process, and since one photon is always around, the system will then start to produce light on its own without outside help.
- Since the gain coefficient is a strong function of the *carrier density*, this also means that light production will start automatically as soon as the carrier density (= electrons in the conduction band) reaches a **threshold value n^e_{th}**. And that density is larger than the density needed for inversion or transparency.

- Now to the second questions: What determines the stationary state, or how much light is actually produced? Naively, we would expect that after the start, the intensity will go up in every cycle, and if nothing changes, it will go through the roof to *infinity* .
  - This, of course never happens, because it would imply that you inject an infinite amount of new carriers for stimulated emission to occur at the required rate. Clearly then, the limited supply of carriers will bring down the gain coefficient and some steady state can be expected for some specific carrier density.
  - Steady state simply means that your gains are exactly identical to the losses, and this means $I_4 = I_0$.
  - This is essentially the same equation as for the start of the process (only the **">"** sign is missing) and we obtain the final result for the gain coefficient in stationary state, g_{stat}, and by inference for the carrier densities n^e_{stat} :

$$\begin{gathered} g_{stat} = \alpha_i - \frac{1}{L} \cdot \ln R = g_{th} \\ \\ n_{stat} = n^e{}_{th} \end{gathered}$$

- While this looks deceptively simple, it provides a lot of open questions. First of all, we have only met our first requirement from above for an oscillator producing coherent light at a defined frequency; the other ones are still open. Then we might ask ourselves, exactly how the crystal manages to regulate the gain coefficient, or how the intensity evolves with time?
  - Since these questions are interrelated, we first look at how requirements **2–4** can be met.

- Requirement **4** is easy now:
  - The photons travel according to the laws of geometric optics (in a first approximation). With planar mirrors perpendicular to the z-direction, they just run back and forth.
  - If we use inclined mirrors, or bent mirrors, or fibre optics, things may become complicated, but in principle we know how to treat it.
  - We therefore will not worry about this point any more, but simply stick to the simple back-and-forth light path arrangement shown in the drawing .

- Requirements **2** and **3** can be met with the same basic trick:
  - *Chose the (optical) distance between the mirrors to be a multiple of half of the wave length you want*. If you use external mirrors you must take into account that the wave length in air is different from that in the crystal, that's were the qualifier "optical" comes in.
  - If we simply use the surfaces of the crystal as mirrors, the length between the mirrors is L = length of the crystal and the condition given above then is

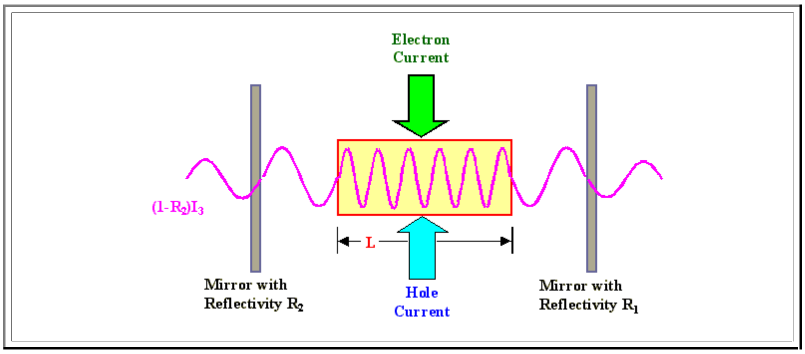
$$L = \frac{m \cdot \lambda_{air}}{2n_{ref}} = \frac{m \cdot \lambda}{2} \qquad m = 1, 2, 3, 4, \ldots$$

- With λ_{air} = wave length in air, λ = wave length in the crystal, n_{ref} = refractive index of the crystal.

- If we do this, we will have a coherent beam of light travelling in z *and* $-z$-direction with a wave length λ_{air} that is
1. given by the equation above, and
2. lies in the wave length region where the gain factor is sufficiently large.

- While the first condition would still allow many wave lengths, the second conditions normally admits just one. But why?
  - Simply because any two mirrors at any distance define a resonant structure and that is why such a system is called a **Fabry- Perot resonator** (or interferometer). Only light with wave lengths given by the *Fabry-Perot resonance condition* $\lambda = 2nL/m$ can exist inside a "Fabry-Perot" as a standing wave.
  - This is best seen by looking at what would happen to light with a "wrong" wave length. Every time it travels through the system, its phase is shifted to some extent, and pretty soon you have a wave with the phase $-\pi$ for any wave with the phase **0** and destructive interference will cancel everything except waves with phases that fit.
  - This is the same old principle that governs diffraction of electrons or **X**-ray beams in crystals, all musical instruments, and, if you believe Richard **Feynman** , just about everything else, too.

- In other words, we already met the **2.** and **3.** requirement (without noticing, perhaps), but *not necessarily* at the optimal frequency which is of course the frequency with the highest gain factor $g(\nu)$.
  - Or in yet other words: While the picture of light waves travelling in and out of the crystal is not wrong, what we really have after a very short time is a *standing wave* inside the Fabry-Perot resonator with usually just *one* dominating wavelength from the multitudes possible. It looks like this



  - Shown is the intensity, i.e. the *square* of the amplitude (and not the amplitude as a function of time) of a standing wave with a wave length considerably smaller than the length of the crystal – as we will encounter it in reality.
  - The wave length is determined by the condition that $\hbar\omega \approx E_g$, or, with $\omega = c/(n_{ref} \cdot \lambda)$

$$\boxed{\frac{h \cdot c}{n_{ref}} \approx E_g}$$

  - Whichever way we describe the light – by its wave length λ, its angular frequency ω , or its energy $\hbar\omega$, we can always index these quantities now with a "**g**" for "gap" and know how to calculate the numerical values for, e.g. ω_g.

- Taking the requirement for the threshold gain and the admissible wave lengths together is called "**second laser condition**", i.e.

$$g_{stat} = \alpha_i + \alpha_R = \alpha_i - \frac{1}{L} \cdot \ln R = g_{th}$$

$$L = \frac{m \cdot \lambda_{gair}}{2n_{ref}} = \frac{m \cdot \lambda_g}{2}$$

  - Since g is a function of the wave length λ_g or the frequency ν_g, respectively, and the carrier density n^e (*do not mix it up with the refractive index* n_{ref} !); $g(\lambda, n^e)$, we can combine both equations into

$$\boxed{\alpha_i + \alpha_R = g(\nu_g, n^e_{th})}$$

- What do we know about the quantities in this equation?
  - We know that α_i with its various components is primarily a function of the carrier density; we need its value at the **threshold density** n^e_{th}.

- We can expect that the optical losses described by α_R are pretty much constant, but $\boldsymbol{g(\nu, n^e)}$ is a rather complicated function defined by integrals over densities of states times Fermi distributions and the like; we thus have a complex (integral) equation for the determination of $n^e{}_{th}$, our only unknown parameter at this point.

Computing $n^e{}_{th}$ from the second laser condition can only be done numerically and requires good knowledge of the relevant quantities. We can get a rough estimate, however, by neglecting the frequency dependence and taking the maximum value of g at the fixed frequency, g^{max}. And for $g^{\,max}$ we had the empirical equation

$$g_{\nu}{}^{max} \;=\; a \cdot (n^e \;-\; n^e{}_T\,)$$

- $n^e{}_T$ was the transparency density, and ***a*** the *differential gain factor*, a material constant.

Inserting this equation for $\boldsymbol{g(\nu_g, n^e{}_{th})}$ in the second laser condition from above yields a kind of (approximate) master equation for semiconductor lasers

$$n^{\,e}{}_{th} \;=\; n^{\,e}{}_T \;+\; \frac{\alpha_i \;+\; \alpha_R}{a}$$

- It includes the first laser condition (which defined $n^e{}_T$) in the conditions for self-induced oscillations at the "right" frequency, parts of which are released to the outside world (this is the α_R part).

And, of course, what we will have as soon as the carrier density that we inject in our semiconductor crystal contained within a properly spaced Fabry-Perot resonator reaches the threshold is a **LASER** in the specific meaning discussed before.

- This leaves us now with the big question: How do we make a semiconductor laser? Or, for that matter, a simple light emitting diode, which will turn into a laser if we put it inside a Fabry-Perot and crank up the injected carrier density "somehow".

6.2 Specific Topics

6.2.1 Turning on a Laser Diode

It would be a quite intesting module. Too bad, I never got around to doing it.

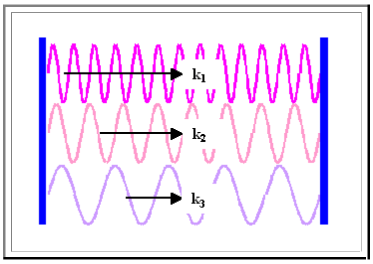
- Sorry!

6.2.2 Laser Modes

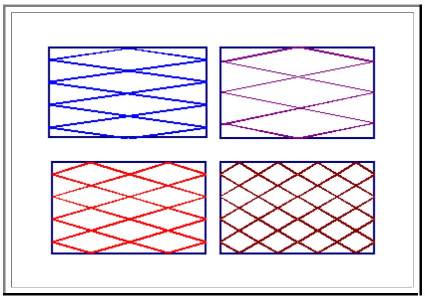
This module is not finished but you can get a rough idea of what it's all about.

- The Fabry Perot resonator introduced in chapter 6.1.3 is an oversimplification of the situation in a *real* semiconductor Laser.
  - Without mentioning it, we have assumed an infinitely extended system in the illustrations, i.e. a one-dimensional situation.
  - The active region in a real Laser, however, is finite. Often, it consists of a particular material embedded in an other material with a *different* index of refraction; in any case it ends somewhere. In a most simple approximation we may consider it to be a box of length ***l***, thickness ***d*** and width ***w***.
- This simply means that *many* standing waves - with different wavelengths and different wave vector directions - satisfy the resonance condition.
  - In *other words* - and that is the common lingo - the Laser cavity may contain many internal **modes** and thus does not automatically emit monochromatic light in one direction only.
  - We may distinguish between **axial** or **longitudinal modes**, and **transverse** modes. The figures below illustrates this

Longitudinal Modes



Transverse Modes



  - Many wavelengths fit in the *longitudinal* direction which we define to be the direction where we want emission We have $\boldsymbol{l = m \cdot \lambda/2n_r}$ and $\boldsymbol{m = 1,2,3,...}$

Many transverse modes are possible as shown. They are undesirable and should be avoided.

  - Only wavelengths compatible with the band gap energy, i.e. $\lambda = c/n_r \cdot \nu \approx c \cdot h/n_r \cdot E_g \approx$ **µm** will become amplified, i.e. ***m*** is large since ***l*** is typically many **µm**.
  - The distance between allowed frequencies is $\Delta\nu = c/2l \cdot \nu \approx$ **80 GHz** for $l = 500$ **µm**. The emission lines of the longitudinal, modes are thus very close together.
- Laser modes, what to do with them, and how to make a Laser working in only *one* mode - this is what we naively expect a Laser to be - is clearly a science in itself.
  - We will not go into details, suffice it to say that **monomode Lasers** are possible by optimizing the resonating properties of the cavity to the local gain inside it.

7. Light Emitting Devices

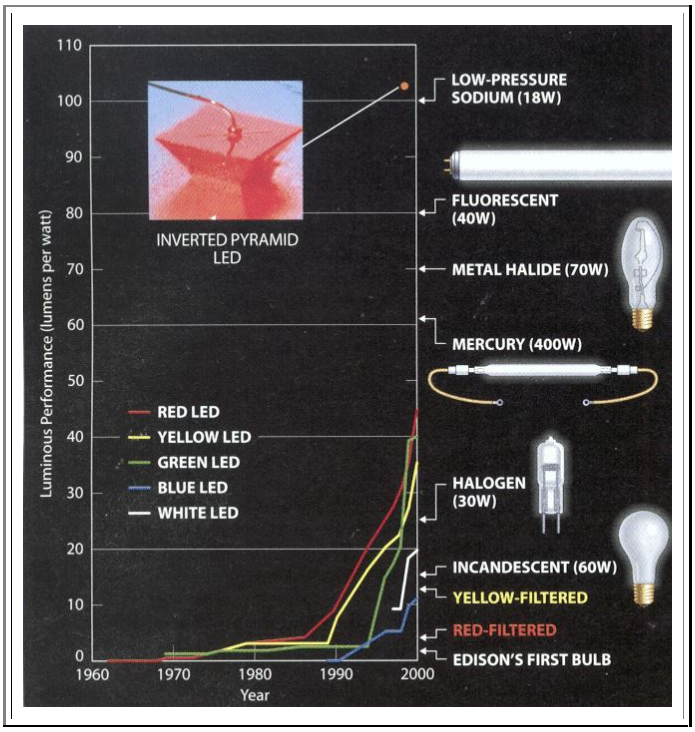
7. Light Emitting Devices

7.1 Basic Requirements and Design Principles

7.1.1 Products, Market, Materials, and Technologies

General Considerations

- All light emitting devices – **LEDs** or laser diodes – have some device principles and requirements in common:
  - There is a defined volume in the device – the *recombination zone* (or active volume) – where the generation of light takes place.
  - The increase in the minority carrier density necessary for radiative recombination is obtained by injecting electron and hole currents across suitable junctions.
  - The device is made so that most of the injected carriers recombine radiatively in the active volume – i.e. the quantum efficiency and the current efficiency should be as large as possible. This is exactly the opposite of the regular **Si p–n** junction, where we try to keep recombination in the **SCR** (and thus leakage currents) as low as possible.
  - For lasers, an optical feedback mechanism is added (e.g. a Fabry–Perot resonator). In addition, the geometric shape is important: Should the laser emit along a line, or just from a "point"?
  - The optical efficiency must be optimized, too
  - And, not to forget, an important consideration neglected so far: For many applications the modulation **frequency range** should be large. In other words, we want to modulate the light intensity by modulating the injection currents at high frequencies – **GHz**, if possible.
- This is a demanding list of specifications; it cannot be met with just a few basic device architectures.
  - Considering the wide range of available semiconductors and the extremely diversified product spectrum, there is a bewildering multitude of devices from many materials involving often ternary and quaternary semiconductors from the **III-V** zinc-blende lattice set.
  - While the most complicated devices concern laser diodes (always with spectacular physics involving all kinds of quantum well structures and tricky resonators), the humble **LED** is not to be sneered at either. It is always the base of laser: If you cannot make an **LED** for a certain wavelength, you sure like hell will also not be able to make a laser.
  - The field was revolutionized some years ago when Shuji **Nakamura** , almost single-handedly, made working **blue LEDs** based on **GaN**; a feat that seemed to be impossible since all the big players in the field could not do it.
  - At present, a race for a **12 · 10^9 $/year** market is gaining in speed (and expenditures for research): *Cheap* **LEDs** suitable to **replace light bulbs**, emitting *white* light at high intensity may be around the corner! But maybe they are not. Only time (after considerable research and development) will tell.
- For light emitting diodes (and of, course, for laser diodes even more so), efficiency is of supreme importance: How much light can you get for **1 W** of electrical power that goes in the device?
  - Much progress has been made, and more will have to be made for light bulb replacement. The figure below illustrates that.



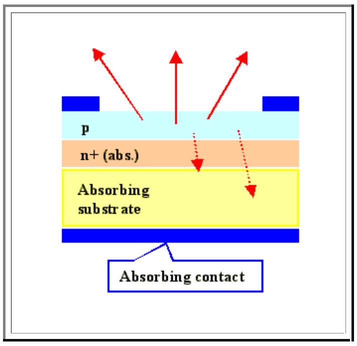
- The record holder, the inverted pyramid LED, is described in some detail in the link.

How large is the market for III-V devices?

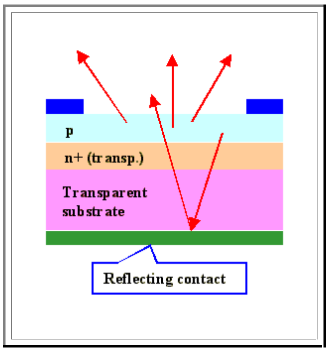
- Here is a link to some information.

7.1.2 Some LED Concepts

- **LEDs** come in many variants, satisfying needs form being cheap to being "super". The figures show some common devices

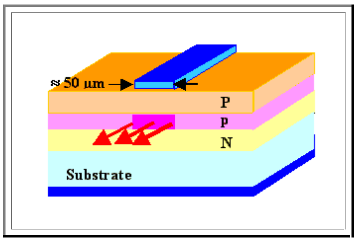


- This is possibly the most simple **LED** device. Electrons are injected into the top **p**-layer, and only the photons that manage to escape will be seen. Of course, you will try to keep the top layer as thin as possible.
  - Not very good, but then there are many applications were you do not want particularly bright light, but cheap products, e.g. for indicator lights in stereo systems, dashboards, etc.
  - You may use the same kind of material – which will be automatically absorbing the light flowing into the depth of the device. For red light, you use **GaAlAs**, for green **GaP**, or anything that comes in handy from the table of possible mixtures.
  - The necessary layers you make with some kind of epitaxy, which allows you to work with relatively cheap substrates.



- A somewhat better device uses the light emitted to the back side by reflecting it back to the front side.
  - If the light has sub-bandgap energies because it stems from excitons, you do not have large absorption effects in the basic material. So for **GaP LEDs**, it pays to make the back contact reflective and keep the layers thin.
  - Generally, however, this approach requires heterojunctions where the **n$^+$** layer and the substrate material must have a larger bandgap than the active layer, so they are transparent to the light.
  - The **N$^+$p** heterojunction may have the added benefit that the injection of electrons becomes more efficient, but it also has the added problem that now you must watch out for lattice constant compatibility, otherwise you may encounter misfit dislocations

- These kinds of **LEDs** emit over the whole area of the active layer; they are called **surface-emitting LEDs**. They are good enough for most general light source applications.
  - With quite some additional effort, such an **LED** can be further developed into a laser, then known as a **VCSEL** (= vertical-cavity surface-emitting laser). Meanwhile, despite the additional effort (mainly for making the mirrors), **VCSELs** are already widely used. (For more information, crawl the web yourself.)
- If you want really high **intensities** , i.e., not just a lot of photons but a lot of *photons per area*, you must confine the light emission to a small area where you realize high injections ratios. This is particularly important if the emitted light is to be coupled to a fiber or wave guide for optical communication purposes. This can be done with an **edge-emitting LED**:

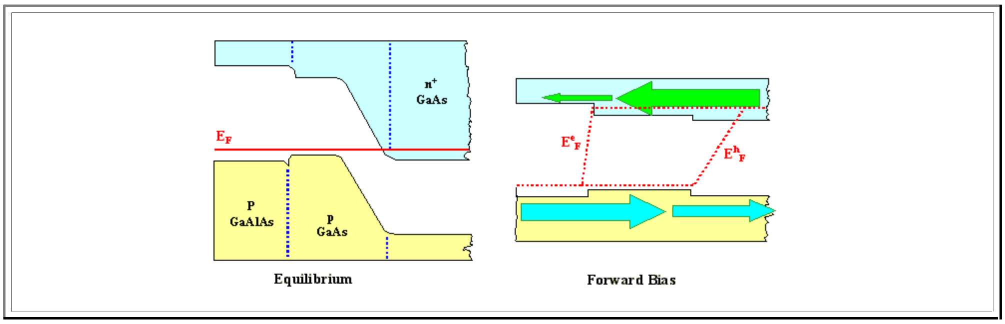


  - The active **p**-layer is confined in its lateral extension and holes and electrons are injected through a "**double heterojunction**". One will be of the diode type, the other one necessarily of the isotype.
  - As already outlined in the chapter about heterojunctions, it is possible to achieve very large injection ratios – essentially only the wide band gap semiconductor injects its majority carriers and the injected carriers can not easily escape.
  - We will look into this situation in more detail for laser "diodes".
  - All things considered, we have a considerably larger efficiency with this design and **LEDs** of this kind are sometimes called **"superradiant" LEDs**.

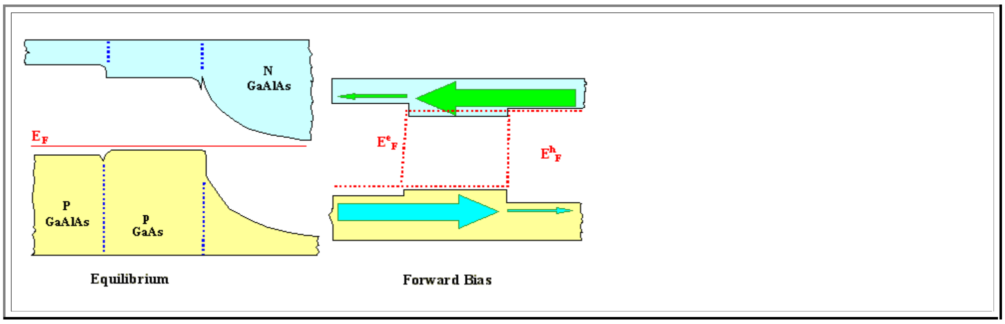
- Much more could be said about the design of **LEDs**, some special or recent developments are discussed in advanced modules.
  - Standard LED structures
  - Recent developments

7.1.3 Double Heterojunctions

- We have not paid too much attention to the injection of carriers into the active volume so far. The more simple **LED** structures shown before employed homo-junctions, i.e., simple **p–n** junctions, but while this may be cheap, it has several disadvantages
    - The active volume for the gain coefficient is ill defined. It is not simply given by the **SCR**, but by the diffusion length of the carriers. It is therefore difficult to keep the carriers in a small volume since they can diffuse away.
    - The active volume for ***u*(ν)** is not defined at all. The photons can go wherever they like in lateral directions; and this will simply not be good enough for laser diodes. Essentially, we would loose a lot of photons and the efficiency of a laser would be low.
- Homojunction lasers therefore are only usable at low temperatures. The solution, of course, are heterojunctions. They offer several general advantages:
    - We may obtain very large injection efficiencies, essentially only injecting the majority carriers from the wide-gap material into the small-gap material.
    - The carriers can be kept confined to the small band-gap material by the energy barriers due to the discontinuities of the band structure. This means that the active gain volume can be well defined.
    - If the refractive indices of the material surrounding the active volume is "right", we may achieve light confinement, too.
- The problems with heterojunctions are clear, too:
    - Only rather perfect heterojunctions will work. If you introduce misfit dislocations or other defects – forget it. This severely restricts the possible combinations of materials.
    - Optimizing all parameters at the same time with a limited choice of materials may be tricky, if not impossible. There are good reasons, after all, why there is no blue (or ultraviolet) semiconductor laser at present (the first blue ones are appearing just now = **2002**).
    - The technology for making a laser may become quite involved, meaning expensive.
- Let's look at some heterojunctions.
    - First we consider the case of a single heterojunction of the **P-p-n$^+$** type.



-
    - For the example chosen, practically all of the injected electrons from the **n$^+$** part are confined the the active **p-GaAs,** while a considerable part of the injected holes can escape (by diffusion in forward direction).
    - This will get better with a double heterojunction which necessarily must consist of one diode type and one isotype junction. The same kind of situation as above, but now with a **pN** junction diode on the right is shown below.



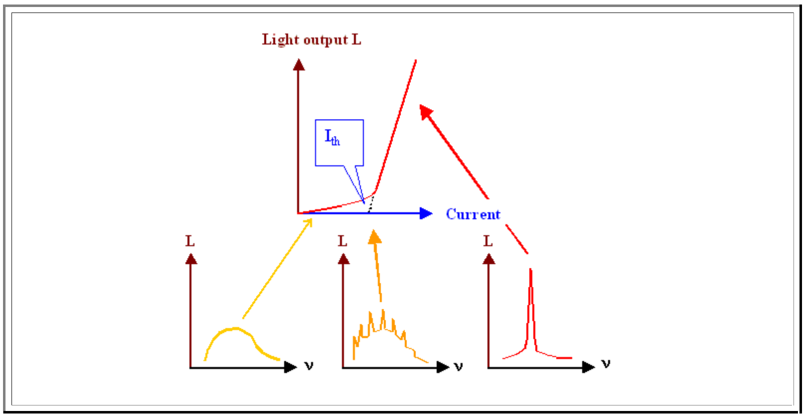
-
    - Now the escape of both injected carrier types is blocked by the band offset due to the large band-gap material on the opposite side of the active region.
    - Large efficiencies can be obtained in this way, but the technology becomes rather complicated – compare the "index-driven" laser diode shown in the following sub-chapter.

7.1.4 Optimizing Light Confinement and Gain in Laser Diodes

- Essentially, edge-emitting **LEDs** will almost automatically work as lasers, too, if simple conditions are met
  - The active region must end in *"mirror" surfaces*, which are most simply obtained by ***cleaving*** the crystal. Diamond and zinc-blende type crystals will always cleave along **{110}** planes, which contains the fewest bonds (that's how diamonds are processed, anyway). The cleavage planes are often almost atomically flat.
  - Than parts of the light will always be reflected back and the ends of the crystal act as a Fabry–Perot resonator. The reflectance ***R*** (for $n_{\text{air}} = 1$) is given by

$$R = \left(\frac{n_{\text{semi}} - 1}{n_{\text{semi}} + 1} \right)^2$$

  - With n_{semi} = index of refraction of the semiconductor ≈ **3.6**, we have ***R* = 0.32,** i.e. almost a third of the light is reflected back into the active zone.
  - Theoretically, we also should have the *total length of the active zone to be a multiple of the wave length* desired. However, since the total length is much larger than the wave length, some wave length will always "fit" and lasing will occur as soon as the gain is large enough to compensate the losses.
- We will have a rather poor laser. Nevertheless, it will show the general behavior of semiconductor lasers as illustrated below:

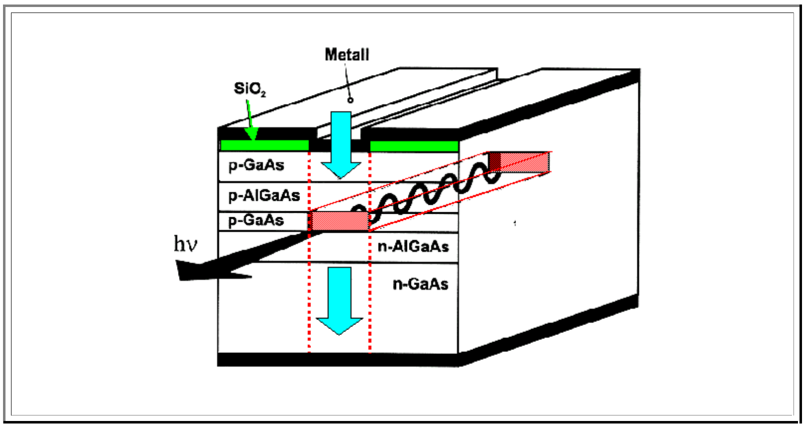


  - For currents below some threshold current I_{th}, the device will be a simple **LED** emitting light with a rather wide frequency distribution.
  - As soon as enough carriers are injected to cause sufficient inversion, some modes of the resonator with the "right" wavelengths will become amplified and appear as small peaks on the spectral distribution of the light.
  - Well above threshold, the frequency with the highest coefficient "wins" and the laser might emit only one wavelength.
- In order to have high efficiency *and* a single mode, we must maximize the density of photons [i.e., ***u***(ν)] *and* the gain coefficient [***g*** (ν)] in the *same* active area of the device.
  - This is *not* a condition automatically met – quite the opposite. The gain coefficient is mostly a function of the positions of the quasi Fermi energies, i.e., the electron densities. While the laser is in operation, it is essentially a function of the carrier injections across some junction. We cannot expect that this is homogeneous everywhere in the active region; ***g*** is thus a function of **(*x*,*y*,*z*)**, too.
  - Light is reflected, diffracted, and absorbed according to the (complex) refractive index n_r of the medium. This is foremost a function of the material itself, but also of the densities of available electrons. In regions with a high gain coefficient we have a high density of electrons, too, and thus a changed index of refraction.
  - If, in a thought experiment, we would keep the semiconductor at equilibrium (no currents, no inversion) and feed the light into the resonator from the outside, ***u***(ν) would be a function of the refractive index only. However, since we absorb and produce light by high densities of electrons in non-equilibrium, we change the optical properties of the resonator and ***u***(ν) couples to ***g***(ν) – we have rather complex conditions not amenable to simple analysis.
- Still, we have *two* basic options looking at the extremes: We can define the active area – or better, volume – by large values of ***u***(ν) *or* ***g***(ν) – irrespective of what the other quantity does.
  - If we choose ***u***(ν), all we have to do is to surround the active volume by "mirrors" on all **6** sides, a feat that can be achieved by enclosing the active volume with material that has a lower index of refraction. At the same time we make sure that in the active volume – and possibly around it – we have a large gain coefficient by injecting carriers all over the place.

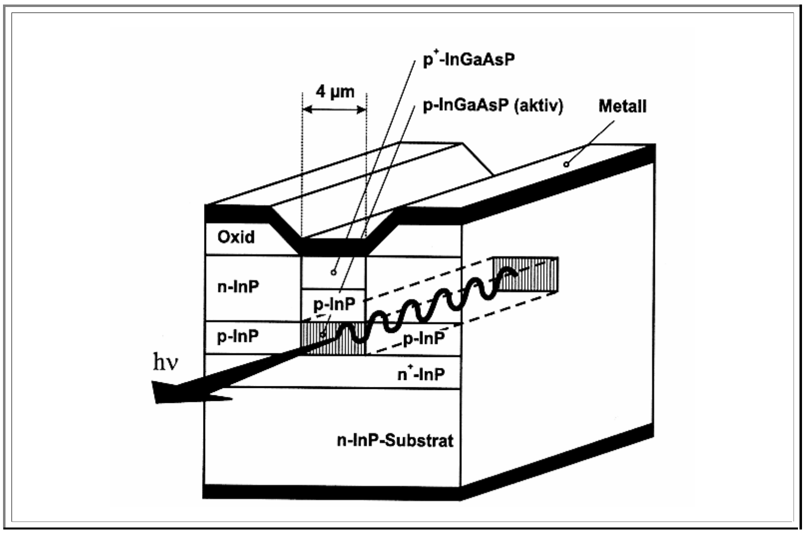
- If we choose ***g*(ν)** for the definition of the active volume – letting the light wander around wherever it likes as long as the longitudinal modes are in the active volume – we simply restrict current flow to the regions where we want it to go.

Two laser structures based on these two principles are shown below:

- First a **"gain-driven" laser**. The current is fed into only a small area by simple geometrical means as shown. Without particular attention to light confinement, we still get some positive effect, because the index of refraction in the active region is higher than in the surroundings due to the high electron density there.



- Next, an **"index driven" laser**. As you can see, it is actually also gain driven to some extent.



7.2 Specialities

7.2.1 Laser Specialities

- There would be plenty of specialities. But not here.
  - Sorry!

8. Speed

8. Speed

8.1 Some Basics to Device Speed

8.1.1 General Device Response to Input Modulation

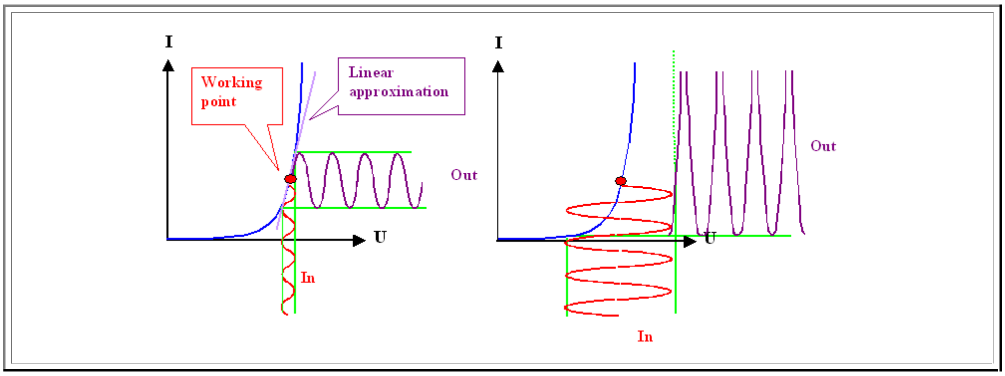
- If we take a transistor (bipolar or **MOS**), a light emitting device made form **GaAlAs**, **GaP**, whatever, a Laser diode, a simple rectifying diode (**pn**-junction or Schottky junction), i.e. just about any device made from semiconductors, we may modulate any of its input parameters (either a little bit or a lot), and see what happens to all other parameters. The paradigmatic experiments, of course are
  - ***MOS** transistors*: Modulate the gate voltage, see what the source-drain current does.
  - *Bipolar transistors*: Modulate the base current, see what the emitter-collector current does.
  - *Rectifying diodes*: Modulate the terminal voltage, see what the device current does.
  - *Solar cell*: Modulate the light flux, see what the photo current or the photo voltage does.
  - *Light emitting diodes*: Modulate the injection current, see what the light output does
  - *Laser diodes*: Modulate the pumping (i.e. the injection current), see what the light output does.
  - The list could be expanded, and many variants are possible.
- Generally, we have highly non-linear systems and a simple sinus modulation of the input parameter ***In*** in the form of

$$\boxed{In(t) = In_0 + In_m \cdot \sin(\omega \cdot t)}$$

  - Or, expressed more generally using complex notation

$$\boxed{In(t) = In_0 + In_m \cdot \exp(i \cdot \omega \cdot t)}$$

  - This simple input function, however, *in general*, will produce responses that are no longer sinus shaped, but contain higher harmonics.
- It is thus useful to distinguish between *two basic modes* of frequency responses, illustrated below for a simple rectifying diode.
  - As long as the input modulation is kept *small*, the response will be linear. This is called the **small signal response** or behavior.
  - Large modulations or signals then obviously give the (non-linear) **large signal response**.



  - Shown is the small signal response on the left, and the large signal response on the right. Note that the reponse in any case is directly given by the shape of the ***I-U*** characteristic and thus is *not* directly dependent on the frequency.
  - That, however, must be an oversimplification. At high frequencies we must expect deviations from the **1:1** correspondence of input and output via the characteristics, and it this behavior we after. Still, the distinction between small signal and large signal behavior (or linear and non-linear response) is still valid.
- The distinction between the two cases is simple: As soon as you find significant deviations in the *form* of the output signal from the from that of the input signal, you have the large signal case.
  - In other words: You have the small signal case, *if* for an input signal $In(t) = In_0 + In_m \cdot \exp(i \cdot \omega \cdot t)$ your output signal can be adequately described by

$$\boldsymbol{O} = \boldsymbol{O}_0 + \boldsymbol{O}_m \cdot \exp(\mathrm{i} \cdot [\omega \cdot t + \varphi)]$$

$$\boldsymbol{O}_m = \boldsymbol{V} \cdot \boldsymbol{In}_m$$

$$\boldsymbol{V} = \text{Amplifikation} = \mathbf{d}\boldsymbol{I} \,/\mathbf{d}\boldsymbol{U}$$

$$\varphi = \text{arbitrary phase shift}$$

  - Note that for a given input signal, it may depend very much on the choice of the the working or operating point, if you observe small signal, or large signal behavior.

- While *digital* real devices usually operate in the large signal mode (the current/voltage is either on or off with the largest possible amplitudes the system permits), we are only going to look at *small signal behavior* here.
  - Generally, we are now entering the very large world of electronic engineering and system analysis, but here we will only ask ourselves one question: What constitutes the *basic limits* of frequency response for the *paradigmatic "ideal" devices* as listed above.
  - A *real device* - always coming with wires, series resistances, parasitic capacitors and inductors, and in most cases consisting of many connected single devices - might have a quite different frequency response; but it is always determined by the frequency response of its individual elements.
  - Generally, we expect that for small frequencies ω at the input, the output will have no problem following the input.
  - Contrariwise, for high frequencies, the device will be to sluggish, and the output amplitude must decrease with frequency until there is practically no more response.

- What we are interested in are answers to the following questions:
  - What is the general (small signal) frequency response of a given basic devices as listed above?
  - What are the important factors, in particular *material properties*, that determine the maximum usable frequencies; and what kind of specific frequency response curve do we obtain?
  - What can we do about it? How can we optimize frequency response; i.e. how must we design materials and devices usable at very high frequencies?

- This is not going to be easy. There are several mechanisms that influence "**device speed**" and their combined effects may result in complex behavior.
  - In what follows we will first look at some general mechanisms that might limit device speed and than apply this to some specific devices.

8.1.2 Basic Time Consuming Processes

- The first essential point to note is that a modulation of an output signal obtained by modulating some input *always* requires a *change or modulation in some internal state* of the device.
    - And changing something always takes some time. Nothing happens instantaneously, changing something *consumes some time*. We thus may start by listing the *time consuming processes* that we already encountered.
- What kind of typical *time constants* in semiconductors did we encounter so far? Think about it for a minute. Well, we had
    - The *minority carrier life time* τ . It measures the average time that a minority carrier "lives" before it recombines with a majority carrier. It can be rather large for very clean indirect semiconductors (**ms**), and rather small for indirect semiconductors (**ns**). The numerical value of a minority carrier life time implies that you cannot change the minority carrier concentration at a frequency much larger than $\mathbf{1/\tau}$.We have a *first limit* to how fast you can change an internal state.
    - The *dielectric relaxation time* τ_d. It measures the average time that *majority carriers* need to respond to some disturbance of their distribution. It was rather small, typically in the **ps** range and given by

$$\tau_d = \frac{\epsilon\epsilon_0}{\sigma}$$

- Those were the two fundamental material related time constants that we encountered so far. But there are more time constants which are not so directly obvious:
- *First*, we have the "trivial" *electrical time constant* τ_{RC} inherent in any electrical system, simply given by the ***R · C*** product. ***R*** is the ohmic resistivity, and ***C*** the capacitance of the circuit (part) considered.
    - ***R*** and ***C*** need not be actual resistors or capacitors *intentionally* included in the system, but unwanted, nevertheless unavoidable, components. The resistivity of **Al** metallization lines together with the parasitic capacitance of this line in a **Si** integrated circuit. e.g., gives a τ_{RC} of roughly $\mathbf{10^{-9}}$ **s**, and this value (per **cm** line length) is directly determined by the product of the specific resistivity ρ of the conducting material times the relative dielectric constant $\epsilon_r$of the dielectric separating individual wires - it is thus a rather intrinsic *material property*.
    - The *physical meaning* of τ_{RC} is clear: It is the time needed to charge or discharge the capacitors in the system. Clearly, you cannot change internal states very much at frequencies much larger than $\mathbf{1/\tau_{RC}}$. And note that space charge regions, or **MOS** structures *always* have a capacity ***C***, too.
- *Second*, if we turn to Lasers for a moment, we have seen that we need to feed some of the light produced by stimulated emission back into the semiconductor by using a suitable mirror assembly.
    - Light bounces back and forth between the two mirrors in the simple system considered - and that means that even after you turned off the current through the Laser diode, some light will still bounce back and forth and thus come out until everything eventually calmed down. There is an obvious time constant

$$\tau_Q = \frac{N_r \cdot L \cdot n_r}{c}$$

    - With N_r **=** average number of reflections, L **=** distance between the mirrors, n_r **=** refective index of the material, and **c =** vacuum velocity of light.
    - If, for an order of magnitude guess, we take ***L* = 100 µm** and consider **10** reflections; the "last" photons to come out would have to travel **10 · 100 µm = 1 mm**, which takes them a time $\tau_Q = N_r \cdot L \cdot n_r/c \approx 10^{-11}$ **s = 10 ps**.
    - In other words, for the example given, it would not be possible to modulate the light intensity with frequencies in excess of about **100 GHz**. This seems to be a respectable frequency, but keep im mind that data can now (**2001**) be transmitted through fibre optics at frequncies in the **THz** regime.
- This example, while a bit far-fetched, gives us an important insight: There is a general relation between a *time constant* of a system and a *typical length* of a system mediated by the speed with which things move. This means that the *size of a device* may be important for its frequency response.
    - In other words, we can always ask: How much time does it take to move things over a distance ***l***? And whenever the output ***O*** is some distance away from the input ***In***, the question of how long it takes to move whatever it takes from ***In*** to ***O*** produces a typical time constant of the system.
    - In straight-forward simple mechanics ***l*** is linked to its time constant τ_l by the *speed* of the moving "things" - for the photons considered above this was clearly the speed of light (in the medium, to be correct).
    - For our moving statistical ensembles, we have somewhat more involved relations, e.g. .

$L = \left(D \cdot \tau\right)^{1/2}$	for the relation between the diffusion length of the minority carrier and their lifetime
$L_{Dn} = \left(D \cdot \tau_d\right)^{1/2}$	for the relation between the Debye length L_{Dn} and the dielectric relaxation time

- What are the moving things? Well, besides photons, we essentially are left with electrons and holes; everything else that might be of interest is usually immobile (dopants, localized excitons), or so slow that it should not matter for *electronic* signals (phonons, mechanical movements (e.g. vibrating parts) in **MEMS** devices)
  - This brings us to a first simple and important question: How long does it take electrons or holes to move from the source to the drain in a **MOS** transistor. Clearly, this will give us another maximum frequency for operating said transistor.
  - The relevant velocity in this case is the **drift velocity v_D** of the carriers, usually proportional to the field strength ***E*** as driving force for the movement, and better expressed via the carrier mobility

$$\mu = \frac{v_D}{E}$$

  - With the source-drain distance l_{SD} , and the source drain voltage U_{SD} , we have $E = U_{SD}/l_{SD}$ and a "travel time"

$$\tau_l = \frac{l_{SD}}{v_D} = \frac{l_{SD}^2}{\mu \cdot U_{SD}}$$

  - To get a feeling for orders of magnitude, we take a source-drain distance l_{SD} **= 1 µm** and a source-drain voltage U_{SD} **= 5V**, obtaining a field strength of $E_{SD} = 5 \cdot 10^4$ **V/cm**. Typical mobilities are μ_{Si} **= 1000 cm²/Vs** for **Si**. This gives us a drift velocity of

$$v_D = 1000 \frac{\text{cm}^2}{\text{Vs}} \cdot 5 \cdot 10^4 \frac{\text{V}}{\text{cm}} = 5 \cdot 10^7 \frac{\text{cm}}{\text{s}}$$

  - Is that a large or small velocity? It might be good to look up at an old exercise at this point
  - The "travel time" τ_l then is

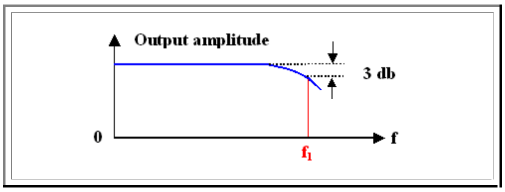
$$\tau_l = l_{SD} \cdot v_D = \frac{10^{-4}}{10^7}\,\text{s} = 10^{-11}\,\text{s}$$

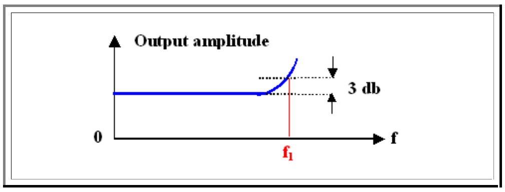
  - A "**1 µm** " **Si MOS** transistor thus would not be able to switch frequencies beyound about **10^{11} Hz = 100 GHz** *if* τ_l would be the only limiting time constant of the system.
- Last, there are some ultimate limits that we should be aware off:
  - Nothing moves faster than **c**, the the speed of light (in vacuum). The consideration for the Laser from above already gives an example for this limit.
  - The movement of electrons and holes has some intrinsic constant of its own: The *average time between scattering processes* and the average distance or mean free path in between. While we are not very aware of the values for these parameters, the mean free path is in the order of **100 nm**.
- This has an important consequence: We only can use *average* quantities like drift velocities, if individual carriers could have many collisions.
  - Turning this around implies: If we look at travel scales around and below **100 nm**, everything may change. For transistors this small, electrons (or holes) might just speed from source to drain without any collisions in between - much faster than at larger distances. This is the case of **ballistic carrier transport** which must be considered separately.

8.2 Dynamic Behavior of pn-Junctions

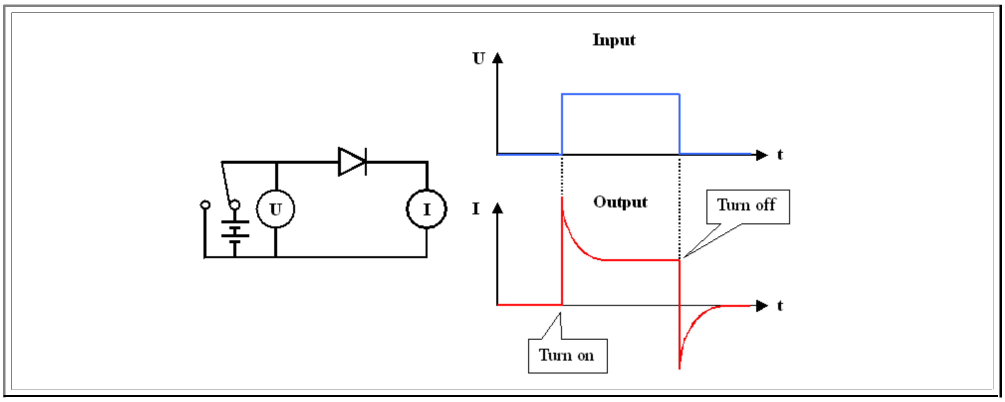
8.2.1 General Observations

- First, lets consider the dynamic behavior of a simple **pn**-junction.
  - Whereas the *small* signal behavior at low frequencies is exactly that of the example given before, the *large* signal response in practice is simply what we (and everybody else) calls the **switching behavior** of the diode. In other words, we are looking at either
    - Suddenly switching the voltage from **0 V** to some forward value U_f , or
    - Suddenly switching the current from **0 A** to some constant forward value I_f .
- In order to have some idea of what we want to find out, here are some typical results of experiments with modulated inputs. First, we look at the *small signal* behavior by plotting the amplitude of the *current output* versus the amplitude of the *voltage input* as a function of the frequency. What we will find looks like this:

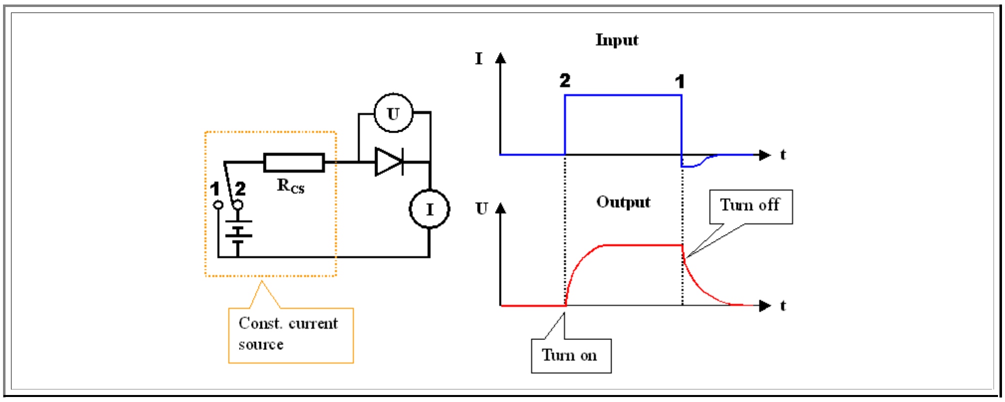




  - While real curves may look quite different, we always will find that with increasing frequency the output amplitude will eventually come down or go up. Often, the frequency where the output value is **3 db** below or above its low frequency value is identified as some limit frequency f_l, giving us a time constant $\tau = 1/f_l$ which we want to understand in terms of materials properties.
- Now we look at the *switching behavior for the voltage*, i.e. we suddenly switch off the voltage across the diode from **0 V** (or some reverse voltage) to some forward voltage, e.g. **0,7 V**.
  - This is experimentally easy to do and to measure. What we obtain looks like this:



  - We have pronounced current "**transients**" describable by one (or possibly more) time constants; and we want to know how these time constants relate to materials properties of the **pn**-junction.
- Next, we *switch the current* "suddenly". The question is how to do this. In formal electrical terms, we just take a "constant current source" and turn it on. In reality this might be a voltage source with a resistor R_{CS} much larger than anything else in the circuit. The current than is simply U/R_{CS} and switching is obtained by switching the voltage which now could be rather large.
  - If we do the experiment, we will observe the following behavior:



- Again, we have **transients** in the diode behavior, this time the voltage is affected. But we also have a kind of transients in the current - we will not be able to switch it off abruptly, but for some time a **reverse current flow** will be observed
- Since the current pulse was "made" by a (rectangular) voltage pulse and the external voltage is zero after the switching, we are forced to conclude that the diode acts for some time as a *voltage/current source* after the primary voltage was turned off. More details can be found in an advanced module

Once more we measure some time constants and we must ask ourselves how they relate to the time constants of the voltage switching experiment and to material parameters.

Accepting all these "experimental" findings, we even must conclude that there might be *several* time constants which, moreover, might depend on the particulars of the experiment, e.g. the working point, or the amplitudes of the input signal.

- This does not just *look* a bit involved; it really *is*. And the "ideal" **pn**-junction diode is about the most simple device we have.

We are going to look at it in some detail in the next subchapters, even so we are not particuarly interets in plain **pn**-junctions. But this will help to develop some basic understanding of the underlying processes and make the dicussion of more involved devices easier.

8.2.2 Small Signal Response of p-n Junctions

Equivalent Circuit Description

- If we pretend for a moment that we are hard-core electrical systems engineers, we do not care at all about what a **pn-**junction diode consists off, how it works, or how it is made. We simply describe it as a **black box** and use the two equations from before, but now with a possible *frequency dependence* thrown in for the output current. For small signal behavior we have by definition

$$\begin{aligned} U(t) &= U_0 + U_m \cdot \exp(i \cdot \omega t) \\ I(t) &= I_0 + I_m(\omega) \cdot \exp(i \cdot [(\omega t) + \varphi(\omega)] \end{aligned}$$

  - i.e. we consider a frequency dependent amplitude $I_m(\omega)$ of the current output signal and a frequency dependent phase shift $\varphi(\omega)$.
- The relation between U_0 and I_0 is simply given by the (**DC**) current voltage characteristics of our black box.
  - In principle, the output signal amplitude (i.e. the current amplitude) is also a function of U_0 or I_0; in other words of the **working point** chosen. All we have to do then is to tabulate (or represent graphically) the functions $I_m(\omega, I_0)$ and $\varphi(\omega, I_0)$ for our black box.
  - This is all. We always have an output signal looking exactly like the input signal, and with an amplitude proportional to the input amplitude because we have *by definition* a linear system.
- We know even more. We always will have

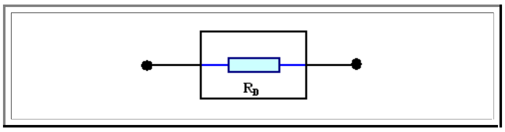
$$I_m(\omega) = \frac{dI}{dU} \cdot U(\omega)$$

$$\varphi = 0$$

  - as long as the output signal can be directly obtained from the **DC** characteristics, i.e. at *small* frequencies (leaving open at present what "small" means in numbers).
- Looking at our diode (or the small signal (i.e. linear) behavior of about anything else) in this way has a *large advantage*:
  - *All* black boxes can now be described by a suitable network of *basic linear electronic elements* - resistors, capacitors and inductors. In the most simple case all elements have constant values; more realistically their value depends on external parameters (e.g. the voltage).
  - Calculating the behavior of such a network is relatively simple - for a computer, that is.
- However, just looking at some network or **equivalent circuit diagram** as it is called, has *disadvantages* too:
  - Many different kinds of networks will give exactly the same response; their is no *unique* choice - but often *intelligent* choices.
  - If some network does behave like the actual device, it is not necessarily easy to extract the important physical parameters of the device from it. In other words, if some network behaves exactly like the diode you are interested in, it does tell you *how* you must change resistors and capacitors if you want to make it faster, but not which physical parameters of the device (doping, lifetime, dimensions, ...).
- The best way therefore is to construct an equivalent circuit - keeping it as simple as possible - by looking at the physical characteristics of the real device first. We are going to do this now for our junction diode.

Equivalent Circuit Construction for a Junction Diode

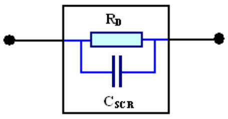
- If we start with the low frequency behavior of a junction diode, the equivalent circuit diagram is exceedingly simple:
  - It consists of a single resistor R_D with a value that depends on the voltage and is given by $R_D = dU_{diode}/dI_{diode} = R_D(U)$, i.e. by the (inverse) slope of the **DC** characteristic of the diode as illustrated before. This is shown below



  - This "network" cannot possibly give *any* frequency dependence, so it can not be all there is to a diode. We now must remember that any **pn**-junction has a **capacitance** C_{SCR} associated with the space charge region. It arises from the **ionized dopants** that provide the net charge on both sides of the junction needed for any capacitor and its value (for a symmetric junction) is given by

$$C_{SCR} = \left(\frac{2e \cdot \epsilon\epsilon_0 \cdot N}{U}\right)^{1/2}$$

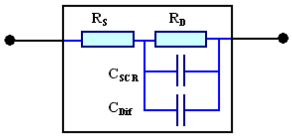
  - C_{SCR} is obviously switched in parallel to R_S; our equivalent circuit diagram now transforms into this:



- This network will give us a simple frequency response: High frequencies are essentially short-circuited by the capacitor; and the current amplitude will increase with frequency (accompanied by a phase shift of **90°** at the maximum).
  - The current amplitude will have increased twofold if the **AC** resistance R_C of C_{SCR} equals R_D , this happens for

$$R_C = \frac{1}{\omega \cdot C_{SCR}}$$

- With our formulas for $I = I(U$**, doping, ...)** and for $C_{SCR} = C_{SCR}$ **(***U***, doping, ...)** , we could calculate the limit frequency as a function of the prime material parameters like doping, lifetime and so on - but we are not yet done in constructing the equivalent circuit diagram:
  - There is a *second capacitance* hidden in the junction diode called "**diffusion capacitance** ". Lets see what it is.
  - *Any* net charges to the left and right of the junction form a capacitance. Above, we considered the space charge region capacitance stemming from ionized dopants in the **SCR** which are not compensated by majority carriers as we have it in the bulk. For diodes biased in *reverse* direction, i.e. with only a small reverse current flowing, this is all there is.
  - But if we look at the distribution of *all* carriers for a diode in *forward* direction, we notice that we have an *excess of minority carriers at the edge of the SCR*; a schematic picture was given before.
  - These excess minorities - only present while current flows - are coming from the *injection of majority carriers* over the potential barrier which become minorities in the other part of the junction. Their concentration is increased because a concentration gradient is needed to remove these carriers by diffusion. They are *not* compensated by other charges and thus form a **diffusion capacitance** C_{Dif} that increases with current.
  - This diffusion capacitance must also be switched *in parallel* to R_D and C_{SCR} .
- Moreover, the ohmic resistance of the bulk **Si** is not zero, but has some finite value R_S which is constant and easily determined by the resistivity of the **Si** and all other ohmic resistors in the circuit.
  - Clearly, R_S must be switched *in series* to everything else, and we obtain the final equivalent circuit diagram of a junction diode valid for all cases.



  - Of course, you always could lump together the two capacitors into just one, but then you loose the connection to physical reality. Note that their numerical values will depend on physical parameters in very different ways - as we will see below.
- If we would know the value of C_{Dif} , we could now calculate the small signal response of a junction diode. We have no ready formula, so we must derive one.
  - We will do that right away; it will prove to be particular enlightening for the understanding of the correspondence of *primary material parameters* and their transformation to *equivalent circuits*.

The Diffusion Capacitance and its Relation to Time Constants of the Device

The excess Q_{min} at to the left and right of the space charge region increases with increasing (forward) voltage U. If we have a linear relation between U and Q_{min}, the capacitance C_{Dif} would be given by

$$C_{Dif} = \frac{Q_{min}}{U}$$

- Since this is not necessarily the case, C_{Dif} is a function of U and must be defined *differentially* as .

$$C_{Dif} = \frac{dQ_{min}}{dU}$$

- Considering that the current I is a function of U (given by the basic diode equation), we may rewrite this expression as

$$C_{Dif} = \frac{dQ_{min}}{dU} = \frac{dQ_{min}}{dI} \cdot \frac{dI}{dU}$$

- The second term is simply the derivative of the ***I-U*** characteristics, and thus equal to the inverse diode resistance.

dQ_{min}/dI is the interesting term. This differential quotient must have the dimension of time, and thus can be seen as the *time constant* connected to C_{Dif} . In order to compute it, we need Q_{min} as a function of the current flowing through the junction, i.e.

$$Q_{min} = Q_{min}(I)$$

- We have not encountered this functional relationship so far - but we came close.

In the "Useful Relations" subchapter we have an equation giving the excess minority carrier density $\Delta n(x)$ as a function of the distance x from the edge of the depletion layer:

$$\Delta n(x) = \Delta n_0 \cdot \exp - \frac{x}{L}$$

- With Δn_0 = excess density at the edge of the depletion zone, L = diffusion length.
- In the "Junction Reconsidered" subchapter we derived an equation relating the current density j (for one of the two current components) through the junction to the excess carrier density Δn_0 at the edge of the depletion layer:

$$j = \frac{e \cdot D}{L} \cdot \Delta n_0$$

So all that remains to do is to express the charge Δn_0 at the edge of the **SCR** as a function of the total excess charge Q_{min}, substitute it in the current equation, and do the differentiation dQ_{min}/dI. This is easy:

- Q_{min} is simply the integral over $e \cdot \Delta n(x)$ taken from $x = 0$ to $x = \infty$; i.e.

$$e \cdot Q_{min} = \int_{x=0}^{x=\infty} e \cdot \Delta n(x) \cdot dx = \int_{x=0}^{x=\infty} e \cdot \Delta n_0 \cdot \exp - \frac{x}{L} \cdot dx = e \cdot \Delta n_0 \cdot L$$

- Inserting $\Delta\, n_0 = Q_{min} / L \cdot e$ in the current equation from above gives

$$j = \frac{e \cdot D}{L} \cdot \frac{Q_{min}}{e \cdot L} = \frac{D \cdot Q_{min}}{L^2}$$

- The differentiation (turning to current densities j instead of currents I and interpreting Q_{min} as charge density) finally yields

$$\frac{dQ_{min}}{dj} = \frac{L^2}{D} = \tau$$

And this, of course, is nothing but the minority carrier life time τ!

- The time constant associated with the diffusion capacitance thus has a very clear physical meaning - it is the minority carrier life time. It is, if you like, the fundamental property that *causes* the diffusion capacitance.
- Interpreted in other words: While separated charges Q always cause a (static) capacitance C given by $C = Q/U = \epsilon\epsilon_0 Q \cdot A/d$, you always will find a (dynamic) capacitance, too, if it *takes time* to change charge concentrations. The capacitance associated with some *given* time constant τ than is more appropriately expressed as

$$C_{dyn} = \frac{\tau}{R}$$

- With R being some ohmic resistor which limits current flow.

Now we have all terms; the diffusion capacitance can be finally expressed as

$$C_{Dif} = \frac{dQ_{min}}{dI} \cdot \frac{d\,I}{dU} = \frac{\tau}{R_D}$$

Finally, we look at the diffusion capacitance from yet another angle: It results from the necessity to "store" some charge on both sides of the junction for forward current flow.

- The concept of **stored charge** and the time constant associated with its removal will come up again later.

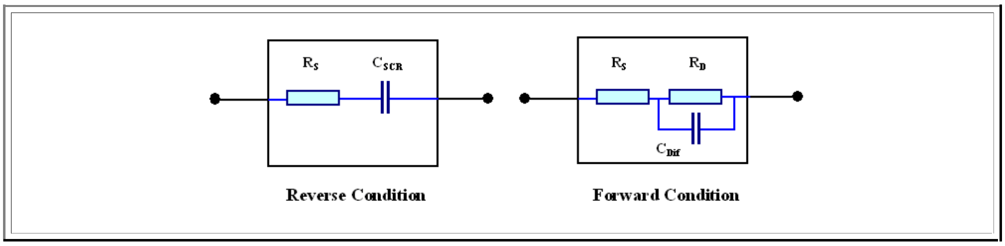
Consequences

We now have all ingredients to analyze the small signal response of a **pn**-junction - as long as it is ideal, infinitely extended, and symmetric.

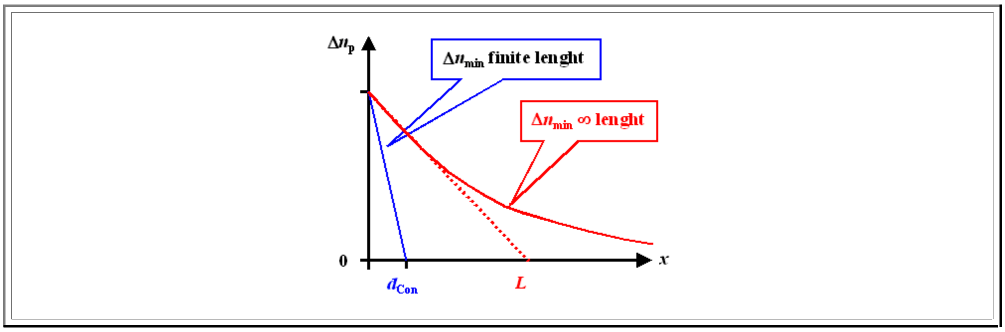
The first question coming to mind is the relation between the two capacitances. Electrically, they add up - but which one dominates?

- This is most easily seen if we consider the forward and reverse current direction separately.
- The *reverse* current mode is easy: The diffusion capacitance is small because dI/dU is practically zero and can be neglected. We only will find the junction capacitance C_{SCR}. Moreover, R_D is very large, certainly much larger than R_S , and may be taken to be infinite.
- In *forward* direction, C_{SCR} increases with $U^{1/2}$ while C_{Dif} increases *exponentially* with U (look at the formulas!). This means that in forward direction C_{Dif} will always win and dominate the total capacitance.

We thus can simplify the equivalent circuit diagrams for the two extremes:



- This means that the small signal behavior of a **pn**-junction is quite different in reverse or forward direction, and that it also depends very much on the working point (the **DC** current) in forward direction. It also means that we only can transmit **AC** signals in reverse direction.
- If we now try to deduce some numbers for limiting frequencies of real diodes, we run into problems:
  - The diffusion capacitance will short-circuit all signal with frequencies considerably larger than **1/τ**. Since the minority carrier life time in good **Si** can be as large as **1 ms**, we must expect problems in the **kHz** region **-** and even for life times of **10 µs** we would not get far beyond the **MHz** region. This is obviously *far below* limiting frequencies in actual devices.
- *What went wrong*? Not so easy to unravel, lets just consider the most important point:
  - Dimensions in real (**Si**) devices, especially integrated circuits, are *much, much smaller* than the diffusion length ***L***. For the excess minority carrier distribution as shown for a large diode, this has profound consequences.
  - The excess concentration must go down to zero at the contact, i.e. after a distance $\boldsymbol{d_{Con}}$ that gives the length of the (almost) neutral piece of **Si** (the distance from the edge of the space charge layer to the contact). In a *first approximation* we have to replace the diffusion length ***L*** (which is in the order of magnitude of **100 µm**) by $\boldsymbol{d_{Con}}$, which can be a fraction of a **µm**.
  - The minority carrier excess density now decays rapidly and reaches zero at the contact, i.e. after a distance $\boldsymbol{d_{Con}}$. For all practical purposes it is sufficient to assume a linear relationsship as shown below. In practice, the difference between ***L*** and $\boldsymbol{d_{Con}}$ is much larger.



  - Accordingly, we have to replace the bulk lifetime **τ =** $\boldsymbol{L^2/D}$ by

$$\tau_{\text{tran}} = \frac{(d_{Con})^2}{D} = \textbf{base transit time}$$

  - This is something we have encountered before by looking at real diodes.
- A small device thus could be faster by several orders of magnitude.
  - And this is where, quite generally, the *dimensions* come in. We always need to move carriers from here to there, and this takes some time determined by the distance and the velocity (or mobility), respectively, as pointed out before.
  - For our case of *small diodes*, we may use the equation from before in the form

$$\tau_{\text{tran}} = \frac{(d_{Con})^2}{\mu \cdot U}$$

  - This tells us that the small signal frequency response of a small junction diode is ultimately controlled by the dimensions of the device and the mobility of the carriers that carry the current.
- Of course, *real* small devices are even more complicated. But all kinds of parameters neglected in this simple consideration will make the frequency response worse, not better. The absolute limits are always determined by dimensions and mobility.

A few last words of caution:

- While we replaced the (bulk) lifetime by a transit time in this simple consideration of the real device world, it would be premature to conclude that the life time is of no importance in small devices.
- It also would be premature to conclude that we should reduce τ as much as possible, because this carries heavy penalties in other areas - see the links 1, 2,
- However, that does not mean that "*lifetime killing*" is not used on occasion. For some devices - particularly (large) power devices - some **Au** might be diffused into a junction to reduce the life time as much as possible.

9. Compound Semiconductor Technology

9. Compound Semiconductor Technology

9.1 General Remarks

9.1.1 Major Differences to Silicon Technology

General Remarks

- In this lecture course it is assumed that you are tolerably familiar with main stream **Si** technology.
  - If you are not really acquainted with that subject, or you forgot most or parts of it, turn to the "**Electronic Materials Hyperscript**", specifically to chapter 4 and chapter 5.
  - We will not cover any part of compound semiconductor technology that is essentially identical to **Si** technology. This includes in particular lithography, ion-implantation, basic chemical vapor deposition, sputtering techniques, or basic plasma etching.
  - Instead we will focus on problems and solutions that are unique to the non-**Si** world. Of course, the backbone part will only contain a rather superficial glance on this subject - the details fill not only libraries but are developing at a fast pace.
- In this subchapter we will try to get a rough overview of the more general properties of compound semiconductors that we have to deal with.

Starting Material

- Huge **Si** crystals with diameters of **400 mm** weighing up to **250 kg** can be routinely grown - absolutely free of "large" defects like dislocations or precipitates, *and* with impurity levels in the low ***ppt*** (= parts per trillion) range.
  - This feat is simply not possible for compound semiconductors. This is not for lack of trying, but for *basic* thermodynamic and mechanical reasons - and this means that this deplorable state of affairs is not going away with more research and experience.
  - While the **GaAs**, **GaAlAs**, **GaP**, **InP**, ... (and so on) crystals will certainly become increasingly better with more and more applications (and thus money for crystal growth research and development), we may safely assume that no compound semiconductor crystal will ever get close in quality to a **Si** crystal. This sad fact, of course, limits some applications, while others are less affected.
- To understand the basic limitations, lets consider the growth of a **GaAs** crystal. Most problems are typical for its relatives, too.
  - We cannot just melt some **GaAs** and do some crystal pulling. As has a very high vapor pressure at the melting temperature (**1238 ^{o}C**) and would just evaporate off, leaving a **Ga** rich melt. And a **Ga** rich melt solidifies not far from room temperature into **GaAs** and **Ga** - until then it consists of *solid* **GaAs** with *liquid* **Ga** inclusions).
  - It is even worse for **GaP** or **GaN**. Whatever you do for crystal production, you must do it in a *high pressure* environment to keep the group **V** elements in the system.
  - Next, the **III-Vs** are generally much "softer" than **Si**, i.e. the have a smaller yield point. Compared to **Si**, small mechanical stresses are enough to cause plastic deformation - even at low temperatures. Since thermal gradients are *always* linked to mechanical stress *and* absolutely unavoidable during any kind of crystal growth, it is much harder to avoid dislocations in **III-Vs** than it is in **Si**. In fact - *it is impossible*.
  - Finally, any deviation from perfect stoichiometry must result in defects by necessity - the surplus atoms of whatever kind are included as point defects, agglomerates or precipitates. Compared to **Si**, where the density of intrinsic point defects is below **10^{-6}** even close to the melting point, you would have to have the mixture of the two elements precisely at **(1 : 1) +/- 10^{-6}** to achieve comparable point defect levels.
- Still, you can buy **GaAs** wafers with diameters of **150 mm** and dislocation densities below about **10^3 cm^{-2}** , a remarkable achievement of the material scientists involved. How this is done will be described in subchapter 9.2.2

Oxides

- **Si** microelectronics owes as much to **SiO_2** than to perfect **Si** crystals.
  - Not only is **SiO_2** a near perfect dielectric for many applications (its dielectric constant is large enough to make **MOS** transistors and capacitors efficient, but not so large as to produce large signal delays because of parasitic capacitors; its break-through field strength is among the highest measured; it is chemically extremely resistant yet easily etched in special chemicals and in plasma, and its interface properties are exceedingly good), but it can be produced in extremely good quality simply by thermally oxidizing the **Si**.
  - For compound semiconductors, however, we have a simple rule: *Forget about **III-V** oxides*
  - They are neither "good" (they may dissolve in water), nor can they be produced by oxidizing the crystal (there are always exceptions, of course).
- This doesn't make compound semiconductor technology easier! While there are many ways to get around this problem, it still is a problem.

Doping

- Semiconductor technology depends on being able to dope the crystal *both ways*: **p**-type conductivity and **n**-type conductivity is needed for most, if not all applications. Doping a semiconductor requires that *two* conditions are met:
  - *First*, dopants must exist - or more generally defects - that introduce electronic energy levels $\boldsymbol{E_{\text{dop}}}$ in the bandgap that are *very close* to the band edges. The difference in the energy levels, $\boldsymbol{E_C - E_{\text{dop}}}$ and $\boldsymbol{E_{\text{dop}} - E_V}$, must be comparable to **kT** (about **40 meV** at room temperature), otherwise the transfer of electrons between the bands and the dopant level does not occur at room temperature.
  - *Second*, the Fermi energy must be able to move freely **-** in must not be "pinned" by defect states. In other words, the energy states of the intentionally introduced dopants should be the only, or at least the dominating ones. If there are a lot of energy states stemming from defects like dislocations, precipitates, interfaces and so on, your Fermi level will be "pinned" to these states and doping is simply not possible or severely restricted.
  - In **Si**, this is not a problem. In many compound semiconductors, however there are big problems with this, especially for the **II-VI** compounds. As an example, it was the impossibility to produce ***p**-type doping* in *GaN* that prevented the use of this "blue" direct band gap material until recently.
- Doping thus is often a problem, and solving it in some hitherto unused semiconductor material is always a major break-through in materials science.

Contacts

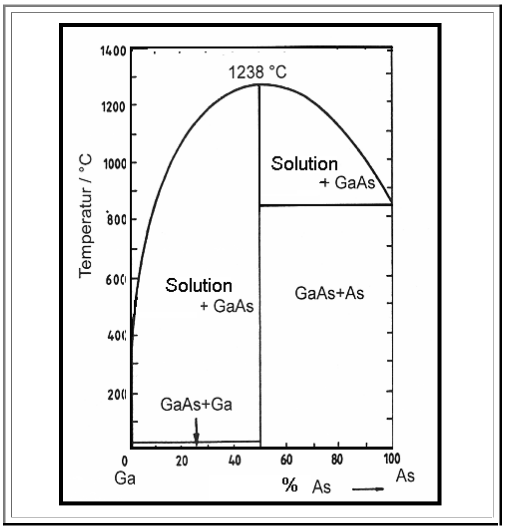
- Any device needs at least two **ohmic contacts** for electrical communication with the "outside" world. An ohmic contact must meet two requirements:
  - Its $\boldsymbol{I(U)}$ characteristics must be linear for both polarities (This defines "ohmic").
  - The ohmic resistance, $\boldsymbol{dU/dI = U/I}$, must be sufficiently low so that the voltage drop across the contact is negligible.
- Just putting some metal on a semiconductor does not necessarily produce a contact at all, you also must have some intimate interaction - usually you must heat the system somewhat. Otherwise you have a semiconductor - dirt/oxide - metal contact; mostly not what you want. Now consider the following list of **GaAs** features:
  - **GaAs** decomposes into **Ga** and **As_2**-gas at about **580 ^{o}C**.
  - The Schottky barrier height of **n-GaAs** is always large - about **0,8 eV** for any metal.
  - It is difficult to dope **n-GaAs** to a carrier concentration $\boldsymbol{> 5 \cdot 10^{18}\ \text{cm}^{-3}}$ (larger concentrations would bring down the Schottky barrier height).
  - Annealing at temperatures up to **850 ^{o}C** for **30 min** is needed to activate implanted dopants.
- Put together, the list essentially says that there are no good ohmic contacts to **n-GaAs**. Still, we have working devices, so compromise must be possible. But metallization is a big issue for many compound semiconductors.

9.2 Bulk Crystals

9.2.1 GaAs

Basic Properties

- Producing **GaAs** crystals must starts with a consideration of its phase diagram.
  - Here it is. It is already sufficient to show that there is only a very small region where you can get solid and stoichiometric GaAs, essentially a line. Small deviations to the left or right will produce some liquid encasements - right after solidification and sone Ga or As related defects after complete solidification.



  - If you think you could avoid or a least minimize those defects (that cannot possibly be good for a device) by melting a perfect 50 : 50 mix of Ga and As, you must think again. Ga will start to evaporate out of your mix as soon as it melts, changing the compositions. and son on.....
- The message should be clear: It is far more difficult to produce a defect-free **GaAs** crystal than it is possible for **Si**. It is actually impossible. And that is true for all compound semiconductors.
  - The problems with III-V technology start right here!

--- To be continued (or possibly not) ---

9.3 Epitaxial Layers

9.3.1 General Remarks

Epitaxial layers play a very big role in compound semiconductor technology, while in **Si** it was only used for the bipolar technology. **MOS** and **CMOS** technologies for many generations could do with the (for **Si** expensive) epitaxial layer deposition techniques. In compound technology it is different for several reasons:

- The substrates are far less perfect than **Si** single crystals, and epitaxial layers usually have a better crystal quality than the substrates (this was also a main driving force behind **Si** epitaxy in the early days of **Si** technologiy in the seventies).
- There are more (and sometimes relatively cheap) methods to grow epitaxial layers of compound semiconductors.

10. Specialities

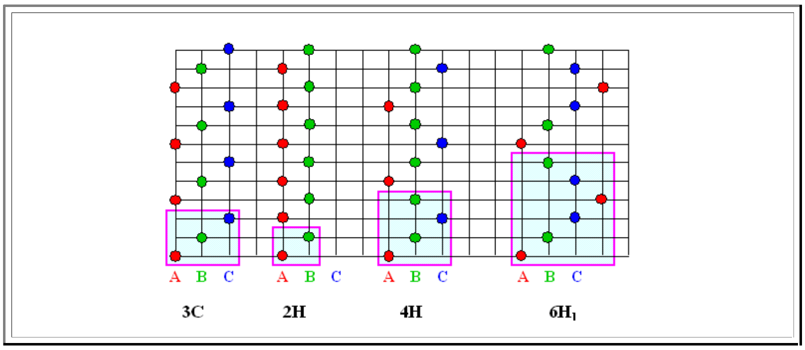
10. Specialities

10.1 Silicon Carbide

10.1.1 Silicon Carbide - Material Aspects

Structure and Basic Properties

- There is no such thing as plain **SiC**!
  - Instead, whenever you look in the literature, you will find names like **3C-SiC**, **6H-SiC**, **4H-SiC**, or **2H-SiC**. In other words: There are many different **polytypes** of **SiC**.
  - **Polytypism** is a special case of **Polymorphism**, which means that a given element or compound can assume more than one crystal structure. Polytypism simply is the one-dimensional variant of polymorphism.
- **SiC** (unfortunately) is sort of the paradigmatic material for polytypism. The always identical hexagonal two-dimensional **SiC** layers can form many crystal structures by different ways of stacking the layers on top of each other - that's why it is one-dimensional. See "MatWiss I" (in German) for the basics of how crystals can be formed by stacking atomic layers.
  - **SiC** does not just have a few polytypes, it has more than **200**! Now you have a problem: which one is the best for the application you have in mind, and if you know that, can you actually make it all by itself (and not in a mix with all the others)?
  - **SiC** is also the only stable group **IV - IV** *compound* semiconductor. No other combination of the elements **C**, **Si**, **Ge**, **Sn** exists in a defined lattice (and not just as mixed crystal like **Si-Ge**).
  - All polytypes have a rather *large* indirect bandgap and other properties, which make **SiC** a very interesting material for many applications.
  - As you may have guessed, **SiC** is devilishly difficult to grow as a (large) single crystal of *one* polytype with low defect density. **SiC** actually boasts a particular (and very bad) lattice defect all of its own - so-called micropipes - the likes of which have not (yet) been found in other crystals.
- The formal way of identifying polytypes, ie.e. the nomenclature of the polytypes, is explained in the link (basic module); here we just look at the more important variants in terms of the familiar "**ABC**" stacking definition. The basic building block (not necessarily a unit cell) is highlighted in light blue.



  - Upon contemplation, you should be able to notice that the "**3C**" structure is nothing but the **ABC** stacking sequence of a close-packed **fcc** lattice; the **2H** is the corresponding simple **hcp** structure resulting from an **AB** stacking sequence.
  - Your guess then that **C** stands for "cubic"; **H** for hexagonal, is correct. If an "**R**" comes up in variants not shown here, it stands for rhombohedral.
  - In some older nomenclature, cubic **SiC** is also known as **β-SiC**; the hexagonal phase (**6H-SiC** more or less) is the **α-SiC**
- Not only the structure of **SiC** polytypes is different, but so are their **electronic properties**.
  - The always *indirect* band gap varies from **2.4 eV** for the cubic **3C-SiC** to **3,3 eV** for the simple hexagonal **2H-SiC** variant. Other relevant parameters like carrier mobility might be quite different, too. Some values (mostly adpated from the publications or presentations of Erlangen (Germany) **SiC** group) are shown below

		4H-SiC	6H-SiC	15R-SiC	3C-SiC
Band Gap [eV]		3.265	3.023 3.03	2.986	2.390
Lattice Constant [Å]	**a**	3.08 3.073	3.08	3.08	4.36
	c	10.05	15.12	37.70	-
Effective Mass [m_c]	m_e	0.37	0.69	0.53 - 0.28	0.68 - 0.25
	m_h	0.94	0.92	-	-
Mobility (@ 300K) [cm^2/Vs]	μ_e	500	300	400	900
	μ_h	50	50	-	20
Thermal conductivity (RT) [W/cm · K]		3.0 - 3.8	3.0 - 3.8		

- If you look at the table long enough, you should now actually have a question!

Anyway, besides the rather large bandgap, the effective masses and the mobilities are not so remarkable compared to the more standard semiconductors.

- However, if you compare on a more specific level, there are definite advantages. Activate the link if you are interested.

Crystal Growth, Wafers and Defects

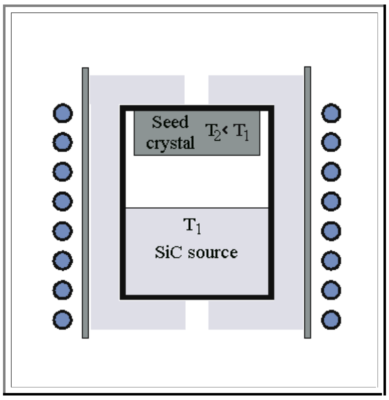
Here we will only look at the basics; details are left to an advanced module.

In most cases, large single rystals are grown from a melt (e.g. Silicon) or some solution (e.g. quartz, or sugar if you leave you coffee cup around too long), but this is not a feasible option for **SiC** single crystal growth since **SiC** does not have a liquid phase under normal conditions (i.e. without applying a large pressure). **SiC** is also extremely hard (close to diamond) and therefore has a high melting point (or is it the other way around?).

- There is also, in principle, no crucible material that could contain molten **SiC** at is nominal melting point temperature of **< 2.500 ^{o} C**. Nevertheless, **SiC** was grown from a melt at **2200 ^{o}C** and **150 bar** in a recent study, but this is probably not a commercially viable process.

We need a basically new method of crystal growth. Some "older" techniques are described in the link, the main method used nowadays is **physical vapor transport** (***PVT***) also known as *seeded sublimation growth* or *modified Lely method*.

- A piece of **SiC** is heated to **(1800-2600) ^{o} C** at low pressure. Due to the high sublimation rate, **SiC** vapor forms and deposits itself on a cooler single-crystalline seed crystal.
- Straightforward and basically simple, as shown in the schematic picture on the right.



However, pondering the situation, some questions should come to mind:

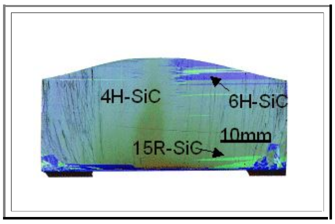
What materials can you use for the crucible and everything else that gets hot? After all, not many materials can cope with temperatures above **2000 ^{o}C**!

- Well, you are basically stuck with graphite, and maybe a bit of **Ta** here and there. That means, of course, that you are forming **SiC** also on your crucible walls and everywhere else. If it flakes off, you will have a defect problem.

What kind of *growth rate* can your get?

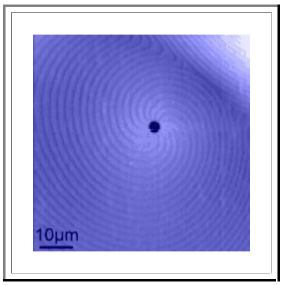
- Well, as you would expect: Not much! Growth rates depend on many parameters, but are in the range of **0.2 - 2 mm/hr**. That's about a factor of **50** slower than the growth rates for **Si** crystal pulling and that makes **SiC** crystal growing automatically expensive

What *polytype* will you get (hoping that it will not be a mixture)? What determines what you get? Can you control it and, if yes, how?



- Good question! First, you might get mixtures as shown in the picture (courtesy of the Erlangen group). Otherwise, the following parameters are essential:
    - Polytype of the seed crystal (as you might have guessed).
    - "Face" of the seed crystal; i.e if the surface is a **C**- or a **Si** layer. If you start with a **4H-SiC** seed crystal, for example, you tend to get **4H-SiC** if you have a **C**-face, and **6H-SiC** if you have a **Si** face. Why? Nobody really knows.
    - Temperature difference and - gradient betwen **SiC** source and seed. Small values tend to favor **4H-SiC**, larger values **6H-SiC** growth.
    - Gas composition. Whatever gas you add will influence the polytype you obtain. **C**-rich gases, for example, promote **4H-SiC** growth
    - The pressure, oddly enough, seems not to have a large influence on polytypie.
- Note that while the polytype **6H** is the easiest to grow, **4H** would be favored by the power electronics industry.

- Last not least: What kind of crystal quality do you get? What is the dislocation density?
    - The bad news is: the dislocation density is high. The good news is, you do not worry too much about that - you worry about something weird called "**micropipe** (and mixtures of polytypes, and all kinds of stacking faults or special boundary faults, and carbon inclusion, or **Si** inclusion, or big voids, ...).
    - To quote an Internet source: "*Problems with micropipes and polytypes dominate to such a degree that the research of dislocations, vacancies and impurities still remains an academic activity*".
- What are micropipes? Well, micropipes are hollow channels running through the lattice; the diameter of these pipes is **(0.1 - 5) µm**.
    - It is not totally clear what micropipes are, how they are formed, and why they exist at all. The probably best way to think about these defects is to consider them to be screw dislocations with a "giant" Burgers vector (violating the rather general rule that Burges vectors always are the shortest possible lattice vectors) and a *hollow core*.
    - The hollow core actually makes sense. If you accept the "giant" Burgesvector bit, it is energetically far more favorable to have the dislocation core hollow instead of extremely strained. What you pay in terms of surface energy, you easily gain in avoided elastic energy.
    - But this is not gospel yet. Micropipes are at present simply not completely understood.
    - Micropipes are also somehow connected to the growth mechanism of the crystal. This is neatly illustrated in the picture on the right (taken with a scanning force microscope, courtesy of H. Strunk; Uni Erlangen) where typical growth spirals are visibly centered around a micropipe.



- Micropipes also will definitely kill any device that contains one of them. They thus must be avoided as much as possible!

- Let's look at the state of the art of what is around. To quote from the product sheet of the major **SiC** supplier Cree, Inc. (located somewhat ironically in Silicon Drive 4600 in Durham, North-Carolina, **USA**):
    - At present, wafer diameters are **50.8 mm** or **76.2 mm**; doping (usually with **N** for **n**-type and **Al** for **p**-type) at high levels produces resistivities in the **0.0x mΩcm** region. Or there is no doping for semi-insulating stuff.
    - **4H**- and **6H-SiC** polytypes are sold; for a more detailed look of some of the products that are available use the link.
- The **2003** state of the art (mostly in the laboratories and not necessarilly on the market) is summarized in the following table:

Diameter		**100 mm** "Four-inch"	For **Si**, **100 mm** was the standard back in the late **70**ties/early **80**ties).
Defects	**Micropipes**	**< 1 cm^{-2}** for **3"** **< 30 cm^{-2}** for **100 mm**	Increasing wafer size usually dramatically increases micropipe density
	Dislocations	**3 · 10^{3} cm^{-2}** achieved	Factor **10** reduction

- Of course, in the many laboratories (university and industrial) devoted to **SiC**, some data might be even better.

Electronic Properties

- The basic electronic properties were listed above and in an illustration module, here we briefly consider **doping** and **optical properties**.
- First let's ask ourselves a question that should have come up by now: Why is **SiC** interesting for **optoelectronic** applications? How could Siemens make light-emitting diodes back in **1977** from an indirect semiconductor?
  - Well, maybe there are bound excitons as in the case of **GaP**? Right - maybe! To quote one of the recommended Books (published **1995**):
    "*The emission (of α-**SiC**; i.e. probably **6H-SiC**) occurs in a wide band from about **400 nm - 600 nm** with a maximum at **480 nm** (blue). So far it is not clear what kind of transition causes the **SiC** emission*".
  - Now you should be glad: There is something left to do - for you!
- The situation becomes a bit clearer, maybe, by pondering another quote (from a very good source in Sveden):
  - " *The viewpoint of a crystal grower differs largely from that of a spectroscopist. The work of a crystal grower is often to provide material of very high quality and sometimes also of high purity. A **PL** spectrum which may look excellent for a crystal grower* (i.e. shows nothing for an indirect semiconductor), *may perhaps not create any higher emotional feelings for a spectroscopist. Indeed, samples which for a crystal grower may be the outcome of failed experiments will be the samples of greatest interest for the spectroscopist*".
  - In other words: **SiC** crystals usually are full of defects with some energy level in the band gap. Besides the levels form the (usually heavy) doping, and all kinds of exciton levels, there are all kinds of atomic defects, spanning the range from simple vacancies to impurity atoms and clusters of atomic defects with levels in the big and roomy band gap of **SiC**.
  - There are thus many possible transitions or recombination channels for electron and holes, and some of those transitions might will emit light.
  - Some more information about the photo luminescence properties of **SiC** can be found in an advanced module, here we simply note that the light emission properties of **SiC** are not so much a property of the ideal (doped) perfect material, but of crystal lattice defects in a general sense.
- However: Whatever recombination events produce light - the quantum efficiency is never very good - the over-all efficiency of the early **LEDs** was **< 1%**. Nevertheless, before the advent of **GaN** in the nineties, **SiC LEDs** were the only ones emitting in the blue.

10.1.2 Silicon Carbide - Applications

- Suffice it to say that now (2019) SiC is finally taking off. The Bosch company is heavily investing in SiC devices for handling power, presumably inside the "electric" car (and in trains, of course).
  - We also see SiC in the power modules handling solar power and in the ultra-high speed power modules needed for the "G5" technology in communications.

10.2 Gallium Nitride - GaN

10.2.1 Basics and History

- Another module that would be interesting to write
    - Do it!

10.3 II - VI Semiconductors

10.4 Semiconducting Polymers

10.4.1 Basics

The Language Barrier

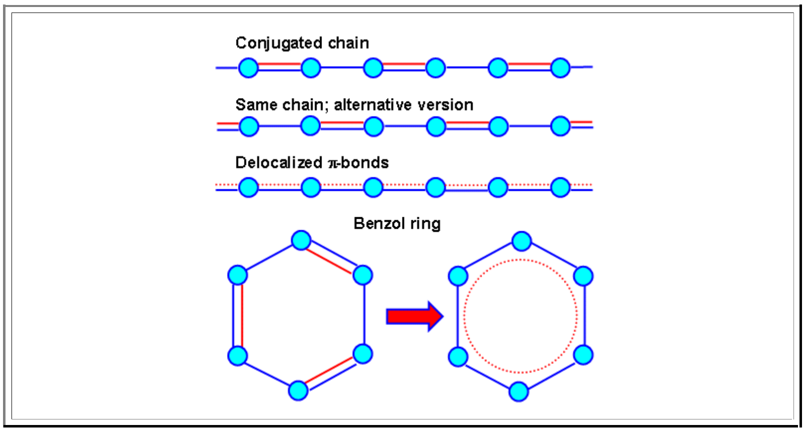
- When we look into the rapidly increasing field of semiconducting polymers, we look into something rather new and rather exciting. We also look at some problems!
  - Knowing a lot of solid state physics in general, and semiconductor physics in particular, as we do if we made it this far in the course, will not be sufficient to understand semiconducting polymers. You will now be biased to consider *all* semiconductors in terms of electrons and holes, which move more or less freely in three more or less equal directions. Their movement happens either in a conduction or in valence band, which have some density of states and exist because of a three-dimensional periodic potential. The overruling parameter is the Fermi energy of the system.
- There is nothing wrong with this particular way of modeling crystalline semiconductors - in terms of equations and in terms of thinking about them. There would be nothing wrong by trying to fit semiconducting polymers into this mold either - except that they do not fit into that mold all that well, and that practically nobody out there, who is working with semiconducting polymers, is doing it.
  - Instead, the language used is heavily biased from chemistry ("oxidation state", ,), and merrily interlaced with terms describing quasi particles ("soliton", polaron, or bi-polaron") that originally were toys from the arsenal of theoretical physicists.
  - Of course, we will also find some semiconductor terms like "holes", "doping", or "recombination", too.
  - This seems to be bad enough. But it is even worse: All those terms may not imply what *you* might think they imply, but something a bit or a lot different.
- We have neither the time, nor the prerequisite knowledge (in this statement I include myself), to delve deeply into the subject. We will just scratch the surface and look at some major points with respect to semiconducting polymers and their applications.
  - In time some more depth might be added to this via advanced modules, and some basic modules may provide more background.
  - At present, however, you may want to look at the polymer modules of other scripts or a short vocabulary of special terms:
    - Einführung in die Materialwissenschaft I
    - Polymers II
    - Vocabulary for Semiconducting Polymer
  - The last link refers to a basic module contained within this Hyperscript offering a short (and not yet complete) vocabulary of the essential terms.

Conducting , Semiconducting, and Insulating Polymers

- When we use the word "polymer" from now on, we only consider so-called **conjugated polymers**. Only these single-bond - double-bond chains are at the heart of conducting and semicondcuting polymers as we will see shortly.
- First, however, we ask ourselves a simple question that is, however, not all that easy to answer:

Why are conjugated polymers not always one-dimensional conductors?

  - Why did we all just "know" that polymers are insulators? Why a Nobel prize for the discovery of *conduction polymers*?
- You probably never thought about it. Nor was this question raised in the more basic polymer stuff you may have learned. What you learned was that polymers are insulators since there are no free electrons. Period.
  - But now we are more advanced and realize, if our nose is rubbed into the matter, that the double bonds along a conjugated chain should not be localized - they should occur to the left or to the right of any given atom with the same probability.
  - In other words, things in conjugated chain *should* be like shown below; in other words exactly as we know it to be in a benzol molecule which is shown with the same symbolism.



- In crystal physics terminology, the π-electrons all overlap and thus must form a band. This band has twice as many states as we have electrons (one for the electron on the left, one for the electron on the right of any chain atom)
  - We thus *should* have (one-dimensional) metallic conductivity along the chain!
- Well, we don't. And the reason for that is a universal pricnciple in physics, called "**symmetry breaking** " at its most general form, and "**Peierls transition**" or "**Peierls instability**" in a more specialized version.
  - A special feature of he picture above, that only emerges because we have a *long chain* , is that the bond length would be the *same* for single and double bonds on average, because the double bond is not localized. For localized double bonds, of course; as we have then in small molecules, we do not expect equal bond lengths for single-, double-, and triple bonds.
  - Equal bond lengths in a chain are thus an expression of a particular symmetry in long chains: Every carbon - carbon pair is equally likely to have the double bond (= have the π-electron) at any given time. But that particular kind of symmetry is *not* a required property of the chain. Symmetries, as we know from many examples, can be broken if we gain some (free) enthalpy by doing it.
  - Bond lengths *could* be different in long chains. Even if there is no directly evident electronic reasons for this breaking of symmetry, all "we" have to do to achieve this effect, is to invest some *elastic energy* that will change the bond lengths via some elastic strain.
  - In a thought experiment, we can easily make the bond lengths alternatingly somewhat shorter and somewhat longer, producing the simple-minded picture of localized (short) double bonds in a comjugated chain. The only question now is if we would be rewarded for doing this. Or in other words: Will we get more energy back by breaking the symmetry than we have to invest for the elastic deformation.
- For a benzol molecule, the answer clearly is *no*! The π-bonds are delocalized (as proven by measurements); there is no symmetry breaking. How about a long conjugated chain?
  - Well - the answer must be *yes* - we *must* have symmetry breaking. Simply because the "experiment" of every day experience tells us that a conjugated polymer chain is an *insulator* and not a conductor, as it would be if the π-electrons could move freely along the chain.
  - The question than is why do we have this **Peierls instability** of conjugated polymer chains - and of many other systems including, e.g. superconductors?
  - Obvioulsy because we get more electronic energy back than we have to invest in elastic energy; there can be no other reason.
- That is not only true, but it is relatively easy to understand in principle what goes on - provided you understood chapter 2 of this lecture course.
  - Here we just note that the simple-minded picture of a conjugated polymer chain with *localized* π-bonds that do *not* allow easy electron movement along the chain is actually rather correct, but for rather involved reasons, which we will consider in more detail in an advanced module
  - Conjugated polymer chains are not conductors, but insulators, or, to be more specific, *semiconductors* with a band gap sufficiently large not to show any intrinsic conductivity at **300 K** or any temperature that does not destroy the molecule anyway.